Mollen, Anne; Hondrich, Lukas

Working Paper
Von der Risikobegrenzung zur Beteiligung - automatisierte Entscheidungssysteme am Arbeitsplatz: Beteiligung von Beschäftigten entlang der Machine-Learning-Pipeline als Perspektive für die europäische Regulierung

Working Paper Forschungsförderung, No. 313

Provided in Cooperation with:
The Hans Böckler Foundation

Suggested Citation: Mollen, Anne; Hondrich, Lukas (2023) : Von der Risikobegrenzung zur Beteiligung - automatisierte Entscheidungssysteme am Arbeitsplatz: Beteiligung von Beschäftigten entlang der Machine-Learning-Pipeline als Perspektive für die europäische Regulierung, Working Paper Forschungsförderung, No. 313, Hans-Böckler-Stiftung, Düsseldorf

This Version is available at:
https://hdl.handle.net/10419/281785

Hans Böckler Stiftung
Mitbestimmung · Forschung · Stipendien

WORKING PAPER FORSCHUNGSFÖRDERUNG

Nummer 313, Dezember 2023

Von der Risikobegrenzung zur Beteiligung – automatisierte Entscheidungssysteme am Arbeitsplatz

Beteiligung von Beschäftigten entlang der Machine-Learning-Pipeline als Perspektive für die europäische Regulierung

Anne Mollen und Lukas Hondrich

Auf einen Blick
Automatisierte Entscheidungssysteme am Arbeitsplatz drohen das Machtverhältnis zwischen Arbeitgebern und Beschäftigten zu verschieben, und zwar zulasten der betroffenen Beschäftigten. Die bisherigen Ansätze zur Risikobegrenzung sind unverzichtbar, doch sollten Beschäftigtenvertreter*innen auch in der Lage sein, ihre Interessen in die automatisierten Entscheidungsprozesse mit einfließen zu lassen. Wie das in der Praxis aussehen kann, zeigt dieser Bericht anhand der Machine-Learning-Pipeline und weist zugleich auf die strukturellen Voraussetzungen für eine erfolgreiche Beteiligung hin.

FORSCHUNGS-
FÖRDERUNG

Aus dem Englischen von Ilja Braun

Georg-Glock-Straße 18, 40474 Düsseldorf
www.boeckler.de

ISSN 2509-2359

Inhalt

Zusammenfassung

Sogenannte People-Analytics-Systeme finden zunehmend Einzug in die Arbeitswelt und unterwerfen Arbeitnehmer*innen verschiedenen Formen des algorithmischen Managements. Algorithmische Entscheidungssysteme (ADM-Systeme) werden für automatisierte Entscheidungen eingesetzt und basieren häufig auf Machine-Learning-Algorithmen (ML-Algorithmen). So werden bei Bewerbungsverfahren Lebensläufe automatisch nach bestimmten Begriffen durchsucht, den Beschäftigten werden Arbeitsschichten zugewiesen, ihre Arbeitsleistung wird automatisch evaluiert, sie werden für Fortbildungen oder eine Beförderung ausgewählt, und manchmal entscheiden diese Systeme sogar darüber, wer auf die Straße gesetzt wird.

Das Versprechen solcher Systeme besteht darin, es den Unternehmen auf Grundlage eines enormen Fundus an Beschäftigtendaten zu ermöglichen, Verfahren effizienter zu gestalten und Entscheidungen, die Beschäftige betreffen, faktenbasiert und objektiver ausfallen zu lassen. Doch obwohl automatisierte Entscheidungen am Arbeitsplatz schwerwiegende Auswirkungen auf die Beschäftigten und ihre Arbeitsumgebung haben, lässt ihre Funktionsweise sich meist nur schwer nachvollziehen, Es unklar bleibt, wie Entscheidungen tatsächlich zustandekommen. Deshalb geht automatisierte Entscheidungsfindung am Arbeitsplatz mit beträchtlichen Risiken einher. Auch der Europäische Gesetzgeber hat das erkannt.

In derzeitigen Gesetzgebungsverfahren, in denen es um Künstliche Intelligenz (KI) geht, wie etwa der geplanten EU-Verordnung zu Künstlicher Intelligenz (KI-Verordnung), werden ADM-Systeme am Arbeitsplatz als sogenannte Hochrisikosysteme eingestuft und deshalb als regulierungsbedürftig angesehen. Wirksame Schutzmechanismen sind hier besonders wichtig, weil ADM-Systeme traditionelle Formen der Interessenvertretung und Mitbestimmung von Beschäftigten auszuhebeln drohen. Intransparent und schwierig zu beaufsichtigen, verstärken sie das Machtungleichgewicht zwischen Arbeitgebern, die die Systeme einsetzen, und Beschäftigten, die deren Entscheidungen ausgesetzt sind.

Aus Beschäftigtenperspektive kann ein Ansatz, der die Risiken dieser Systeme eingrenzen soll, nur als Schutz angesehen werden, um die schwerwiegendsten Risiken zu verhindern. Darüber hinaus sollten die Interessen von Beschäftigten von vornherein in die automatisierten Entscheidungsprozesse der People-Analytics-Systeme einfließen. Beschäftigtenvertreter*innen sollten also schon bei der Entwicklung solcher Systeme für den Einsatz am Arbeitsplatz eingebunden und ihre Interessen berücksichtigt werden.

Vor allem die Intransparenz von ML-Systemen muss vor diesem Hintergrund diskutiert werden. Zum Teil wissen nicht einmal die Entwickler*innen selbst, nach welchen Kriterien die Entscheidungen ihrer Systeme zustandekommen. Das bedeutet aber nicht, dass keinerlei Aufsicht, Kontrolle oder Transparenz möglich wäre. Ein Beteiligungsprozess bei der Entwicklung ist daher nicht ausgeschlossen. In diesem Papier greifen wir auf das Konzept der Machine-Learning-Pipeline zurück, um zu demonstrieren, wie Beschäftigte und ihre Vertreter*innen eigene Interessen in ADM-Systeme integrieren können.

Die ML-Pipeline unterscheidet fünf Schritte im Lebenszyklus eines üblichen, datengetriebenen ADM-Systems:

- Problemdefinition
- Daten
- Modelltraining
- Einsatz
- Retraining

Im Rahmen der Planung und Entwicklung eines ADM-Systems sollten Beschäftigte und ihre Vertreter*innen für jeden dieser Schritte eigene Positionen und Anforderungen entwickeln, um ihre Interessen in diesem Prozess integrieren zu können.

In der Problemdefinitionsphase werden die Aufgaben und Ziele des ADM-Systems bestimmt. Hier sollten Beschäftigte sich dazu äußern können, wozu das System eingesetzt wird und ob die Problemstellung für eine automatisierte Entscheidungsfindung geeignet ist. Auch der Grundansatz und die anzuwendende Methode sollten zur Diskussion stehen. Dabei muss den Beschäftigten bewusst sein, dass es zu anhaltenden Verschiebungen im Machtgefüge einer Organisation kommen kann, wenn das Wissen darüber, wie Organisationsentscheidungen getroffen werden (beispielsweise über Beförderungen), in ein ADM-System ausgelagert wird und somit für Beschäftigte nicht mehr ohne Weiteres verfügbar ist.

Die Datenphase dreht sich um Fragen der Datenerhebung, der Operationalisierung von wesentlichen Konstrukten der Entscheidungsfindung des Systems und um Datenverarbeitung. Hier kommen wichtige Fragen zur Privatsphäre der Angestellten und zum Grad der Überwachung, der sie ausgesetzt sein werden, ins Spiel. Während es vermutlich generell im Interesse der Betroffenen liegen wird, Überwachung am Arbeitsplatz zu begrenzen, könnten ADM-Systeme beispielsweise sensible Informationen sammeln, um später Diskriminierung vorzubeugen. Dies wird stets eine schwierige Abwägung sein und Einzelfallentscheidungen erfordern, bei denen die Betroffenen nicht außen vor bleiben sollten.

In der Phase des Modelltrainings wird die mathematische Funktion extrahiert, die dem in der Problemstellungsphase definierten Zweck des Systems am besten dient. Solche Funktionen können mehr oder weniger intransparent und komplex sein. Insofern nützliche Muster in ML-Systemen manchmal gleichzeitig mit schädlichen Mustern auftreten, sollten Beschäftigte an dieser Stelle darauf achten, dass die extrahierte Funktion ihren Interessen nicht zuwiderläuft – weil sie etwa bestimmte Verzerrungen aus der Verarbeitung früherer Trainingsdaten übernimmt.

Bei der Einsatzphase wird das System in Anwendung gebracht. Hier bedarf es wirksamer Vorkehrungen, um auf eine mögliche Verschlechterung der Leistung des Systems reagieren zu können. Die Ergebnisse des Systems müssen kontinuierlich überprüft werden, um zu verhindern, dass schädliche Muster sich ausbreiten. Wenn die Leistung des Systems nachlässt, erfordert dies in der Regel eine erneute Trainingsphase (Retraining) – auch bei dieser Entscheidung sollten die Beschäftigten einbezogen werden.

Sicherzustellen, dass Beschäftigte und Beschäftigenvertreter*innen ihre Interessen in den Workflow einer ML-Pipeline einbringen können, wenn die Einführung eines ADM-Systems am Arbeitsplatz geplant wird, setzt allerdings ein Umdenken in den Debatten um KI-Gesetzgebung und die mit algorithmischen Prozessen verbundenen Risiken voraus. Während aktuelle Regulierungsvorhaben sich auf Risikobegrenzung konzentrieren, drehen gewerkschaftliche Diskussionen sich häufig um die Gewährleistung ethischer Standards für ADM-Systeme am Arbeitsplatz.

So gewinnbringend diese Debatten waren, so dringend ist nun auch ein nächster Schritt hin zu eher praxis- und prozessorientierten Ansätzen geboten. Gesetzliche Regulierung muss stärker als bisher darauf abzielen, Beschäftigten im Kontext algorithmischer Entscheidungen Mitbestimmungsmöglichkeiten zu verschaffen. Zu denken ist hier etwa an umfassende Transparenzanforderungen, an eine Unterstützung von Governance-Strukturen, die Beteiligung mitdenken, oder an finanzielle Unterstützung für Fortbildungen der Beschäftigtenvertreter*innen im ADM-Bereich.

1. Automatisierung im Personalmanagement: Über Risikobegrenzung hinausdenken

Personalmanagement ist eine vielschichtige Aufgabe mit Verteilungs- und Evaluationsverantwortlichkeiten. Dazu gehört die Einteilung von Schichten und die Zuteilung von Aufgaben, Entscheidungen über Beförderungen und Fortbildungen, Evaluierungen der Arbeitsleistung, das Beenden von Verträgen etc. Angesichts der Komplexität dieser Aufgaben und in der Hoffnung auf Effizienzgewinne ziehen immer mehr Unternehmen den Einsatz von regelbasierten oder Machine-Learning-Algorithmen (ML-Algorithmen) in Betracht.

Darauf haben Gewerkschaften weltweit reagiert, indem sie ihrerseits die Wirkungen solcher algorithmischen Managementansätze untersucht haben. Die größte Befürchtung ist dabei ein zunehmendes Machtungleichgewicht von Arbeitgebern und Beschäftigten, das sich aus der Intransparenz algorithmischer Entscheidungssysteme ergibt, aber auch aus den Datenbeständen, mit denen sie arbeiten sowie aus der zunehmenden Überwachung von Beschäftigten, die sie voraussetzen.

Die Anbieter von Algorithmischen Entscheidungssystemen (ADM-Systemen) legen die Parameter für Entscheidungen fest, die für die Beschäftigten weitreichend sein können. Aber wie diese Entscheidungsprozesse funktionieren, bleibt oft intransparent. 2022 konnte jede*r dritte Beschäftigte in Europa nicht angeben, ob sie am Arbeitsplatz mit einem algorithmischen Managementsystem konfrontiert war oder nicht (Holubová 2022).

Damit sich die Interessen von Beschäftigten berücksichtigen lassen, wenn ADM-Systeme am Arbeitsplatz eingesetzt werden, ist es unerlässlich, ihre Interessen in den Entscheidungen der ADM-Systeme zu integrieren: durch neue Formen der Mitbestimmung, des sozialen Dialogs oder im Rahmen von Tarifverhandlungen.

ADM-Systeme am Arbeitsplatz können schwerwiegende negative Auswirkungen auf Beschäftigte haben, von Diskriminierung über Fehlentscheidungen ohne entsprechende Maßnahmen der Schadensbehebung bis hin zu intransparenten Entscheidungprozessen, für die es keinen Mechanismus der Verantwortungszuschreibung gibt. In gewerkschaftlichen Kreisen scheint ein Konsens über bestimmte Grundsätze und ethische Prinzipien zu bestehen, an denen sich Beschäftigtenvertreter*innen orientieren sollten, wenn sie mit algorithmischen Personalmanagement-Systemen zu tun bekommen.

Grob zusammengefasst, scheinen sich viele Gewerkschaften darauf einigen zu können, dass Transparenz, Fairness, Privatsphäre, Daten-

schutz, Informationsrechte, menschliche Aufsicht, Vorgaben für Hochrisikoanwendungen und Erklärbarkeit grundsätzliche Anforderungen an ADM-Systeme am Arbeitsplatz sein sollten (AlgorithmWatch 2023). Das Bestehen auf diesen Prinzipien ist in erster Linie eine Strategie der Risikobegrenzung.

Auch aktuelle Legislativvorschläge, wie die geplante EU-Verordnung zu Künstlicher Intelligenz (KI-Verordnung), konzentrieren sich vornehmlich auf eine Minimierung der mit ADM-Systemen verbundenen Risiken. So sieht die vorgeschlagene europäische Verordnung eine Selbstauskunft der Hersteller von KI-Systemen vor, mit der sichergestellt werden soll, dass Hochrisiko-KI-Systeme mit Grundrechten vereinbar sind.

KI-Systeme für die Arbeitswelt werden als hochrisikoreich eingestuft, wenn es um die Bekanntmachung freier Stellen, das Sichten oder Filtern von Bewerbungen, das Bewerten von Bewerber*innen in Vorstellungsgesprächen oder Tests, um Kündigungen von Arbeitsvertragsverhältnissen, die Aufgabenzuweisung sowie die Überwachung und Bewertung der Leistung und des Verhaltens von Personen in solchen Beschäftigungsverhältnissen geht (Europäische Kommission 2021a). Während die KI-Verordnung also die mit ADM-Systemen am Arbeitsplatz verbundenen Risiken anerkennt, sieht es nur sehr grundlegende Schutzmaßnahmen vor.

Beschäftigtenvertreter*innen sollten über den Ansatz der Risikobegrenzung hinausdenken und Gelegenheiten zu Interventionen im Interesse der Beschäftigten identifizieren. Ihre Beteiligung sollte darauf abzielen, ADM-Systeme im Interesse der Betroffenen mitzuprägen. Derzeitige Gesetzgebungsprozesse wie die KI-Verordnung sind in dieser Hinsicht aber unzureichend, sodass es weiterer Vorgaben auf europäischer oder nationaler Ebene bedarf.

Wir plädieren für einen Paradigmenwechsel in der europäischen Regulierungsperspektive beim Einsatz von ADM-Systemen am Arbeitsplatz. Risikobegrenzung bietet einen unverzichtbaren, grundlegenden Schutz, aber darüber hinaus müssen Beschäftigte in die Lage versetzt werden, ihre Interessen bei der Entwicklung von ADM-Systemen integrieren zu können. Die dringende Frage lautet also: Wie können Beschäftigte ihre Interessen mit Blick auf ADM-Systeme am Arbeitsplatz artikulieren und in die Systeme einbinden? Oder genauer gefragt: Wo gibt es mit Blick auf People-Analytics-Systeme für Beschäftigteninteressen einen Handlungs- und Gestaltungsspielraum?

Diese Frage ist nicht trivial. ADM-Systeme werden oft als Black Boxes beschrieben, deren Arbeitsweise und Ergebnisse schwer nachzuvollziehen und zu beeinflussen sind. ML-Systeme lernen autonom, oft unbeaufsichtigt und ohne weitere Überprüfung. Grundsätzlich kann sich die Logik, nach der sie arbeiten, ständig verändern. Deshalb wird mitunter befürch-

tet, es könnte unmöglich sein, die Interessen von Beschäftigten wirksam zu implementieren. Solche Vermutungen basieren auf einer Reihe von Mythen, die sich um ADM-Systeme ranken. Sie müssen aufgeklärt werden, um das Gefühl von Apathie und Machtlosigkeit im Hinblick auf die Interessenvertretung Beschäftigter im ADM-Kontext zu überwinden.

Im folgenden Kapitel 2 werden wir einige wichtige Begriffe definieren, um dann zu eruieren, wo sich bei ADM-Systemen Spielräume für die Interessenvertretung eröffnen. Wir nutzen das Konzept der ML-Pipeline, um Transparenz zu schaffen und ADM-Systeme in Workflows und Abschnitte innerhalb ihres gesamten Entwicklungs- und Einsatzzyklus herunterzubrechen sowie um Punkte zu identifizieren, an denen Beschäftigte ihre Interessen einbringen können.

Unser Ziel ist zu zeigen, wie Beschäftigte diese Einstiegspunkte auf technischer sowie organisationeller Ebene identifizieren können, um ihre Interessen einzubringen. Sie zu nutzen, setzt voraus, dass Beschäftigtenvertreter*innen in der Lage sind, ihre Interessen präzise und anwendungsorientiert zu formulieren.

Wir greifen auf fiktionale, aber realistische Beispiele aus der Arbeitswelt zurück und zeigen, dass die ML-Pipeline dafür genutzt werden kann, Transparenz zu schaffen und die Interessen der Beschäftigten zu inkorporieren. Im nächsten Schritt blicken wir in die Zukunft und zeigen auf, wie eine größere Partizipation von Beschäftigten erreicht und ihrer Stimme mehr Gehör verschafft werden kann – und in welchem Zusammenhang dies mit Verpflichtungen für Hersteller und Nutzernde in zukünftiger KI-Gesetzgebung steht.

2. People Analytics: Risiken und Transparenz für Betroffene

Neue technische Entwicklungen werden oft als übermächtig, nahezu magisch dargestellt. Das gilt besonders für die sogenannte Künstliche Intelligenz (Campolo/Crawford 2020) oder ADM-Systeme. Die Mystifizierung der Technologie ist in zweierlei Hinsicht problematisch. Erstens führt sie zu einer Überschätzung der tatsächlichen Leistungsfähigkeit von ADM-Systemen. Zweitens steht sie einem klaren Verständnis davon im Wege, was solche Systeme tatsächlich tun und wie sie arbeiten.

Deshalb ist es von kaum zu unterschätzender Wichtigkeit, genau zu definieren, wovon die Rede ist, wenn es um Automatisierungsprozesse am Arbeitsplatz geht. Wir müssen sehr genau verstehen, um welchen Untersuchungsgegestand es sich handeltn, wenn wir ADM-Systeme am Arbeitsplatz und People-Analytics-Vorgänge analysieren.

2.1 ADM-Systeme

Um nicht über die Fallstricke zu stolpern, die der mystifizierte Begriff der Künstlichen Intelligenz uns legt, soll hier die Rede von automatisierten Entscheidungssystemen, von ADM-Systemen sein. Automatisiert heißt in diesem Fall: Automatisierung entweder durch regelbasierte Algorithmen oder durch Maschinelles Lernen, also ML-Systeme, worunter ebenfalls neuronale Netze und Deep-Learning-Systeme fallen. Obwohl viele Probleme auch beim Einsatz einfacherer, regelbasierter Algorithmen auftreten können, konzentrieren wir uns in diesem Papier auf ADM-Systeme, die auf Machine Learning basieren.

Maschinelles Lernen bezeichnet „einen Ansatz, der auf dem Erlernen komplexer Muster in bereits vorhandenen Daten beruht und diese Muster für Voraussagen über zukünftige Daten verwendet“ (Huyen 2022, eigene Übersetzung aus dem Englischen). Solche Muster können z. B. auf simplen Korrelationen beruhen, wie etwa zwischen Dienstjahren und Gehalt, oder komplexe, nichtlineare Beziehungen abbilden, etwa zwischen der Arbeitserfahrung, den Fähigkeiten und der Arbeitsleistungsbewertung von Beschäftigten und ihrer Eignung für freie Stellen. Im Allgemeinen neigen Systeme letzterer Art zu größerer Komplexität im Hinblick auf Transparenz, Kontrollierbarkeit und Mitbestimmungsmöglichkeiten.

2.2 People-Analytics-Software

Die Grundlage dieses Papiers sind unsere Erkenntnisse zum Einsatz von ADM-Systemen am Arbeitsplatz, insbesondere solcher zum algorithmischen Personalmanagement. Wir bezeichnen solche Systeme als People-Analytics-Systeme, aber andere Begriffe wie Human Resource Analytics oder Workforce Analytics sind ebenfalls verbreitet. Gemeint sind stets softwarebasierte Systeme, die Datenanalyse und Automatisierung für typische Aufgaben aus dem Bereich der Personalplanung und des Belegschaftsmanagements einsetzen.

Solche Systeme verarbeiten Beschäftigtendaten, um Erkenntnisse und Informationen über diese Beschäftigten zu gewinnen und grafisch darzustellen. Deskriptive, prädiktive und präskriptive Elemente können hier allesamt enthalten sein. Aufgrundlage einer Analyse der erhobenen Daten können die Systeme dann die Beschäftigten betreffende Hypothesen aufstellen oder Entscheidungen treffen (Gießler 2021). Um ein People-Analytics-System verwenden zu können, ist es deshalb unverzichtbar, systematisch Daten über die Belegschaft zu sammeln.

People-Analytics-Systeme können für eine Vielzahl von Zwecken eingesetzt werden: bei der Personalbeschaffung (indem Bewerbungen algorithmisch gescannt und gefiltert werden), für die automatisierte Planung von Arbeitsschichten, die Zuweisung von Aufgaben, die Evaluierung der Produktivität und der Arbeitsergebnisse, für die Verbesserung der Arbeitssicherheit, um bestimmte Beschäftigte für Fortbildungen oder Beförderungen auszuwählen, um ihre Loyalität zum Arbeitgeber oder die Wahrscheinlichkeit eines Arbeitsplatzwechsels einzuschätzen etc.

Die Versprechungen der Anbieter von People-Analytics-Systemen sind ebenso weitreichend wie die Einsatzmöglichkeiten. Einige davon haben potenziell schwerwiegende Auswirkungen auf die Organisationsstruktur und die Machtverhältnisse zwischen Arbeitgebern und Beschäftigten (Jarrahi et al. 2021).

2.3 Risiken von People-Analytics-Systemen

In seiner derzeitigen Fassung benennt die geplante KI-Verordnung der EU Risiken, die mit ADM-Systemen am Arbeitsplatz einhergehen. Besonders hoch sind die Risiken nach dem Gesetzesvorschlag in den folgenden Bereichen:

„a) KI-Systeme, die bestimmungsgemäß für die Einstellung oder Auswahl natürlicher Personen verwendet werden sollen, insbesondere für die Bekanntmachung freier Stellen, das Sichten oder Filtern von Bewerbungen und das Bewerten von Bewerbern in Vorstellungsgesprächen oder Tests;
b) KI-Systeme, die bestimmungsgemäß für Entscheidungen über Beförderungen und über Kündigungen von Arbeitsvertragsverhältnissen, für die Aufgabenzuweisung sowie für die Überwachung und Bewertung der Leistung und des Verhaltens von Personen in solchen Beschäftigungsverhältnissen verwendet werden sollen“ (Europäische Kommission 2021a, S. 5).

Die Liste konzentriert sich vor allem auf Einsatzmöglichkeiten, die die persönlichen Karrierechancen der Betroffenen beeinträchtigen könnten. Damit erfasst sie einige der wichtigsten Risiken solcher Systeme, etwa im Bereich der Aus- und Weiterbildung, der Jobsuche, der Anstellung, der Zuweisung von Aufgaben, der Leistungsevaluation und schließlich der Beendigung des Arbeitsverhältnisses.

Das KI-Gesetz zielt darauf ab, dass Hersteller von Hochrisiko-KI-Anwendungen bestimmte Bedingungen erfüllen sollen, um die Risiken der Nutzung solcher Systeme zu begrenzen. Dabei geht es um Vorgaben zu Daten und Daten-Governance (Artikel 10), technische Dokumentation (Artikel 11), Aufzeichnungspflichten (Artikel 12), Transparenz und Bereitstellung von Informationen für die Nutzer (Artikel 13), menschliche Aufsicht (Artikel 14) sowie um Genauigkeit, Robustheit und Cybersicherheit (Artikel 15).

Eine wesentliche Unzulänglichkeit dieses risikobasierten Ansatzes liegt in seinem Fokus auf der Marktfähigkeit von Produkten. KI-Systeme sollen bestimmte Anforderungen erfüllen, damit sie in der Europäischen Union zugänglich gemacht werden dürfen. Damit wird die Verantwortlichkeit hauptsächlich den Anbietern von KI-Systemen und nicht den Nutzern zugewiesen. In der Logik des KI-Gesetzes ist ein Unternehmen, das ein People-Analytics-System für seine Personalverwaltung einsetzt, nämlich kein Anbieter, sondern ein Nutzer.

Was KI-Systeme am Arbeitsplatz betrifft, dreht sich die geplante KI-Verordnung also um Risikominimierung und nicht darum, Beschäftigte in die Lage zu versetzen, beim Einsatz von People-Analytics-Systemen ihre eigenen Interessen einzubringen. Aber Risikobegrenzung reicht nicht. Vielmehr muss das Problem angegangen werden, dass die Stimmen der Beschäftigten, wenn ADM-Systeme am Arbeitsplatz eingeführt werden, keinen Eingang in die Entscheidungsfindung dieser Systeme finden.

2.4 Von Risikobegrenzung zu Mitbestimmung

Es ist schwer vorstellbar, dass die Vorgaben aus Artikel 10 der KI-Verordnung zu Datenerhebung, Datenaufbereitung und Datenmanagement aus Beschäftigtenperspektive ausreichend sind, insbesondere vor dem Hintergrund der diversen Überwachungspraktiken, die bereits gang und gäbe in der Arbeitswelt sind. Das Ausmaß, in dem Daten gesammelt und nachverfolgt werden, stimmt ausgesprochen nachdenklich im Hinblick auf Überwachung am Arbeitsplatz und den Rückschlüssen, Analysen und Profilbildungen, die durch People-Analytics-Systemen am Arbeitsplatz durchgeführt werden können.

Ein Bericht der gemeinnützigen Organisation Cracked Labs aus dem Jahr 2021 gibt einen ebenso umfassenden wie besorgniserregenden Überblick über die derzeit auf dem Markt (und im Einsatz) befindlichen People-Analytics-Systeme (Christl 2021). Der Bericht listet Überwachungspraktiken auf, die für die Nutzung bestimmter ADM-Systeme am Arbeitsplatz notwendig sind. Diese reichen von Sensoren, die Beschäftigte am Körper tragen, über Kameras, die ihre Bewegungen am Arbeitsplatz erfassen, bis hin zu Systemen, die Mausbewegungen und Clicks nachverfolgen, Kommunikationskanäle durchleuchten (darunter auch private Social-Media-Kanäle) und zur Nutzung von Logfiles (siehe auch Krzywdzinski et al. 2022).

Vermutlich sind all diese weitreichenden Überwachungspraktiken nur schwerlich mit der derzeitigen oder der zukünftigen EU-Gesetzeslage in Einklang zu bringen. So oder so sollte man sich aber mit ihren Risiken auseinandersetzen. Mit dem immer intensiveren Sammeln von Beschäftigtendaten geht das Risiko einher, dass verschiedene Datenbestände über Systeme und Plattformen hinweg miteinander verbunden werden, wodurch es in großem Maßstab möglich wird, Profile von Beschäftigten anzulegen und komplexe Analysen ihrer Daten durchzuführen.

In dieser Hinsicht sind die fehlende Transparenz der Systeme, die fehlende Rechenschaftspflicht von Herstellern und Anwendern sowie die fehlenden Mitgestaltungsmöglichkeiten von Beschäftigtenvertreter*innen ausgesprochen problematisch. Es ist oft unklar, wie People-Analytics-Systeme funktionieren und wie ihre Entscheidungen zustandekommen, ob diese solide, fair und begründet sind. Aber auch, ob Beschäftigte über die automatisierte Entscheidungsfindung in Kenntnis gesetz wurden und ob problematischen oder falschen Entscheidungen in irgendeiner Weise vorgebeugt wurde.

Die KI-Verordnung sieht für Produkte, die neu auf den Markt kommen, im Vorfeld eine Risikoabschätzung vor. Das reicht aber nicht aus. ADM-

Systeme müssen vielmehr einzelfallbezogen im Rahmen der realen Nutzung untersucht und bewertet werden. Dafür liegt es nahe, bestehende Interessenvertretungsstrukturen zu stärken, damit Beschäftigte beim Einsatz von ADM-Systemen ihre Perspektive einbringen können. Frühere Erfahrungen haben jedoch gezeigt, dass die Intransparenz dieser Systeme ein großes Problem darstellt und dass im Hinblick auf die Einbringung der Beschäftigteninteressen in die Konstruktion der Systeme Schwierigkeiten bestehen.

2.5 Transparenz als Vorbedingung

Wie Transparenz geschaffen werden kann, und wie hierüber Informationen über ADM-Systeme Beschäftigten sowie ihren Interessenvertreter*innen zugänglich gemacht werden können, wird im Weißbuch der Europäischen Kommission zur Künstlichen Intelligenz skizziert (Europäische Kommission 2020a). Transparenz ist ebenfalls eine der wichtigsten Anforderungen, die in der KI-Verordnung (u. a. in Artikel 13) an Hochrisikosysteme gestellt werden, um Risiken dieser Systeme zu minimieren. Jenseits eines Risikobegrenzungsansatzes kommt es aber darauf an, dass Beschäftigte ihre Interessen wirksam in das Design und den Implementierungsprozess von People-Analytics-Systemen einbringen können.

Im folgenden Kapitel 3 wird auf das Konzept der ML-Pipeline zurückgegriffen, um aufzuzeigen, wie Transparenz hergestellt werden kann. Außerdem zeigen wir Möglichkeiten auf, die Interessen von Beschäftigten in das Design und den Implementierungsprozess von ADM-Systemen am Arbeitsplatz einfließen zu lassen. Vorher soll jedoch noch der Stand der Diskussionen zusammengefasst werden, die derzeit im gewerkschaftlichen Raum über People-Analytics-Systeme geführt werden.

3. Gewerkschaftliche Perspektiven auf Beschäftigteninteressen bei ADM-Systemen

Die derzeitigen Verhandlungen zur KI-Verordnung der EU bieten eine gute Gelegenheit zusammenzufassen, an welchem Punkt gewerkschaftliche Debatten über die Regulierung von ADM-Systemen derzeit stehen. Im Rahmen der vorgelagerten öffentlichen Konsultation hatten interessierte Stakeholder und Einzelpersonen die Möglichkeit, Stellungnahmen zum geplanten Gesetzesvorhaben abzugeben. Eine Reihe von Gewerkschaften und gewerkschaftlichen Dachverbänden sowie einige zivilgesellschaftliche Organisationen haben dabei spezifisch auf die Arbeitswelt bezogene Stellungnahmen abgegeben. Diese haben wir analysiert.

Ebenso haben wir uns einen Überblick über die gewerkschaftlichen Debatten in acht europäischen Ländern verschafft: Tschechien, Estland, Deutschland, Ungarn, Italien, Polen, Spanien und Schweden. Im Folgenden fassen wir zusammen, wie gewerkschaftliche Verbände und Organisationen sich mit Transparenzfragen und Verantwortungsstrukturen beim Einsatz algorithmischer Systeme am Arbeitsplatz auseinandersetzen, teils mit, teils ohne direkten Bezug zur KI-Veordnung.

Unser Ziel war es zunächst die Positionen wichtiger Stakeholder aus dem Bereich der Arbeitswelt zu legislativen Ansätzen, die dem Schutz der Beschäftigten im Zusammenhang mit der Einführung von People-Analytics-Systemen dienen, zu systematisieren. Auf dieser Grundlage haben wir skizziert, wie Beschäftigte beim Design und bei der Einführung von ADM-Systemen ihre Interessen einbringen können, ohne sich auf Risikobegrenzung beschränken zu müssen.

Wir stützen uns dabei auf die Arbeit mit Stakeholdern, die sowohl auf konzeptioneller als auch im Alltag der Beschäftigtenvertretung Erfahrung mit ADM-Systemen haben. Unser Ziel ist, aus der Theorie in die Praxis zu gelangen und Ideen für konkrete Instrumente und Prozesse zu entwickeln, etwa dazu, wie Transparenz bei ADM-Systemen erreicht werden kann und wie Beschäftigtenvertreter*innen bei der Einführung solcher Systeme die Interessen der Betroffenen sinnvoll einbringen können.

Basierend auf unserer Arbeit mit gewerkschaftlichen Interessenvertreter*innen haben wir Konzepte dafür entwickelt, wie ADM-Systeme im Prozess der ML-Pipeline so gestaltet werden können, dass Beschäftigteninteressen adäquat in sie einfließen. Im Folgenden fassen wir die wichtigsten Erkenntnisse und Schlussfolgerungen zusammen.

3.1 Geteilte, aber oft vage Vorstellungen von Risiken

Viele Risiken werden von gewerkschaftlichen Stakeholdern ähnlich wahrgenommen. Ihre Befürchtungen kommen in den Antworten auf die öffentliche Konsultation (Europäische Kommission 2020b) zur KI-Verordnung zum Ausdruck, finden sich aber auch in ähnlichen Veröffentlichungen von Gewerkschaften wie industriAll (2021), UNI Europa (2019), des Deutschen Gewerkschaftsbunds (DGB 2021), des Europäischen Gewerkschaftsinstituts (ETUI; Ponce del Castillo 2021) und anderer. Auch die Arbeit weiterer gewerkschaftlicher Organisationen zu Algorithmen am Arbeitsplatz, die wir über acht europäische Länder hinweg analysiert haben (AlgorithmWatch 2023), unterstreicht ähnliche Bedenken.

Typischerweise drehen sich die Befürchtungen um den Verlust von Arbeitsplätzen und um eine Dequalifizierung der Beschäftigten in Folge zunehmender Automatisierung, aber auch um die Gefahr von Diskriminierungen und Fehlentscheidungen von ADM-Systemen. Auch eine Zunahme von Überwachung am Arbeitsplatz beunruhigt die Akteure, ebenso wie mögliche Verletzungen der Privatsphäre Beschäftigter.

Hinzu kommen abstraktere, allgemeinere Sorgen im Hinblick auf ML-Systeme, insbesondere, dass deren Lernprozesse ohne menschliche Aufsicht zu einem Kontrollverlust und fehlender Nachvollziehbarkeit ihrer Entscheidungsfindung führen könnten, und zwar sowohl bei jenen, die sie einsetzen, als auch bei den von ihrem Einsatz betroffenen Beschäftigten. Die Diskussion über solche Gefahren verläuft anscheinend recht einheitlich, obwohl die Risiken eher allgemein benannt werden.

3.2 Fokus auf Risikobegrenzung

Da die geplante KI-Verordnung von vornherein einen risikobasierten Ansatz wählt, ist es wenig überraschend, dass sich auch die Stellungnahmen zur öffentlichen Konsultation darauf konzentrieren, wie Risiken des Einsatzes von KI-Systemen am Arbeitsplatz eingedämmt werden können. Zum Beispiel haben mehrere Gewerkschaften Transparenzanforderungen, Auditverfahren für Hochrisikosysteme und Maßnahmen zum Schutz der Privatsphäre vorgeschlagen, insbesondere ein Verbot des Trackings von Beschäftigten, etwa mit biometrischen Technologien und Körpersensoren – siehe die Stellungnahmen der „Association of Nordic Engineers“ (2021) und des Deutschen Gewerkschaftsbunds (DGB 2021).

Auch jenseits der Verhandlungen über das KI-Gesetz weisen Stellungnahmen und Publikation aus dem gewerkschaftlichen Bereich vor allem

auf die Notwendigkeit hin, den Risiken vorzubeugen, die mit dem Einsatz von Automatisierung am Arbeitsplatz einhergehen. Als entsprechende Risikobegrenzungsstrategien werden etwa ethische oder regulatorische Prinzipien vorgeschlagen, die für die Entwicklung und Implementierung von ADM-Systemen am Arbeitsplatz verbindlich werden sollen (Algorithm Watch 2023).

Der Schwerpunkt liegt dabei meist auf einer Eingrenzung der Risiken von ADM-Systemen am Arbeitsplatz. So wird z. B. vorgeschlagen, die Anwender auf menschliche Aufsicht und Datenschutz zu verpflichten und Anforderungen an die Transparenz, Unabhängigkeit, Fairness und Sicherheit der Systeme festzuschreiben.

3.3 Erste Forderungen nach partizipativen Ansätzen

Eher im Hintergrund stehen bislang Forderungen nach Mitbestimmungs- und Beteiligungsansätzen, die Beschäftigten und ihren Vertreter*innen die Möglichkeit verschaffen würden, Risiken und Gefahren von ADM-Systemen am Arbeitsplatz vorzubeugen, indem sie diese ihren eigenen Interessen entsprechend mitprägen. Gleichwohl haben der Deutsche Gewerkschaftsbund (DGB 2021) und die „Association of Nordic Engineers“ (2021) in ihren Stellungnahmen zur öffentlichen Konsultation der KI-Verordnung sehr wohl die Notwendigkeit solcher Governance-Ansätze hervorgehoben.

Jenseits der KI-Verordnung hat etwa die Schweizerische Gewerkschaft Syndicom auf die Bedeutung von Sozialpartnerschaft und Mitarbeiterbeteiligung hingewiesen (2020). Bei Syndicom taucht dies als eines von neun Leitprinzipien für die Entwicklung und Einführung von ADM-Systemen am Arbeitsplatz auf.

Solche Forderungen werden durch Verweise auf etablierte Mitbestimmungspraktiken unterstrichen. Beschäftigte, so die Kernaussage, müssen sich einbringen können, insbesondere wenn ADM-Systeme fundamentale ethische Fragen aufwerfen. Doch während es zahllose Forderungen nach umfassender Transparenz gibt, lassen sich die Vorschläge, wie eine solche partizipatorische Governance in der Praxis aussehen könnte, an einer Hand abzählen. Die britische „Trade Union Confederation“ (TUC 2022) ist einer der ersten gewerkschaftlichen Akteure, die Verhandler*innen praxisorientierte Leitlinien für Kollektivvereinbarungen über ADM-Systeme an die Hand gibt.

Genau solcher praktischen Leitlinien bedarf es, um über die Artikulation unspezifischer Ängste und über Risikobegrenzungsstrategien hinauszukommen. Im folgenden Kapitel 4 wird das Konzept der ML-Pipeline herangezogen, um zu demonstrieren, dass partizipatorische Governance-Ansätze bei People-Analytics-Systemen durchaus realisierbar sind.

4. Beschäftigteninteressen entlang der ML-Pipeline

Die mit People-Analytics-Systemen verbundenen Risiken wirksam zu begrenzen, ist von kaum zu unterschätzender Bedeutung. Für den Einsatz solcher Systeme am Arbeitsplatz sollte man aber noch einen Schritt weitergehen und thematisieren, wie Beschäftigtenvertreter*innen sich produktiv für die Interessen von Beschäftigten in Bezug auf solche Systeme einsetzen können. Denn mit ADM-Systemen können für Beschäftigte auch Vorteile verbunden sein.

Sie müssen also in die Lage versetzt werden, diese nutzbar zu machen. Das wiederum setzt voraus, dass ihre Interessenvertreter*innen die Anliegen der Beschäftigten übersetzen und in People-Analytics-Systemen wirksam werden lassen können. Im Folgenden werden wir anhand des Konzepts der ML-Pipeline zeigen, wo Ansatzpunkte für Mitbestimmung bei der Entwicklung von People-Analytics-Systeme liegen.

Um die Transparenz, die Überprüfbarkeit (etwa im Rahmen eines Audits) und die Verwaltbarkeit von ML-Systemen zu verbessern, kann das Konzept der ML-Pipeline genutzt werden (für eine Diskussion über systematische Verzerrungen von ADM-Systemen siehe Schelter/Stoyanovich 2020). Es beinhaltet fünf Schritte, die sich über den gesamten Entwicklungs- und Einsatzzeitraum eines üblichen datenbasierten ADM-Systems verteilen:

- Problemdefinition
- Daten
- Modelltraining
- Einsatz
- Retraining

Diese Schritte werden hier generisch definiert, damit sie auf möglichst viele datenbasierte ADM-Systeme anwendbar sind. Die Differenzierung kann dann genutzt werden, um die Implikationen eines ADM-Systems und seiner Auswirkungen auf die Betroffenen weiter zu analysieren (siehe Abbildung 1). Vor allem aber können diese Schritte genutzt werden, um ML-Systeme von vornherein so zu entwerfen, dass die Anliegen und Grundrechte von allen involvierten Parteien gewahrt werden.

Abb. 1: Beschäftigteninteressen entlang der ML-Pipeline

/ ML-PIPELINE	/ BEISPIELE FÜR BESCHÄFTIGTENINTERESSEN ENTLANG DES ML-LEBENSZYKLUS
1 PROBLEMDEFINITION	Welches Problem soll gelöst werden und wie?
2 DATEN	Welche Daten werden verwendet, und wie werden sie interpretiert?
3 MODELLTRAINING	Welche Methoden sind am besten geeignet, um das definierte Ziel zu erreichen?
4 EINSATZ	Wie werden die Ergebnisse des Systems operativ verwendet?
5 RETRAINING	Welche Maßnahmen zur Qualitätssicherung werden ergriffen?

Quelle: eigene Darstellung

Derzeit werden ML-Systeme in People-Analytics-Verfahren in der Regel eingesetzt, um die wirtschaftlichen Interessen eines Unternehmens voranzubringen, also etwa zur Kostenreduktion oder Möglichkeiten auszuweiten. Es ist wichtig, sich bewusst zu machen, dass jeder Schritt, der innerhalb des Entwicklungs- und Einsatzzyklus eines solchen ML-Systems unternommen wird, auf dieses Ziel ausgerichtet ist.

Daraus folgt, dass die Interessen der Beschäftigten ebenfalls bei jedem dieser Schritte berücksichtigt werden müssen: von der Definition des Anwendungsfalls für ein ADM-System über die Wahl der richtigen Methode, die Operationalisierung wichtiger Konzepte, die Datenerhebung und das Modelltraining bis hin zu konkreten Handlungsableitungen, zur Messung der Ergebnisse und zur Systempflege. Daher können vereinzelte Interventionen an bestimmten Schritten in diesem Prozesse im Interesse der Beschäftigten nicht ausreichen.

Vielmehr kann die ML-Pipeline eine sinnvolle Unterstützung dafür sein, die Perspektive der Beschäftigten systematisch in allen Stadien und Entwicklungsschritten zu integrieren. Interessenvertreter*innen, die sich dieses Konzepts bedienen, werden dadurch in die Lage versetzt, die Anliegen der Beschäftigten bei der Planung, der Entwicklung und schließlich auch beim Einsatz von ADM-Systemen in geeigneter Weise einfließen zu lassen.

4.1 Problemdefinition

Die Problemdefinitionsphase ist dazu da, die Fragestellungen auszuloten, greifbare Ziele zu definieren und festzulegen, wie ein ADM-System entworfen und in die Organisation eines Unternehmens integriert werden kann, um die zuvor identifizierten Ziele zu erreichen. Dazu werden in aller Regel Verhandlungen mit unterschiedlichen Stakeholdern geführt, wobei die von dem System Betroffenen unbedingt mit am Tisch sitzen sollten. Außerdem werden die computationalen und personellen Ressourcen definiert, die benötigt werden. Spezifische methodologische und technische Erfordernisse, wie etwa Mindestanforderungen an die Leistungsfähigkeit des Systems und sein Einsatzbereich, werden ebenfalls erörtert.

Der explorative Charakter der Entwicklung von ML-Systemen bedingt, dass diese Parameter bei manchen Projekten eher vage definiert oder zunächst nur auf Zwischenergebnisse bezogen werden. Während des gesamten Entwicklungsprozesses ist deshalb Flexibilität ebenso gefragt wie die Fähigkeit, auf Probleme reagieren zu können, die man nicht vorhergesehen hat. Beschäftigtenvertreter*innen sollten ihre Stimme gerade innerhalb der Problemdefinitionsphase erheben, denn zu diesem Zeitpunkt werden fundamentale Fragen im Hinblick auf den Zweck und den Einsatzbereich eines ADM-Systems verhandelt, ebenso wie daraus folgende Veränderungen in der Organisationsstruktur.

4.1.1 Grundlagen schaffen

Wegen ihres nichttechnischen Charakters wird die Problemdefinitionsphase häufig nicht als integraler Bestandteil der ML-Pipeline betrachtet. Es handelt sich jedoch um die Hauptplanungsphase, in der Richtungsentscheidungen über grundsätzliche Herangehensweisen getroffen werden. Aus unserer Sicht ist sie deshalb eine besonders kritische Phase, in der eine Artikulation der Anliegen wichtiger Stakeholder nicht unter den Tisch fallen darf.

Ein Problem zu definieren und eine brauchbare technische und organisationelle Lösung dafür zu finden, ist ein oft kontroverser und verhandlungstechnisch entscheidender Moment. Eine ungleiche Machtverteilung, die Nichtberücksichtigung oder eine inadäquate Repräsentation berechtigter Interessen in diesem Stadium zieht sich sonst leicht durch den gesamten Entwicklungs- und Verwendungszeitraum eines ADM-Systems hindurch.

Zum Beispiel kann der Einbezug vieler unterschiedlicher Stakeholder, darunter auch Beschäftigte oder Angehörige von in der Vergangenheit marginalisierten Gruppen, entscheidend dafür sein, eine Diskriminierung durch ADM-Systeme zu vermeiden. Wenn die Stimmen solcher Stakeholder bereits in den Planungsprozess einfließen, steigt bestenfalls die Sensibilität für möglicherweise diskriminierende Ergebnisse, sodass angemessene Fairness-Kriterien in das System integriert werden können.

4.1.2 Langfristige Verschiebungen in der Machtverteilung

Die Einführung von ADM-Systemen in einem Unternehmen besteht nicht nur in der Automatisierung von Prozessen, sondern geht häufig mit einer Restrukturierung von Organisationsprozessen einher. Dies kann langfristige Auswirkungen auf Machtverteilungen innerhalb der Organisation haben, die Beschäftigtenvertreter*innen unbedingt im Blick haben sollten.

Wenn eine Organisation z. B. ein ADM-System einführt, um Teile der internen und externen Einstellungsprozesse für Mitarbeiter*innen zu automatisieren, geht Beschäftigten möglicherweise wertvolles Wissen über die Besetzungsverfahren verloren. Wenn ein einziges System den internen Stellenmarkt eines Unternehmens verwaltet – und nicht mehr möglicherweise etliche auf verschiedene Stellen dieses Unternehmens verteilte Personen –, ergeben sich daraus möglicherweise negative Folgen für die Verhandlungsmacht der Beschäftigten.

Während das ADM-System und diejenigen, die damit arbeiten, relevante Daten und Erkenntnisse über Karrierewege und das Potenzial einzelner Beschäftigter zunehmend zentralisieren, bleibt dieses Wissen über den unterstellten Wert jener Beschäftigter für das Unternehmen der Belegschaft verborgen. Solche Wissensungleichgewichte beeinträchtigen dann unter Umständen die Möglichkeit, faire Vergütungen auszuhandeln.

Bei der Problemdefinitionsphase geht es also nicht nur darum festzulegen, wie weit das Anwendungsspektrum eines bestimmten ADM-Systems reichen soll. Sondern es sollte auch darum gehen, alle relevanten Stakeholder in eine Diskussion darüber einzubeziehen, welche Auswirkungen davon möglicherweise auf Informationsflüsse und Entscheidungswege im Betrieb ausgehen, sowie welche langfristigen Veränderungen in der Organisationsstruktur und bei der Machtverteilung möglicherweise zu erwarten sind.

4.1.3 Der Unterschied von Prädiktion und Präskription

Die Erfahrung zeigt, dass der beabsichtigte Anwendungsbereich eines ADM-Systems zunächst sehr viel kleiner sein kann als der spätere tatsächliche Einsatzbereich in der Praxis („scope-creep“). Beispielsweise kann ein System, das mit dem Ziel entworfen wurde, bestimmte Ergebnisse zu prognostizieren, theoretisch später dazu genutzt werden, Handlungsanweisungen abzuleiten.

Das ist insofern problematisch, als dabei Korrelationen als Kausalprädikatoren fehlinterpretiert werden, was zu potenziell diskriminierenden Entscheidungen führen kann. Bei der Arbeit lässt sich z. B. häufig eine Korrelation zwischen dem Geschlecht und einem bestimmten Arbeitsfeld feststellen. So ist es nicht unwahrscheinlich, dass der Anteil männlicher Beschäftigter in der IT-Abteilung höher ist als in anderen Abteilungen.

Es verbietet sich jedoch, daraus einen Kausalzusammenhang abzuleiten. Wenn ein ADM-System, das darauf ausgelegt ist, Bewerber*innen auszuwählen, diese Korrelation als Kausalverhältnis interpretiert, bevorzugt es bei der zukünftigen Personalauswahl möglicherweise männliche Bewerber. Dies kann übrigens selbst dann passieren, wenn die Gender-Identifikation selbst nicht als Input-Variable verwendet wurde, sondern stattdessen eine andere, die mit der Gender-Variable zusammenhängt. Beispielsweise können die Namen bestimmter Schulen auf das Geschlecht schließen lassen und auch Hobbys können eine Korrelation zur Gender-Identität aufweisen.

Insofern besteht immer eine hohe Gefahr, wenn prädiktive Modelle präskriptiv verwendet werden. Korrelationen sind möglicherweise ausreichend, um wahrscheinliche Ergebnisse vorauszusagen. Um jedoch Handlungsanleitungen zu geben, benötigt man kausale Modelle (Barocas/Hardt/Narayanan 2022).

Deshalb kommt es entscheidend darauf an, bei der Planung datenbasierter Projekte realistisch zu bleiben und stets im Auge zu behalten, zu welchem Zweck ein bestimmtes ADM-System entworfen wird, was es also leisten soll und was nicht. Auch sollten die jeweiligen Anwendungsgebiete klar definiert und regelmäßig mit Sorgfalt überprüft werden.

4.2 Daten

Sinn und Zweck der Datenerhebungsphase ist es, die Menge an brauchbaren Informationen, mit denen das ML-Modell vor dem Hintergrund der Zielsetzung des jeweiligen ADM-Systems trainiert werden kann, zu maxi-

mieren. Bevor Trainingsdaten verwendet werden können, müssen sie erhoben, vorbearbeitet und ggf. transformiert werden.

Es ist wichtig, das Zustandekommen eines solchen Datenbestands kritisch zu betrachten. Dafür kann die Rücksprache mit Beschäftigten, die die Daten generiert haben, ein entscheidender Schritt sein, denn diese haben die relevante Expertise und wissen um den für die Interpretation benötigten Kontext. Zum Beispiel könnten Arbeitgeber versucht sein, auf bereits bestehende Daten der Beschäftigten wie Mails oder Chatprotokolle zurückzugreifen, um Rückschlüsse auf die Kommunikation und die Leistung der Arbeitnehmer*innen zu ziehen.

Dies wäre jedoch aus vielerlei Gründen prolematisch. Es könnte nicht nur eine Verletzung der Privatsphäre bedeuten, sondern eröffnet auch Möglichkeiten der Manipulation und führt überdies möglicherweise zu verzerrten Ergebnissen, etwa im Hinblick auf Kolleg*innen, die hauptsächlich andere Kommunikationswege nutzen.

Alle Entscheidungen über Datenauswahl, Datenerhebung und Schutz der Privatsphäre sind von hoher Relevanz für die Beschäftigten, für den Grad an Überwachung, den sie am Arbeitsplatz zu erwarten haben werden, und für den Output des ML-Systems. Damit ADM-Systeme positive Auswirkungen auf die Beschäftigten haben, müssen sie in adäquater Weise beaufsichtigt werden. Dazu gehört auch, dass Beschäftigtenvertreter*innen ein Mitspracherecht über die Art und die Qualität der verwendeten Daten bekommen. Datenbestände in Form sogenannter „data sheets“ zu dokumentieren (Gebru et al. 2021), kann die Transparenz und die Sicherheit der Datennutzung verbessern und ist in jedem Fall ratsam.

4.2.1 Zentrale Konstrukte operationalisieren

Im Zuge des Design- und des Datenmanagement-Prozesses muss entschieden werden, welche Konzepte für das jeweilige ADM-System eingesetzt werden sollen. Die Grundlage für diese Entscheidung ist das Ziel des Systems, wie es in der Anfangsphase der ML-Pipeline definiert wurde. Zentral sind auch die Daten, die zur Verfügung stehen oder in Zukunft erhoben werden und in welcher Weise diese verarbeitet werden können. An diesem Punkt sollten auch Beschäftigte bzw. ihre Interessenvertreter*innen reflektieren, wie ihre Ziele operationalisiert werden können – z. B. wenn eine bestimmte Vorstellung von Fairness in einem ADM-System umgesetzt werden soll.

In anderen Fällen sind Beschäftigte vielleicht daran interessiert, dass die Systeme darauf ausgerichtet werden, die Arbeitskultur zu verbessern. In einem solchen Szenario sind die Beschäftigten die eigentlichen Ex-

pert*innen und sollten als direkt Betroffene unbedingt zu Rate gezogen werden.

Sie haben vermutlich bestimmte Vorstellungen davon, wie bestimmte Schlüsselkonstrukte operationalisiert werden sollten, wie man also z. B. die Arbeitsleistung, die Produktivität oder die Eignung für offene Stellen etc. am besten beurteilt und misst. Dabei können sie auf etablierte psychologische Konzepte wie Gewissenhaftigkeit, Offenheit oder Intelligenz zurückgreifen, die über Verhaltenstests oder Fragebögen auf der Grundlage von Peer-Review-Verfahren und Benchmarks operationalisiert worden sind. Pseudowissenschaftliche Konstrukte wie die Phrenologie, die von Gesichtsmerkmalen Rückschlüsse auf bestimmte Eigenschaften der Persönlichkeit zieht und die bereits in verschiedenen KI-Tools in Bewerbungsprozessen eingesetzt wird, sollten inakzeptabel sein.

Beachtenswert ist in diesem Zusammenhang, dass das ML-Modell selbst in Fällen, in denen fragwürdige theoretische Konstrukte nicht explizit aufgestellt und modelliert werden, möglicherweise gleichwohl auf solche Daten und Konstrukte zurückgreift und sie zur Grundlage von Berechnungen macht. Deshalb muss stets nach der Rechtfertigung für die Verwendung bestimmter Datenquellen gefragt werden, etwa von Fotos oder Namen der Beschäftigten oder der Bewerber*innen.

Um noch einmal auf die Vorstellung zurückzukommen, dass es möglich wäre, die Produktivität durch eine Analyse der Nachrichten zu messen, die Beschäftigte schreiben: Die Verwendung solcher Daten würde wahrscheinlich bei Beschäftigten, die viel mobil arbeiten, zu vorteilhafteren Ergebnissen führen, weil sie schlichtweg mehr Daten produzieren als Kolleg*innen, die sich im Büro gegenübersitzen.

Solche unbeabsichtigten Konsequenzen sollten stets vorab in Betracht gezogen werden. Um bereits vorhandene Daten sinnvoll zu interpretieren und zu kontextualisieren, muss man den Gesamtzusammenhang gut kennen. Deshalb ist es immer sinnvoll, Beschäftigte einzubeziehen, um ein theoretisches Verständnis davon zu entwickeln, wie ADM-Systeme bei ihrer Entscheidungsfindung bestimmte Schlüsselkonzepte definieren, operationalisieren und messen.

4.2.2 Datenerhebung

Daten sind die Grundlage der Muster, die ML-Modelle extrahieren, und damit auch Grundlage des Outputs, den ADM-Systeme generieren. Datenerhebung und -verarbeitung sind also unverzichtbar, um solide ADM-Systeme zu bauen. Wenn jedoch historisch bedingte Verzerrungen in die Trainingsdaten einfließen, führt das – wie viele Beispiele zeigen – zur

Verfestigung und Verstärkung solcher Verzerrungen. Wenn sensible Daten, wie beispielsweise zum Geschlecht, zur ethnischen Herkunft oder zum Alter gesammelt werden, können diese Verzerrungen erkannt und reduziert werden.

Allerdings erfordert dies schwierige Abwägungen zwischen dem Ziel, nichtdiskriminierende und faire Entscheidungen zu treffen, und dem Schutz der Privatsphäre bzw. geltendem Datenschutzrecht. Hier zeigt sich wiederum, dass Betroffene in jedem Einzelfall die Möglichkeit haben müssen, ihre Positionen einzubringen.

Auch über die Verhinderung von negativen Auswirkungen hinaus können Daten ein wichtiges Instrument sein, um bestimmte Beschäftigteninteressen voranzubringen. Wenn ein ADM-System z. B. darauf ausgerichtet ist, Fortbildungsangebote vorzuschlagen, sollten vorher Daten über das Interesse der Beschäftigten an bestimmten Themen gesammelt werden, damit das System auch tatsächlich die relevantesten Angebote auswählen kann. Daher sollten Beschäftigtenvertreter*innen stets beteiligt sein, wenn darüber entschieden wird, welche Beschäftigtendaten erhoben werden, sodass die Nutzung eines ADM-Systems in einem für die Betroffenen positiven Sinne gestaltet werden kann.

4.2.3 Datenverarbeitung

Daten, die für Trainingszwecke erhoben werden, sind üblicherweise unstrukturiert, chaotisch und unvollständig. Um sie nutzen zu können, müssen sie bereinigt, verarbeitet und ggf. in komplexere Merkmale umstrukturiert werden. Rohdaten könnten beispielsweise zu granular sein. Wenn ein ADM-System z. B. die Produktivität von Beschäftigten auswerten soll, registriert es möglichweise Tastaturanschläge, eingehende und ausgehende Nachrichten und die Zeitpunkte dieser Ereignisse. Aber die Zeitstempel sind als solche wahrscheinlich nicht sehr informativ, weshalb die Daten aggregiert werden, um nach verschiedenen Abschnitten des Arbeitstags oder der Arbeitswoche differenzieren zu können.

Für die Datenverarbeitung gibt es verschiedene Methoden und Verfahren. Viele basieren auf Entscheidungen, die die Leistung, aber auch die Fairness eines ADM-Systems in hohem Maße beeinflussen und folglich auch die Interessen Beschäftigter berühren. Beispielsweise sind unvollständige Datenbestände nichts Ungewöhnliches. Daten fehlen, weil es Probleme bei der Erhebung gab oder auch, weil sensible Informationen, etwa zur ethnischen Herkunft oder zur Sexualität, absichtlich nicht erfasst wurden oder nicht verwendet werden sollen. Solche fehlenden Daten haben unter Umständen negative Auswirkungen auf ein ADM-System.

Deshalb werden fehlende Werte in der Praxis häufig durch modellierte Daten ersetzt, mit sogenannten Imputationsverfahren. Bei einem solchen Vorgehen kann aber auch eine Verzerrung zulasten von im Datenbestand unterrepräsentierten Gruppen entstehen (Martínez-Plumed 2021). Um bei sensiblen Attributen eine größere Fairness zu gewährleisten, können dem Datenbestand jedoch kontrafaktische Werte hinzugefügt werden. So kann einer Unterrepräsentation spezifischer demografischer Gruppen und einer entsprechenden Verzerrung entgegengewirkt werden. Perfekt funktionieren solche Methoden allerdings nie, und sie korrekt anzuwenden, ist alles andere als trivial.

In dieser Phase setzt ein erfolgreiches Eintreten für die Interessen Beschäftigter also bereits ein tiefgreifendes technisches Verständnis von Verfahren des Maschinellen Lernens und der Datenverwaltung voraus. Interessenvertreter*innen bedürfen hier also der Hilfe externer Expert*innen – und womöglich auch einer Schulung zu den Grundlagen von ML-Methoden, um kritische Stellen und Fragen identifizieren zu können.

4.3 Modelltraining

Der Zweck der Modelltrainingsphase besteht darin, die Regeln, statistischen Muster und Kontingenzen zu extrahieren, um die vorab definierten Zwecke eines ADM-Systems zu optimieren. Das Ergebnis dieses Prozesses, das mit Maschinellem Lernen erstellte Modell, ist eine mathematische Funktion, die mehr oder weniger komplex und intransparent sein kann, je nachdem, welcher Algorithmus verwendet wurde. Wichtige Subkomponenten sind hierbei der Optimierungsalgorithmus, der die Optimierung des Modells steuert, die Zielfunktion, die das Ziel der Optimierung festlegt, und die Metriken, anhand derer Ingenieure sich ein Bild von verschiedenen Aspekten des Optimierungsprozesses machen können.

Das grundsätzliche Problem ist hier, dass nützliche Muster in ML-Systemen mit schädlichen häufig untrennbar verbunden sind. Außerdem ist es sehr kontextabhängig, ob ein Muster schädlich oder nützlich ist. Mit den derzeit zur Verfügung stehenden Mitteln lassen sich die beiden nur unvollständig voneinander trennen. Hinzu kommt, dass man in vielen Fällen Kompromisse zwischen Lesbarkeit und Leistung eingehen wird. Nichtsdestoweniger lohnt es sich auch bei komplexen und intransparenten Modellen, gewisse Ressourcen zu investieren, um Input-Output-Kontingenzen transparenter, besser erklärbar und robuster zu machen.

Hier bieten sich Beschäftigten zwei Möglichkeiten an, ihre Interessen einzubringen. Zum einen, indem sie verhindern, dass schädliche Verzerrungen in das Modell einfließen, und zum anderen, indem sie darauf hin-

wirken, das Modelltraining auf Muster auszurichten, die sie als nützlich betrachten (siehe die Beispiele in Kapitel 4.4.1 bis 4.4.2). Obwohl ML ein weites und schnelllebiges Feld mit einer großen Vielfalt an Systemen ist, können für die meisten Systeme drei Komponenten identifiziert werden, die das spätere Funktionieren bestimmen: Zielfunktion, Optimierungsalgorithmus und Metriken.

4.3.1 Zielfunktion

Die Zielfunktion ist eine mathematische Darstellung der Ziele des Systems. Bei der Entwicklung eines ADM-Systems sollten Beschäftigtenvertreter*innen sicherstellen, dass diese keine schädlichen Verzerrungen verfestigt und perpetuiert. Sie sollten außerdem Interessen der Beschäftigten einbringen, die über Sicherheitsanliegen hinausreichen, also etwa auf eine Optimierung des Systems hinwirken, die persönliche Interessen oder Karriereziele der Betroffenen berücksichtigt.

Um noch einmal das Beispiel eines internen Stellenbesetzungssystems zu verwenden: Hier könnte ein ADM-System z. B. eine semantische Suche durchführen und die für neu zu besetzenden Stellen gewünschten Qualifikationen mit jenen der Beschäftigten abgleichen. In diesem Prozess würden die Qualifikationen der Beschäftigten und die Anforderungen der Stelle in einem gemeinsamen semantischen Raum codiert. Hierüber würden sie vergleichbar gemacht, da im Codierprozess Tokens (also Wörter in den Qualifikationsbeschreibungen der Beschäftigten und in den Stellenausschreibungen) in Zahlen übersetzt werden, die eine bestimmte inhaltliche Bedeutung haben.

Die Übersetzung in einen Zahlenraum ermöglicht es mathematische Operationen, also beispielsweise Distanzen zu errechnen. Wenn in diesem Raum nun nach einem Beschäftigtenprofil mit der geringsten Distanz zum Anforderungsprofil einer offenen Stelle gesucht wird, wird – näherungsweise – auch die semantische Passung zwischen den beiden Dokumenten optimiert und, so die Hoffnung, dann auch die passende Person für die passende Stelle gefunden.

Ein solcher Einbettungsraum wird üblicherweise geschaffen, indem versucht wird, Wörter auf der Grundlage benachbarter Wörter vorauszusagen – die Zielfunktion des Systems wäre dann die Maximierung der Wahrscheinlichkeit eines Zutreffens dieser Voraussage. Das Problem ist jedoch, dass die Voraussage auf Trainingsdaten basiert. Wenn diese Verzerrungen enthalten, lernt das Modell diese Verzerrungen mit. Es gibt aber Möglichkeiten, ihren Einfluss zu begrenzen.

Eine besteht in der Aufnahme eines zusätzlichen Ziels. Das Ziel des Systems könnte also nicht nur sein, die Wahrscheinlichkeit des Zutreffens der Voraussage benachbarter Wörter zu maximieren, sondern möglicherweise auch, den Abstand zwischen Wörtern mit Gendermarkierung zu minimieren und jenen zwischen gendermarkierten und neutralen, normativen Wörtern im Einbettungsraum auszugleichen. Dies würde im Ergebnis Verzerrungen auf Basis von Gender reduzieren (vgl. Bolukbasi et al. 2016; Caton/Haas 2020).

Um die Interessen der Beschäftigten zu unterstützen, könnten diese Ziele, wie z. B. angestrebte Karriereziele, zusätzlich in den Matchingprozess integriert werden. Beispielsweise könnte die semantische Suche nicht nur die Distanz zwischen Stellenanforderung und Qualifikationen, sondern auch zwischen Stellenanforderung und angestrebten Tätigkeitsbereich der Beschäftigten berücksichtigen.

Theoretisch können Systeme auf beliebig viele Ziele hin optimiert werden. In der Praxis werden jedoch stets Kompromisse gemacht. Beschäftigten und ihren Vertreter*innen sollte bewusst sein, dass die Festlegung der Zielfunktion ein entscheidender Punkt für die Einbringung ihrer Interessen sein kann.

4.3.2 Optimierungsalgorithmus

Die Zielfunktion und der Optimierungsalgorithmus hängen eng miteinander zusammen. Während die Zielfunktion der rechnerische Ausdruck für das ist, was erreicht werden soll, definiert der Optimierungsalgorithmus den sogenannten Lösungsraum und den Weg, den ein Modell während der Optimierung sozusagen durch diesen Raum zurücklegt. Optimierungen sind in aller Regel nicht perfekt, und deshalb hat der Optimierungsalgorithmus einen großen Einfluss darauf, an welchem Punkt das Modell den Optimierungsprozess beendet. Dies hat auch Auswirkungen auf den Grad der Robustheit, der Fairness und der Transparenz des Modells.

Größere Modelle, etwa große Sprachmodelle, sind vielleicht leistungsstärker als kleinere, aber auch schwerer zu testen und zu korrigieren. Für manche Zwecke ist zudem nicht nur die Größe des Modells entscheidend, also die Menge der Parameter, sondern auch die Architektur. Wenn z. B. kontrafaktische Werte eingefügt werden, um die Fairness zu verbessern, können sich die Methoden der Wahl insbesondere darin unterscheiden, mit welchem Grad an Genauigkeit sie solche kontrafaktischen Datensätze modellieren und generieren können, insbesondere an den Schnittpunkten verschiedener demografischer Merkmale wie ethnische Zugehörigkeit, Geschlecht und Alter (Creager et al. 2019).

Wenn Beschäftigtenvertreter*innen in diesem Zusammenhang die Unterstützung von ML-Expert*innen in Anspruch nehmen, sollten diese vor allem im Blick haben, wie bestimmte Modelltrainingsmethoden funktionieren und auf welche Zwecke sie ausgerichtet sein können.

4.3.3 Metriken

Um die von den verschiedenen Stakeholdern formulierten Ziele im Blick zu behalten, sollte die experimentelle und iterative Modellierung anhand von Leistungsmetriken überwacht und justiert werden. Dabei können vor allem Aspekte in den Fokus genommen werden, die von der Zielfunktion noch nicht abgedeckt sind. Metriken sollten die gesamte Breite des Problemraums erfassen und den Optimierungsprozess für Interpretationen öffnen (auch für Aspekte, die von den Beschäftigten eingebracht werden), um unerwünschte, verdeckte Nebeneffekte zu minimieren.

Beschäftigtenvertreter*innen sollten bei den Metriken, die verwendet werden, um die Leistung des Systems zu bewerten, ein Mitspracherecht haben. Teilweise ist es aber gar nicht so leicht, es sinnvoll auszuüben, weil man schlicht nicht alle Aspekte des Problemraums berücksichtigen kann.

Was bedeutet es etwa bei einem automatisierten Stellenbesetzungsverfahren, dass ein Beschäftigter „gut auf die Stelle passt“? Kann man einfach den Beschäftigten selbst oder seinen Vorgesetzten fragen? Reicht es schon aus, dass der Beschäftigte nicht gekündigt hat oder nicht gefeuert wurde? Nach welchem Zeitraum kann so etwas sinnvollerweise evaluiert werden? Manche dieser Aspekte sind naturgemäß schwierig zu bewerten, wie etwa die Qualität von Ablehnungen. Wo das System keine Treffer generiert hat, gibt es auch keine Daten, die man analysieren könnte.

Andere wichtige Daten werden hingegen durchaus vorhanden sein. Es lässt sich etwa überprüfen, wer von bestimmten Empfehlungen, wenn sie denn umgesetzt wurden, tatsächlich profitiert hat, oder ob die Ergebnisse in aggregierter Form den eigens aufgestellten Fairness-Anforderungen entsprechen.

Übrigens ist es oft einfach eine Frage von Ressourcen, ob Daten für eine genauere Analyse negativer Ergebnisse zur Verfügung gestellt werden oder nicht. Solche Maßnahmen zur Informationsbeschaffung können von der Durchführung von Absagegesprächen bis zur Einstellung einer Zufallsstichprobe reichen, um experimentelle, weniger voreingenommene Daten zur Validität des Modells zu erhalten (Bird et al. 2016). Entschei-

dend bleibt jedenfalls, dass den Stakeholdern die Grenzen dessen bewusst bleiben, was Metriken leisten können. Der Einblick in den Optimierungsprozess wird nie vollständig sein.

4.4 Einsatz

Der Zweck der Einsatzphase besteht darin, das Ziel eines ADM-Systems zu materialisieren. Ein ADM-System einzusetzen, bedeutet, das mit Maschinellem Lernen trainierte Modell in ein übliches Software-System zu integrieren, das den Input, den Output und die Aufbereitung sowie die Nachbearbeitung der Daten integriert, eine Benutzeroberfläche bereitstellt und auf Basis einer Hardware die Rechenoperationen durchführt. Jetzt erst werden das Modell und das ADM-System mit echten Daten getestet. Nachdem es mit Trainingsdaten optimiert wurde, wird das System nun vermutlich einen Leistungsverlust aufweisen, was die Metriken angeht, die den vorab definierten Problemraum zu erfassen versuchen.

Deshalb sollte das ADM-System jetzt gründlich darauf geprüft werden, ob es die Leistungsanforderungen weiterhin erfüllt. Hier ist also eine Aufsicht essenziell, besonders von Vertreter*innen der Betroffenen. Auch die Benutzeroberfläche und die *user experience* spielen eine große Rolle für die Auswirkungen, die das System auf seine Umgebung hat. Es ist entscheidend, jetzt stets an best practices festzuhalten und die Wirkung eventueller Veränderungen genau zu beobachten.

Wenn das ADM-System bei einem internen Stellenbesetzungsverfahren eingesetzt werden soll, würden jetzt die „Treffer“ (die Fälle, bei denen die Qualifikationen der Beschäftigten und die Anforderungen des Stellenprofils zusammenpassen) an das Management und die Personalabteilung weitergegeben. In unserem Beispiel ist die letzte Entscheidung also eine von Menschen getroffene.

Es ist erwähnenswert, dass dies erneut die Tür für weitere Verzerrungen und Machtmissbräuche öffnet – z. B. wenn nur selektiv auf die vom ADM-System getroffenen Entscheidungen eingegangen wird. Auch hier kann vorgebeugt werden, indem im Vorhinein klare Richtlinien für menschliche Intervention festgeschrieben werden und die Compliance mit solchen Vorgaben überwacht wird. Doch selbst mit Monitoring, fairen Verfahren und einer ML-Pipeline, die sich an best practices orientiert, gibt es keine Garantie dafür, dass das Ergebnis etwa bestimmten Kriterien für einen fairen Output entspricht.

Hieran wird deutlich, wie wichtig es ist, die Ergebnisse, die ein ADM-System produziert, auch nach der Trainings- und Modelloptimierungsphase noch zu überwachen. Dabei sollte spezifiziert werden, was genau

überwacht wird. Im Interesse der Beschäftigten könnte es beispielsweise liegen, ein bestimmtes Niveau an Fairness der Ergebnisse im Hinblick auf Gender, ethnischen Hintergrund und Alter sicherstellen zu wollen. Gruppen mit solchen sensiblen Attributen (die sich auch überschneiden können), müssen unter Umständen in desaggregierter Form überwacht werden.

Wenn der Output des Systems kein akzeptables Fairness-Niveau erreicht, muss es ggf. nachkalibriert werden, sprich die Grenzen bestimmter gruppenspezifischer Entscheidungen müssen in die eine oder andere Richtung verschoben werden. Allerdings kann ein solcher Eingriff auch wieder die Verfahrensfairness beeinträchtigen, woran man erkennt, dass das Design von ADM-Systemen auf Basis hoher ethischer Standards immer auch Kompromisse mit sich bringt (Kleinberg/Mullainathan/Raghavan 2016).

4.5 Retraining

Das Retraining ist eine Art Wartung des ML-Modells. Das System soll auf die ursprünglichen Ziele ausgerichtet bleiben, auch wenn unerwartete und komplexe sozio-technische Dynamiken auftreten. Denn während das ML-Modell gewissermaßen einen Schnappschuss der Trainingsdaten zum Zeitpunkt des Trainings darstellt, ändern sich die Inferenzdaten, gegen die das Modell ausgeführt wird, in Abhängigkeit von Umgebung und Kontext. Das führt zu einer immer größeren Abweichung zwischen den Daten, für die das Modell optimiert wurde, und jenen, denen es neu begegnet. Es kommt zu einer Verschlechterung der Leistung. Üblicherweise wird das Modell dann mit aktuelleren Daten neu trainiert.

Wie schnell das ML-Modell sich verschlechtert, hängt von der Stationarität der Anwendungsdomäne des ADM-Systems ab. Bei Social-Media-Anwendungen wie TikTok oder Instagram können solche Veränderungen binnen Minuten vor sich gehen. Im Kontext von Tätigkeitsfeldern einer Organisation dauert es vielleicht Wochen oder Monate bis sich signifikante Veränderungen angesammelt haben. Rein physikalische Problemdomänen können auch vollständig stationär sein.

Ein Retraining ist zwar die angeratene Methode, einer Verschlechterung der Systemleistung zu begegnen. Es bringt aber neue Herausforderungen, nämlich durch die Interaktion des Systems mit seiner Umgebung, also mit dem organisationellen und sozialen Kontext. Beispielsweise sind die Daten, die bei einem Retraining des ML-geschulten Modells verwendet werden, immer schon von dem Modell selbst beeinflusst. Signale oder Muster, die es für seine Entscheidungsfindung nutzt, finden sich verstärkt

in den Output-Daten und damit auch in der nächsten Trainingsrunde wieder, was potenziell zu Verzerrungen (Schelter/Stoyanowitsch 2020) und Feedback-Schleifen (Barocas/Hardt/Narayanan 2022) führen kann.

Schon beim ersten Wiederholungsdurchgang kann das Schwierigkeiten verursachen, weil Voraussagen und Entscheidungen sich als selbsterfüllende Prophezeiungen erweisen können, wie sich beim sogenannten „predictive policing", also der vorausschauenden Polizeiarbeit, bereits gezeigt hat (Dobbie 2016; Ensign 2017). In zahlreichen Durchgängen kann sich ein solcher Effekt verstärken und auch in weniger sensiblen Bereichen problematisch werden. Bei Empfehlungsalgorithmen wie jenem aus unserem Beispiel kann das bedeuten, dass bestimmte Beschäftigungsgruppen bevorzugt werden.

Wie man dieser Gefahr begegnen kann, ist derzeit noch Gegenstand der Forschung. Es hängt sowohl von der Domäne ab als auch davon, mit welcher Art von Feedback-Effekten man rechnet. Einige Vorschläge für Gegenmaßnahmen gibt es aber schon, etwa dass man die Verschiebungen in den Input- und Output-Daten genau überwachen und dann den Eigeneinfluss des Modells auf diese Daten modellieren sollte, um entsprechende Korrekturen vornehmen zu können (Krauth/Wang/Jorden 2022). Oder dass man das erneute Training nach Möglichkeit nur mit Daten vornehmen sollte, die vom Modell unbeeinflusst sind (einen guten Überblick über praktische Ansätze gibt Huyen 2022).

Aus Beschäftigtenperspekte sollte jedenfalls auch beim Retraining wieder darauf hingewiesen werden, dass Aufsicht, Kontrolle und Präventionsmaßnahmen unerlässlich sind, und zwar auch nachdem ein ADM-System bereits implementiert und in einem bestimmten Organisationskontext eingeführt wurde. Man braucht also einen Aufsichtsprozess, bei dem Beschäftigtenvertreter*innen kontinuierlich in das Monitoring der bereits im Einsatz befindlichen ADM-Systeme eingebunden sind. Und man braucht Verfahren, die wirksam Abhilfe schaffen, wenn Beschäftigte von Entscheidungen mangelhafter Systeme negativ betroffen sind.

Hier wird wiederum deutlich, dass technische Präventionsmaßnahmen auch organisationell abgesichert werden müssen. Und wie wichtig ein Bewusstsein für die Unvollkommenheiten, Risiken und Gefahren von ADM-Systemen ist – selbst wenn Beschäftigteninteressen zunächst in ihr Design eingeflossen sind.

5. Über Risikobegrenzung hinaus: Wissensaufbau und Beteiligung

ADM-Systeme im Personalbereich werden nur dann im Interesse der Beschäftigten arbeiten, wenn Beschäftigtenvertreter*innen von Anfang an bei der Entwicklung, der Implementierung und schließlich der Anwendungsphase dieser Systeme eingebunden werden. Anhand der ML-Pipeline haben wir die verschiedenen Punkte aufgezeigt, an denen wichtige Entscheidungen getroffen werden, bei denen Beschäftigtenvertreter*innen mitsprechen sollten. Dies ist nicht nur entscheidend für arbeitnehmer*innenfreundliche Technologien, sondern auch Voraussetzung für das Vertrauen der Beschäftigten in die Systeme.

Unsere Reflektionen haben deutlich gemacht, wie ethische Grundsätze, beim Design von ADM-Systemen für den Arbeitsplatz, in der Praxis implementiert werden können. Dieser Übergang vom Prinzip zur Praxis ist dringend nötig. Es braucht gute Beispiele, konkrete Leitlinien und Handreichungen für Beschäftigtenvertreter*innen sowie Maßnahmen zum Wissensaufbau in diesem Bereich. Beschäftigtenvertreter*innen müssen keine Expert*innen in Machine Learning werden, sollten sich jedoch ein grundsätzliches Verständnis von ML-Prozessen aneignen, um die richtigen Fragen stellen und im Einzelfall die wichtigsten Defizite eines bestimmten ML-Systems erkennen zu können. Worauf es also ankommt:

- ADM-Systeme mithilfe der ML-Pipeline zu entmystifizieren,
- Interventionspunkte in der ML-Pipeline zu identifizieren, die sich dafür eignen, People-Analytics-Systeme auf die Interessen von Beschäftigten auszurichten,
- den Wissensaufbau von Beschäftigtenvertreter*innen voranzubringen und sie in die Lage zu versetzen, im Workflow der ML-Pipeline die Interessen der Beschäftigten einzubringen sowie einzelfallbezogen auf Defizite von ADM-Systemen aufmerksam zu machen,
- Beschäftigtenvertreter*innen in die Lage zu versetzen, methodische Detailfragen mit externen ML-Expert*innen besprechen zu können.

Somit ergibt sich eine prozessorientierte Perspektive dafür, Beschäftigteninteressen wirkungsvoll einzubringen, und zwar bei der Planung, der Entwicklung und der Implementierung von People-Analytics-Systemen. Diese Perspektive kann aber auch nützlich sein, um ADM-Systeme, die extern eingekauft werden, richtig zu bewerten. Wenn eine gewisse Transparenz gegeben ist (wenn z. B. sogenannte „data sheets“ oder „model cards“ und weitere Dokumentationen vorliegen), kann die Vereinbarkeit eines bestimmten ADM-Systems mit den Interessen der Beschäftigten durchaus beurteilt werden.

Dafür sollte wieder die ML-Pipeline zu Rate gezogen und überlegt werden, an welchen Punkten diese Interessen konkret eingebracht werden sollten. So kann die ML-Pipeline dafür eingesetzt werden, ein People-Analytics-System im Nachhinein zu überprüfen, etwa indem man das Ziel (die Problemdefinition), die Datenverwaltung, das Training und die Rahmenbedingungen eines späteren Retrainings kritisch hinterfragt.

Obwohl der risikobasierte Ansatz der geplanten KI-Verordnung auf EU-Ebene unvermeidbar zu Risikobegrenzungsansätzen führt, hat er doch bei einigen Gewerkschaften und ihnen nahestehenden Verbänden zu Forderungen nach stärker partizipatorischen Governance-Ansätzen für ADM-Systeme am Arbeitsplatz geführt. Dies zeigt, dass ethische Prinzipien für den ADM-Einsatz am Arbeitsplatz nicht ausreichen. Nötig ist ein Schritt darüber hinaus, hin zu einer konkreten Vorstellung davon, wie Beschäftigtenvertreter*innen solche Prinzipien in der Praxis auch operationalisieren können.

Risiken und Potenziale von ADM-Systemen können stets nur von Fall zu Fall beurteilt und nur sehr grundlegend verallgemeinert werden. Beschäftigte können ihre Interessen aber mithilfe der ML-Pipeline durchaus in People-Analytics-Systeme einbringen. Insofern sind die Ex-ante-Logiken der KI-Verordnung und vergleichbarer Risikobegrenzungsansätze für KI-Anwendungen stets nur begrenzt nützlich. Beschäftigteninteressen zu wahren, ist ein Anspruch, der darüber hinausreicht.

Risikobasierte Ansätze taugen letztlich als Sicherheitsnetz, auf einem sehr grundlegenden Niveau, nämlich um ADM-Systeme, deren Einsatz offenkundig schädlich für Beschäftigte wäre, vom Markt fernzuhalten. Aus Beschäftigenperspektive ist das jedoch nicht genug. Vielmehr sollten Beschäftigtenvertreter*innen und andere Stakeholder über Risikobegrenzungen wie in der KI-Verordnung hinausdenken und sich überlegen, wie sie Beschäftigteninteressen in People-Analytics-Systeme so einbringen können, dass Beschäftigte nicht nur vor Risiken geschützt werden, sondern auch von den Möglichkeiten profitieren, die solche Systeme eröffnen.

Es bedarf zudem weiterer nationaler und transnationaler Initiativen, die beispielsweise auch den nationalen Kontext der Beschäftigtenvertretung einbeziehen. Sie sollten einerseits Risiken begrenzen, andererseits aber auch gewährleisten, dass Beschäftigteninteressen in Bezug auf ADM-Systeme am Arbeitsplatz angemessen repräsentiert sind.

Literatur

AlgorithmWatch (2023): Algorithmic Transparency and Accountability in the world of work. A mapping study into the activities of trade unions. www.ituc-csi.org/Algorithmic-transparency-and-accountability (Abruf am 14.8.2023).

Association of Nordic Engineers (2021): Response of the Association of Nordic Engineers to the European Commission's public consultation on the White Paper on Artificial Intelligence – A European approach to excellence and trust. https://nordicengineers.org/wp-content/uploads/2020/10/ane-response-consultation-eu-commission-white-paper-on-ai-2020-final.pdf (Abruf am 14.8.2023).

Barocas, Solon / Hardt, Moritz / Narayanan, Arvind (2022): Fairness and Machine Learning. Limitations and Opportunities. https://fairmlbook.org/pdf/fairmlbook.pdf (Abruf am 14.8.2023).

Bolukbasi, Tolga / Chang, Kai-Wei / Zou, James / Saligrama, Venkatesh / Kalai, Adam (2016): Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. https://arxiv.org/pdf/1607.06520 (Abruf am 14.8.2023).

Bird, Sarah / Barocas, Solon / Crawford, Kate / Diaz, Fernando / Wallach, Hanna (2016): Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. Workshop on Fairness, Accountability, and Transparency in Machine Learning. https://ssrn.com/abstract=2846909 (Abruf am 14.8.2023).

Bommasani, Rishi / Hudson, Drew A. / Adeli, Ehsan / Altman, Russ / Arora, Simran / Arx, Sydney von / Bernstein, Michael S. / Bohg, Jeannette / Bosselut, Antoine / Brunskill, Emma / Brynjolfsson, Erik / Buch, Shyamal / Card, Dallas / Castellon, Rodrigo / Chatterji, Niladri / Chen, Annie / Creel, Kathleen / Davis, Jared Quincy / Demszky, Dora / Donahue, Chris / Doumbouya, Moussa / Durmus, Esin / Ermon, Stefano / Etchemendy, John / Ethayarajh, Kawin / Fei-Fei, Li / Finn, Chelsea / Gale, Trevor / Gillespie, Lauren / Goel, Karan / Goodman, Noah / Grossman, Shelby / Guha, Neel / Hashimoto, Tatsunori / Henderson, Peter / Hewitt, John / Ho, Daniel E. / Hong, Jenny / Hsu, Kyle / Huang, Jing / Icard, Thomas / Jain, Saahil / Jurafsky, Dan / Kalluri, Pratyusha / Karamcheti, Siddharth / Keeling, Geoff / Khani, Fereshte / Khattab, Omar / Koh, Pang Wei / Krass, Mark / Krishna, Ranjay / Kuditipudi, Rohith / Kumar, Ananya / Ladhak, Faisal / Lee, Mina / Lee, Tony / Leskovec, Jure / Levent, Isabelle / Li, Xiang Lisa / Li, Xuechen / Ma, Tengyu / Malik, Ali / Manning, Christopher D. / Mirchandani, Suvir / Mitchell, Eric / Munyikwa, Zanele / Nair, Suraj / Narayan, Avanika / Narayanan, Deepak /

Newman, Ben / Nie, Allen / Niebles, Juan Carlos / Nilforoshan, Hamed / Nyarko, Julian / Ogut, Giray / Orr, Laurel / Papadimitriou, Isabel / Park, Joon Sung / Piech, Chris / Portelance, Eva / Potts, Christopher / Raghunathan, Aditi / Reich, Rob / Ren, Hongyu / Rong, Frieda / Roohani, Yusuf / Ruiz, Camilo / Ryan, Jack / Ré, Christopher / Sadigh, Dorsa / Sagawa, Shiori / Santhanam, Keshav / Shih, Andy / Srinivasan, Krishnan / Tamkin, Alex / Taori, Rohan / Thomas, Armin W. / Tramèr, Florian / Wang, Rose E. / Wang, William / Wu, Bohan / Wu, Jiajun / Wu, Yuhuai / Xie, Sang Michael / Yasunaga, Michihiro / You, Jiaxuan / Zaharia, Matei / Zhang, Michael / Zhang, Tianyi / Zhang, Xikun / Zhang, Yuhui / Zheng, Lucia / Zhou, Kaitlyn / Liang, Percy (2021): On the Opportunities and Risks of Foundation Models. https://arxiv.org/pdf/2108.07258 (Abruf am 14.8.2023).

Campolo, Alexander / Crawford, Kate (2020): Enchanted Determinism: Power without Responsibility in Artificial Intelligence. In: Engaging Science, Technology, and Society 6, S. 1–19. DOI: 10.17351/ests2020.277.

Caton, Simon / Haas, Christian (2020): Fairness in Machine Learning: A Survey. https://arxiv.org/abs/2010.04053 (Abruf am 14.8.2023).

Christl, Wolfie (2021): Digitale Überwachung und Kontrolle am Arbeitsplatz: Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management? Wien. https://crackedlabs.org/daten-arbeitsplatz (Abruf am 14.8.2023).

Creager, Elliot / Madras, David / Jacobsen, Jörn-Henrik / Weis, Marissa A. / Swersky, Kevin / Pitassi, Toniann / Zemel, Richard (2019): Flexibly Fair Representation Learning by Disentanglement. In: Proceedings of the International Conference on Machine Learning (ICML). https://arxiv.org/pdf/1906.02589 (Abruf am 14.8.2023).

DGB (2021): The German Trade Union Confederation's Position on the EU Commission's draft of a European AI Regulation. www.dgb.de/downloadcenter/++co++9341cf1a-5107-11ec-9432-001a4a160123 (Abruf am 14.8.2023).

Dobbie, Will / Goldin, Jacob / Yang, Crystal S. (2018): The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. In: American Economic Review 108, H. 2, S. 201–240. www.aeaweb.org/articles?id=10.1257/aer.20161503 (Abruf am 14.8.2023).

Ensign, Danielle / Friedler, Sorelle A. / Neville, Scott / Scheidegger, Carlos / Venkatasubramanian, Suresh (29.6.2017): Runaway Feedback Loops in Predictive Policing. In: Conference on Fairness, Accountability and Transparency, S. 160–171. https://arxiv.org/abs/1706.09847 (Abruf am 14.8.2023).

Europäische Kommission (2020a): Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen. COM(2020) 65 final. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52020DC0065 (Abruf am 14.8.2023).

Europäische Kommission (2020b): Künstliche Intelligenz – ethische und rechtliche Anforderungen. Konsultation vom 20. Februar 2020 bis 14. Juni 2020. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/public-consultation_de (Abruf am 14.8.2023).

Europäische Kommission (2021a): Anhänge des Vorschlags für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. COM(2021) 206 final. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_2&format=PDF (Abruf am 14.8.2023).

Europäische Kommission (2021b): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. COM(2021) 206 final. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF (Abruf am 14.8.2023).

Fernando, Martínez-Plumed / Cèsar, Ferri / David, Nieves / José, Hernández-Orallo (2021): Missing the missing values: The ugly duckling of fairness in machine learning. In: International Journal of Intelligent Systems 36, H. 7, S. 3217–3258. DOI: 10.1002/int.22415.

Gebru, Timnit / Morgenstern, Jamie / Vecchione, Briana / Vaughan, Jennifer Wortman / Wallach, Hanna / Daumé, Hal, III / Crawford, Kate (2021): Datasheets for Datasets. https://arxiv.org/pdf/1803.09010 (Abruf am 14.8.2023).

Gießler, Sebastian (2021): Was ist automatisiertes Personalmanagement? https://algorithmwatch.org/de/wp-content/uploads/2021/05/Was-ist-automatisiertes-Personalmanagement-Giesler-AlgorithmWatch-2021.pdf (Abruf am 14.8.2023).

Holubová, Barbora (2022). Algorithmic management: Awareness, risks and response of the social partners (Final report). Brüssel: Friedrich-Ebert-Stiftung – Competence Centre on the Future of Work. https://library.fes.de/pdf-files/bueros/bruessel/19524.pdf (Abruf am 14.8.2023).

Huyen, Chip (2022): Designing machine learning systems. An iterative process for production-ready applications. Beijing/Boston/Farnham/Sebastopol/Tokyo: O'Reilly.

IndustriALL (2021): Artificial Intelligence: Humans must stay in command. Policy Brief 2019-01. https://news.industriall-europe.eu/documents/upload/2019/2/636849754506900075_Policy%20Brief%20-%20Artificial%20Intelligence.pdf (Abruf am 14.8.2023).

Jarrahi, Mohammad Hossein / Newlands, Gemma / Lee, Min Kyung / Wolf, Christine T. / Kinder, Eliscia / Sutherland, Will (2021): Algorithmic management in a work context. In: Big Data & Society 8, H. 2. DOI: 10.1177/20539517211020332.

Kleinberg, Jon / Mullainathan, Sendhil / Raghavan, Manish (2016): Inherent Trade-Offs in the Fair Determination of Risk Scores. https://arxiv.org/pdf/1609.05807 (Abruf am 14.8.2023).

Krauth, Karl / Wang, Yixin / Jordan, Michael I. (2022): Breaking Feedback Loops in Recommender Systems with Causal Inference. https://arxiv.org/abs/2207.01616 (Abruf am 14.8.2023).

Krzywdzinski, Martin / Pfeiffer, Sabine / Evers, Maren / Gerber, Christine (2022): Measuring Work and Workers. Wearables and Digital Assistance Systems in Manufacturing and Logistics. Wissenschaftszentrum Berlin für Sozialforschung – Social Science Research Center Berlin, WZB. Berlin. https://bibliothek.wzb.eu/pdf/2022/iii22-301.pdf (Abruf am 14.8.2023).

LO Sweden (2021): Remiss av Europeiska kommissionens förslag till förordning om harmoniserade regler för artificiell intelligens. www.lo.se/home/lo/res.nsf/vRes/lo_fakta_1366027472949_remiss_eu_k_harmoniserande_regler_artificiell_intelligens_pdf/$File/remiss_EU-K_harmoniserande_regler_artificiell_intelligens.pdf (Abruf am 14.8.2023).

Mitchell, Margaret / Wu, Simone / Zaldivar, Andrew / Barnes, Parker / Vasserman, Lucy / Hutchinson, Ben / Spitzer, Elena / Raji, Inioluwa Deborah / Gebru, Timnit: Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 29–31, S. 220–229. https://arxiv.org/abs/1810.03993 (Abruf am 14.8.2023).

Ponce Del Castillo, Aída (2021): The AI Regulation: entering an AI regulatory winter? Why an ad hoc directive on AI in employment is required. ETUI Policy Brief, Brussels. www.etui.org/publications/ai-regulation-entering-ai-regulatory-winter (Abruf am 14.8.2023).

Schelter, Sebastian / Stoyanovich, Julia (2020): Taming Technical Bias in Machine Learning Pipelines. In: IEEE Data Engineering Bulletin. https://ssc.io/pdf/taming-technical-bias.pdf (Abruf am 14.8.2023).

Syndicom (2020): 9 KI-Leitprinzipien für eine menschenfreundliche Zukunft. Wie die Chancen genutzt und die Risiken gebannt werden können. https://syndicom.ch/unserethemen/dossiers/kuenstlicheintelligenzki/herausforderungen/ (Abruf am 14.8.2023).

TUC – Trade Union Congress (2022): People-Powered Technology: Collective Agreements and Digital Management Systems. www.tuc.org.uk/sites/default/files/2022-08/People-Powered_Technology_2022_Report_AW.pdf (Abruf am 14.8.2023).

UNI Europa (2019): UNI Europa response to the European Commission consultation on AI ethics guidelines. www.uni-europa.org/news/uni-europa-response-to-the-european-commission-consultation-on-ai-ethics-guidelines/ (Abruf am 14.8.2023).

Autor*innen

Lukas Hondrich war wissenschaftlicher Mitarbeiter bei AlgorithmWatch und arbeitete zu Arbeitsrechten und KI-Regulierung in der EU. Er untersuchte Mitbestimmungsmöglichkeiten von Beschäftigten und Gewerkschaften bei datenbasierten soziotechnischen Systemen. Zuvor entwickelte er maschinelle Lernsysteme in der E-Commerce- und Tech-Industrie. Er hat einen Master in Cognitive-Affective Neuroscience der TU Dresden und einen Bachelor in Psychologie der Johannes-Gutenberg-Universität Mainz.

Dr. Anne Mollen ist Senior Research Associate bei AlgorithmWatch sowie Projektmanagerin für „SustAIn – Der Nachhaltigkeitsindex für Künstliche Intelligenz“. Sie arbeitet auch als Medien- und Kommunikationswissenschaftlerin an der Universität Münster und untersucht dort die Beziehung zwischen digitalen Medien, Gesellschaft und Demokratie. Der Schwerpunkt ihrer Arbeit liegt auf automatisierten Entscheidungssystemen in der Arbeitswelt und bei Online-Plattformen sowie auf Nachhaltigkeitsfragen im Zusammenhang mit Künstlicher Intelligenz.

ISSN 2509-2359