Bernd Irlenbusch ist Informatiker, Volkswirt und Kaufmann. Der Professor für Unternehmensentwicklung und Wirtschaftsethik ist Principal Investigator des Exzellenzclusters Econtribute, Sprecher des Centers for Social and Economic Behavior an der Universität zu Köln und Mitglied im Ethikbeirat HR Tech.

KI und Ethik: Was auf HR zukommt

Künstliche Intelligenz bietet dem Personalbereich viele Erleichterungen, bringt aber auch zahlreiche Gefahren mit sich. Welche neuen Verantwortungen auf HR zukommen, welche Handlungsempfehlungen sich dadurch ergeben und wie der Ethikbeirat HR Tech unterstützt, erläutert Professor Bernd Irlenbusch.

Interview Daniela Furkel

Personalmagazin: KI findet immer öfter Einsatz in HR-Prozessen. Welche ethischen Herausforderungen ergeben sich daraus?

Bernd Irlenbusch: Wir unterscheiden fünf Bereiche, in denen sich ethische Herausforderungen aus dem Einsatz von KI ergeben, wobei sich diese überlappen. Erstens Diskriminierung und Fairness: Wie kann sichergestellt werden, dass die durch KI-Anwendungen betroffenen Menschen nicht nach ungewollten Kriterien diskriminiert und unfair behandelt werden? Zweitens Transparenz und Erklärbarkeit: Die von einer KI getroffenen Entscheidungen müssen nachvollziehbar sein und auch erklärt werden können. Drittens Verantwortlichkeit und Kontrolle: Wer ist für die von einer KI getroffenen Entscheidungen verantwortlich? Können diese überprüft und revidiert werden? Viertens Datenschutz und Sicherheit: Wie können private Daten geschützt werden – auch vor böswilligen Angriffen? Fünftens Verlässlichkeit und Schutz: Wie kann sichergestellt werden, dass die KI das tut, was sie soll, und dass keine Schäden von ihr ausgehen?

Ihr vierter Punkt war Datenschutz und Sicherheit. Welche Fragen stehen da im Fokus?

In diesem Kontext gibt es zahlreiche Fragen, die zum Teil auch von der jeweiligen Anwendung abhängen. Tendenziell arbeiten KI-Anwendungen umso präziser, je mehr Daten zur Verfügung stehen. Ist es daher ethisch erlaubt, möglichst viele Daten über Personen zu erheben und diese auch zusammenzuführen, um die Vorhersagekraft der KI zu erhöhen? Betrachten wir zum Beispiel ein Einstellungsverfahren. Hier ist es naheliegend, Daten aus verschiedenen Quellen, wie Lebensläufen, Interviews und sozialen Netzwerken, zu sammeln und zusammenzuführen. Ist sichergestellt, dass die Bewerberinnen oder Bewerber der Erhebung der einzelnen Daten zugestimmt haben? Haben sie auch der Zusammenführung der Daten zugestimmt, durch die ein ganz neuer Blick auf die jeweilige Person geworfen werden kann?

Haben Sie noch ein Beispiel?

Ein anderes Beispiel ist die Erhebung von Leistungskennzahlen und Arbeitsergebnissen. Auch hier muss sichergestellt sein, dass die Mitarbeitenden zustimmen – auch der Zusammenführung dieser Daten mit anderen persönlichen Daten. Hier stellt sich die Frage: Wo sind die Grenzen zur Überwachung? Schwierige Datenschutzprobleme ergeben sich auch immer, wenn personenbezogene Daten verwendet werden, um die KI zu trainieren und zu verbessern. Letzteres kann insbesondere dann problematisch sein, wenn die KI von einem externen Anbieter stammt, der sich den Zugriff auf Eingabedaten offenhält, um sein KI-Produkt zu optimieren. Wird dies vom Anbieter zur Bedingung gemacht, ist Vorsicht geboten.

Sie sprachen auch das Thema Diskriminierung an. Was ist hierbei die größte Herausforderung?

Mit Blick auf Diskriminierung muss man verstehen, dass KI-Systeme mit Daten trainiert werden, welche aus von Menschen getroffenen Entscheidungen resultieren. Diese Daten enthalten im Allgemeinen menschliche Verzerrungen und Vorurteile bezüglich bestimmter Merkmale. Diese werden von der KI übernommen. Falls in den Trainingsdaten zum Beispiel eher männlich klingende Vornamen bevorzugt werden oder wenige Datenpunkte mit weiblich klingenden Vornamen enthalten sind, benachteiligt die KI in Bewerbungsprozessen möglicherweise Lebensläufe mit weiblich klingenden Vornamen, wie es in einem bekannten Fall bei Amazon vor mehr als sechs Jahren bekannt wurde. Es ist also nicht garantiert, dass die KI vorurteilsfreier entscheidet als der Mensch. Auch das Ausblenden der Merkmale vor der KI hilft im Allgemeinen nicht weiter, da bestimmte Merkmale häufig über andere Informationen von der KI erschlossen werden können, etwa über Hobbys, die eher von Frauen ausgeübt werden.

Was kann der HR-Bereich dagegen tun?

Solche Verzerrungen können nur durch nachträgliche umfangreiche Tests, sogenannte Audits, aufgedeckt werden – insbesondere dann, wenn die Trainingsdaten von der Entwicklungsfirma der KI nicht offengelegt werden. Dass dann Korrekturen möglich sind, kann man gleichzeitig aber auch als Stärke der KI ansehen, denn anders als beim Menschen, sind Verzerrungen oder Diskriminierungen mithilfe von Tausenden von Abfragen im Prinzip aufdeckbar. Dies muss dann allerdings auch geleistet werden.

Wozu würden Sie raten: Wie können Unternehmen sicherstellen, dass KI-gestützte HR-Systeme fair und transparent sind?

KI-Anwendungen müssen durch kontinuierliche Audits – gegebenenfalls durch zertifizierte Dritte – auf ihre Anfälligkeit zur Diskriminierung in unterschiedlichen Dimensionen getestet werden. Beim Thema Fairness ist es zunächst wichtig, sich im Unternehmen über die anzuwendenden Fairnessnormen im Klaren zu werden und festzulegen, welche Fairnessnorm bei einer KI-Anwendung für welche Art von Entscheidung zum Einsatz kommen soll. Zum Beispiel: Soll bei Neueinstellungen rein die Qualifikation zählen oder soll der Anteil von Männern und Frauen gewichtet werden? Über diese ergebnisorientierten Fairnessnormen hinaus spielt aber auch prozedurale Fairness, wie Transparenz, eine entscheidende Rolle. Transparenz darüber, welche Daten gesammelt werden, oder Transparenz darüber, dass jemandem gerade von einer KI geholfen wird, verstehen sich aus ethischer Sicht von selbst. Nehmen Sie zum Beispiel durch eine KI individuell gestaltete Stellenanzeigen. In sozialen Netzwerken ist es mittlerweile üblich, Stellenanzeigen zu versenden, die unter gezielter Nutzung von frei zugängigen Daten über Zielpersonen von einer KI gestaltet werden. Bei den Zielpersonen soll der Eindruck entstehen, dass das Unternehmen sich besonders mit ihnen beschäftigt hat und gerade sie gewinnen will. In der Konsequenz fühlen sich die Zielpersonen individuell angesprochen. In diesen Fällen scheint es mir ein ethisches Gebot der Transparenz, dass die Verwendung einer KI kenntlich gemacht wird.

Was ist mit der berühmten „Black Box“ bei KI-Anwendungen? Wie kann da mehr Transparenz hergestellt werden?

Unter dem Begriff KI verbergen sich viele Algorithmen mit unterschiedlichen Komplexitätsgraden. „White-Box“-Modelle umfassen nur wenige einfache Regeln, die zum Beispiel als Entscheidungsbaum oder einfaches lineares Modell mit wenigen Parametern dargestellt werden. Deshalb können die Prozesse hinter diesen Algorithmen normalerweise von Menschen verstanden werden. Im Gegensatz dazu verwenden sogenannte „Black-Box“-Modelle zum Teil Tausende von Entscheidungsbäumen oder unüberschaubar viele Parameter. Sie sind in ihren Entscheidungen und Vorhersagen mitunter deutlich präziser als „White-Box“-Modelle. Allerdings können ihre Ergebnisse viel schwieriger von Menschen nachvollzogen werden. Es ergibt sich also ein Präzisions-/Transparenz-Tradeoff. Bei jeder Anwendung in HR sollte geprüft werden, ob der Einsatz von „Black-Box“-Modellen mit ihrer hohen Intransparenz wirklich deutlich bessere Entscheidungen bringt und somit gerechtfertigt ist. Teilweise ist der Unterschied zu „White-Box“-Modellen nicht groß, sodass mit deren Einsatz eine gute und transparentere Entscheidung gewährleistet ist.

Und wenn doch „Black-Box“-Modelle zum Einsatz kommen?

Dann ist es meist schwierig, die Entscheidungen nachvollziehbar zu machen. Sollte eine solche KI zum Beispiel bei einer Beförderungsentscheidung eingesetzt werden, so wird es einer Kollegin oder einem Kollegen, die oder der nicht befördert wurde, nicht reichen, zu hören, dass die KI so entschieden hat. Vielmehr möchte sie oder er wissen, welche Kriterien ausschlaggebend waren und warum sie nicht als erfüllt angesehen wurden. Neuere Forschungsansätze versuchen, auch für Entscheidungen von „Black-Box“-Modellen Erklärungen zu liefern, die Menschen nachvollziehen können. Diese Ansätze werden unter dem Begriff „Explainable

AI“ zusammengefasst. Dann können zum Beispiel für einzelne Kompetenz-Dimensionen Aussagen getroffen werden, wie: „Wenn Sie auf der Kompetenzdimension X einen Wert von Y erreicht hätten, wären Sie befördert worden.“ Andere Ansätze setzen eine zweite KI ein, um die Entscheidungen einer anderen KI für den Menschen verständlich zu machen. Die Ansätze der „Explainable AI“ befinden sich noch in den Kinderschuhen und es gibt begründete Zweifel daran, dass es je gelingen wird, Entscheidungen der „Black-Box“-Modelle für Menschen verständlich darzustellen, da die hohe Zahl der Parameter geistig nicht fassbar ist.

Welche Rolle spielen menschliche Entscheidungsträger in einer von KI unterstützten HR-Umgebung?

Das hängt stark davon ab, wofür welche KI im HR zum Einsatz kommt. Sollen von der KI Entscheidungen getroffen werden, die die Lebensverläufe von Bewerbenden oder Mitarbeitenden signifikant beeinflussen, sollte die KI aus meiner Sicht ausschließlich als Ratgeber verwendet werden. Das heißt, menschliche Entscheidungsträger haben das letzte Wort. Entscheidungen einer KI sollten im HR-Bereich von einer oder mehreren Personen kontrolliert werden, die auch korrigierend eingreifen können. Zum jetzigen Zeitpunkt ist es aus meiner Sicht aus ethischer Perspektive unmöglich, dass eine KI die Verantwortung für eine Entscheidung übernimmt. Dies müssen menschliche Entscheidungsträger sein. Die Verantwortung komplett auf die Entwickler einer KI zu übertragen, scheint mir zum jetzigen Zeitpunkt auch fragwürdig.

In diesem Zusammenhang sollte auch darauf hingewiesen werden, dass eine KI immer eine vergangenheitsorientierte Entscheidung trifft, da sie auf Daten trainiert ist, die bereits existieren. Aktuelle oder zukünftige Entwicklungen können beim jetzigen Stand der Technik wahrscheinlich nur unzureichend von der KI berücksichtigt werden. So erfordert eine neue Unternehmensumgebung, mit der die KI keine Erfahrung hat, möglicherweise neue Kompetenzen der Mitarbeitenden, über deren Notwendigkeit die KI mangels Daten keine Aussage treffen kann.

KI bringt aber nicht nur Risiken, sondern auch Chancen mit sich. Welche Vorteile bieten KI und automatisierte HR-Tools für die Mitarbeitererfahrung?

Aus meiner Sicht lassen sich die Chancen des Einsatzes von KI grob in zwei Kategorien einordnen: Bessere Entscheidungen sowie Kostenreduktion beziehungsweise Zeitersparnis. In Zeiten von Fachkräftemangel sind Unternehmen immer mehr darauf angewiesen, dass ihre Mitarbeitenden möglichst passgenau bei denjenigen Aufgaben zum Einsatz kommen, für die sie motiviert sind, die ihnen liegen, für die sie gut geeignet und ausgebildet sind. Hier sehe ich zunächst die größten Chancen für den Einsatz von KI. Es ist zu erwarten, dass wir durch KI-Unterstützung diese Zuordnung stark optimieren können. Darüber hinaus kann KI fehlende Kompetenzen bei einzelnen Personen frühzeitig identifizieren und HR kann helfen, diese aufzubauen.

Das heißt, auch für die Beschäftigten gibt es positive Folgen?

Ja. Wenn die Systeme gut funktionieren, sollten sie somit auch die „Employee Experience“ verbessern, denn es können den jeweiligen Mitarbeitenden in kürzerer Zeit passgenauere und individuell zugeschnittene Job- und Qualifizierungsangebote gemacht werden. In diesem Zusammenhang gibt es bereits erste KI-Ansätze für „Frühwarnsysteme“ im Bereich Mitarbeiterbindung. Die KI identifiziert Personen, bei denen es eine höhere Wahrscheinlichkeit gibt, dass sie das Unternehmen zu verlassen drohen – sei es, weil sie eine inhaltlich neue Herausforderung suchen, bessere Gehaltsangebote von außen bekommen können oder den nächsten Karriereschritt machen möchten. Hier kann HR dann gezielt individuell auf die jeweiligen Personen zugehen und gemeinsam Vorschläge erarbeiten, die den Verbleib im Unternehmen ermöglichen. KI-Systeme können auch als digitale Assistenten fungieren, zum Beispiel beim Onboarding.

Und die negativen Folgen?

Entfremdung kann ein Problem werden. Wichtig sind hier aus meiner Sicht zwei Dinge: erstens das Gespräch zwischen den Mitarbeitenden und der persönliche Kontakt, auch zum HR-Bereich. Zweitens ist es wichtig, dass die Mitarbeitenden Vertrauen zu den eingesetzten KI-Systemen bekommen. Um das zu erreichen, müssen die eingesetzten KI-Systeme vertrauenswürdig sein. Nichts ist schlimmer als die Erfahrung, dass die KI schlecht oder sogar falsch berät oder nicht funktioniert. Des Weiteren müssen die eingesetzten KI-Systeme die bereits genannten ethischen Mindestanforderungen erfüllen. Dem Unternehmen kommt auch eine besondere Aufklärungspflicht zu, das heißt, die Mitarbeitenden müssen besser verstehen lernen, wie die eingesetzte KI arbeitet und was von ihr erwartet werden kann. Aus meiner Sicht ist dies entscheidend für Vertrauen zur KI.

Welche Maßnahmen sollten Unternehmen ergreifen, um sicherzustellen, dass KI-gestützte HR-Systeme die Privatsphäre der Mitarbeitenden respektieren und ihre Daten angemessen schützen?

Zunächst sollte bei jeder Verarbeitung personenbezogener Daten geklärt werden, zu welchem Zweck die Daten erhoben werden und ob für diesen Zweck nicht weniger Daten ausreichen. Wenn Prozesse der Datenverarbeitung aufgesetzt werden, ist die Einbindung von Mitarbeitervertretungen und Datenschutzbeauftragten sowie die Einhaltung der DSGVO unerlässlich. Die vereinbarten Bestimmungen sollten in einer verständlichen Datenschutzrichtlinie festgehalten werden. Hier sind die folgenden Fragen zu beantworten: Wer hat das Recht, welche personenbezogenen Daten zu erheben und zu erfassen? Wer darf die Daten in welcher Weise anpassen, verändern oder verknüpfen? Wer darf welche Daten abfragen oder verwenden? Wer hat die Pflicht, welche Daten zu welchem Zeitpunkt zu löschen oder zu vernichten? Zusätzlich müssen geeignete Schutzmaßnahmen die Einhaltung der spezifizierten Rechte und Pflichten ermöglichen und sicherstellen, insbesondere durch das Protokollieren der einzelnen Zugriffe. Es sind sichere Speicherorte zu garantieren, ob intern oder extern, die vor Verlust der Daten und so weit wie möglich vor Angriffen schützen. Für besonders sensible Daten erscheint mir auch eine Verschlüsselung personenbezogener Daten im HR-Bereich zwingend.

Gibt es für die Datenspeicherung eine Empfehlung?

Für Datenbanken personenbezogener Daten, die zum Zweck des Trainierens einer KI gehalten werden, entstehen gerade neuere vielversprechende Ansätze, wie der Standard des „differential privacy“. Die Daten einer Person werden dabei stochastisch so abgeändert, dass die KI zwar noch aus den Daten lernen kann, eine Rückverfolgung auf die einzelne Person aber quasi nicht mehr möglich ist. Unter gewissen Voraussetzungen gilt dies sogar dann,

„Die Auswahl der richtigen KI-Anwendung ist durch den EU AI Act noch wichtiger geworden.“

wenn mehrere Datensätze derselben Person zusammengeführt werden. Meiner Meinung nach sollte dieser Ansatz bei personenbezogenen Daten im HR systematisch zum Einsatz kommen.

Sie sind Mitglied des Ethikbeirats HR Tech. Welche Rolle nimmt dieser bei der Regulierung von KI in HR-Systemen ein?

Der Ethikbeirat HR Tech möchte Hilfestellungen für HR und Mitarbeitende geben, um eine ethisch fundierte und verantwortungsvolle Anwendung von neuen Technologien im Unternehmen zu ermöglichen. Die Hilfestellungen werden zum einen in zehn Richtlinien gegeben, die in einem umfangreichen Diskurs von Theorie und Praxis entwickelt wurden und in Zukunft weiterentwickelt werden. Die Themen der Richtlinien sind: transparenter Zielsetzungsprozess, fundierte Lösungen, Menschen entscheiden, notwendiger Sachverstand, Haftung und Verantwortung, Zweckbindung und Datenminimierung, Informationspflicht, Achten der Subjektqualität, Datenqualität und Diskriminierung, stetige Überprüfung. Aufbauend auf diesen Richtlinien hat der Ethikbeirat HR Tech zu Beginn dieses Jahres den „Ethik Check KI“ veröffentlicht – ein frei zugängliches und kostenloses Tool, mit dem HR eine verwendete KI auf die Vereinbarkeit mit ethischen Grundsätzen prüfen kann.

Inwiefern sind diese zehn Richtlinien mit dem EU AI Act vereinbar?

Die Richtlinien des Beirats sowie der Ethik Check KI basieren bislang auf freiwilligen Empfehlungen und nehmen bereits viele Vorschriften des EU AI Act vorweg. Inwieweit die Empfehlungen des Beirats Überschneidungen mit den zukünftig bindenden Vorschriften des EU AI Act haben, dahinter zurückbleiben oder sogar darüber hinausgehen, muss aus meiner Sicht bis zur Rechtswirksamkeit des EU AI Act in einem nächsten Schritt durchdacht und geklärt werden.

Wie hat der EU AI Act die Diskussion über den Einsatz von KI in HR-Systemen verändert?

Aus meiner Sicht ist die entscheidende Neuerung für HR, dass der EU AI Act risikobasierte Regelungen für KI-Anwendungen vorsieht, das heißt die Einteilung in Risikoklassen. Für jede Klasse werden Anforderungen an Entwicklung und Anwendung gestellt. KI-Anwendungen, die in HR zum Einsatz kommen, werden zum Teil als sehr risikoreich eingestuft. Hierzu zählen Einstellungs-, Beförderungs- oder Kündigungsentscheidungen, die Zuweisung von Aufgaben auf der Grundlage von individuellem Verhalten oder persönlichen Eigenschaften sowie die Überwachung und Bewertung von Leistung und Verhalten. Als Begründungen führt der EU AI Act an, dass Berufsaussichten und Lebensunterhalt von Personen sowie Grundrechte auf Datenschutz und Privatsphäre stark beeinflusst oder existierende Diskriminierungen wie oben beschrieben fortgesetzt werden.

Inwiefern fördert der EU AI Act den verantwortungsvollen Einsatz von KI in HR-Systemen?

Er erlegt den Anbietern und Anwendern umfassende Pflichten auf, wenn die KI als hochriskant klassifiziert wird. Anbieter müssen hohe Anforderungen insbesondere an Transparenz, Datenqualität und Zuverlässigkeit garantieren. Anwendende Unternehmen haben Pflichten bezüglich Aufklärung gegenüber den Beschäftigten über die eingesetzten KI-Anwendungen und regelmäßige Audits. Beide Gruppen von Pflichten sind relevant, wenn eigene KI-Anwendungen entwickelt werden. Es bleibt abzuwarten, wie genau die Pflichten in nationales Recht umgesetzt werden und wie sie überprüft werden. Momentan kann ich mir nicht vorstellen, dass reine Selbstverpflichtungen genügen werden. Ob eine eigens dafür eingerichtete staatliche Behörde oder aber Organisationen nach dem Vorbild eines TÜVs für KI-Anwendungen zur Überprüfung geschaffen werden, bleibt abzuwarten.

Welche Handlungsempfehlungen leiten Sie daraus ab?

Die Auswahl der richtigen KI-Anwendung ist durch den EU AI Act noch wichtiger geworden. Neben altbekannten Forderungen nach Datensicherheit und Compliance, Übereinstimmung mit den Werten und Bedürfnissen des anwendenden Unternehmens sowie Flexibilität und Benutzerfreundlichkeit sollten nur solche Anwendungen in Betracht gezogen werden, die die Einhaltung der Pflichten für Anbieter umfassend und auch für Laien verständlich dokumentieren. Besonderes Augenmerk würde ich darauf legen, dass Anbieter regelmäßige Audits in Bezug auf Diskriminierung und Zuverlässigkeit durchführen. Ein weiterer wichtiger Aspekt scheint mir zu sein, dass ausgeschlossen wird, dass Anbieter durch ihre Produkte uneingeschränkten Zugriff auf die Daten des Unternehmens bekommen. Klar ist aber auch, dass die Einhaltung der Pflichten für Anwender solch spezifisches Fachwissen erfordert, dass wahrscheinlich nur große Unternehmen in der Lage sein werden, eigene Spezialistinnen und Spezialisten dafür einzustellen, während kleinere Unternehmen dieses Wissen extern einkaufen müssen.