NOVARA
CONSULTING GROUP

Dieses Dokument wurde maschinell aus dem Englischen übersetzt und kann Fehler enthalten. Das englische Original finden Sie unter https://www.novaracg.com/2025/05/23/four-numbers-and-no-counterfactual-reading-the-artificial-intelligence-case-study-in-human-resources/.

Vier Zahlen und kein Kontrafaktum: Die Fallstudie zur Künstlichen Intelligenz im Personalwesen lesen

Heather Grizzle · 23. Mai 2025

Vier prozentuale Verbesserungen und kein Kontrafaktum. Was jede gängige KI-Kennzahl im Personalwesen voraussetzt, damit sie zutrifft, und die Fragen, die zehn Minuten dauern.

Inhalt

1. Das Genre
2. Behauptung eins: Verkürzung der Besetzungsdauer
3. Behauptung zwei: Reduzierung manueller Verwaltungstätigkeiten
4. Behauptung drei: Steigerung der Bindung von High Potentials
5. Behauptung vier: Bias-Minderung
6. Was allen vieren fehlt
7. Warum der Anspruchsteller es meist selbst glaubt
8. Was ein Käufer fragen sollte
9. Was eine glaubwürdige Fallstudie enthalten müsste
10. Fazit
11. Quellen

Die Fallstudie zur Künstlichen Intelligenz im Personalwesen hat sich zu einer festen Form verfestigt.

Das Genre

Die Fallstudie zur Künstlichen Intelligenz im Personalwesen hat sich zu einer festen Form verfestigt. Eine Organisation beschreibt eine Einführung, berichtet drei oder

vier prozentuale Verbesserungen, fügt eine Behauptung zur Bias-Minderung hinzu und schließt daraus, dass Talent zu einer Quelle des Wettbewerbsvorteils geworden sei. Die Zahlen sind konkret. Die Methodik fehlt. Die Schlussfolgerung lädt zu einem Gespräch ein.

Diese Form hält sich, weil sie als Überzeugungsmittel funktioniert und weil niemand in der Position ist, sie zu überprüfen. Anbieter veröffentlichen sie, weil sie sich verkauft. Praktiker veröffentlichen sie, weil sie damit Ansehen erwerben. Käufer akzeptieren sie, weil die Alternative bedeuten würde zuzugeben, dass sie keine Grundlage für die Entscheidung haben, die sie gerade treffen wollen. Jeder in dieser Kette hat ein Interesse an der Zahl, und niemand hat ein Interesse am Nenner.

Jeder in dieser Kette hat ein Interesse an der Zahl, und niemand hat ein Interesse am Nenner.

Die Position von Novara Consulting Group ist, dass diese Dokumente als Werbeaussagen und nicht als Beleg gelesen werden sollten, dass gerade die am häufigsten zitierten Kennzahlen jene sind, die sich am leichtesten ohne jede zugrunde liegende Verbesserung erzeugen lassen, und dass ein Käufer, der vier Fragen zu stellen vermag, in rund zehn Minuten ein echtes Ergebnis von einem lediglich berichteten unterscheiden kann. Es folgt diese Methode, angewandt auf die vier am häufigsten auftretenden Behauptungen.

Behauptung eins: Verkürzung der Besetzungsdauer

Die Besetzungsdauer ist die am häufigsten berichtete Kennzahl und diejenige, die sich am leichtesten verändern lässt, ohne dass sich am Einstellungsprozess selbst etwas ändert.

Die Messung erfordert einen festgelegten Startpunkt und einen festgelegten Endpunkt, die über die Vergleichszeiträume hinweg konstant gehalten werden. In der Praxis variiert der Startpunkt je nach Organisation und oft sogar je nach Stellenausschreibung: Er kann das Datum sein, an dem eine Stellenanforderung genehmigt wurde, das Datum der Ausschreibung, der Beginn der Kandidatensuche oder der Zeitpunkt, an dem die einstellende Führungskraft den Bedarf erstmals angemeldet hat. Verschiebt man diesen Startpunkt nach hinten, verkürzt sich das gemessene Intervall, ohne dass sich für irgendjemanden tatsächlich etwas verkürzt. Dasselbe gilt für den

Ausschluss stornierter oder zurückgezogener Anforderungen, den Ausschluss der am längsten dauernden Stellen oder eine veränderte Zusammensetzung der Rollen im Nenner.

Der Störfaktor ist größer als der Effekt selbst. Die Besetzungsdauer reagiert stark auf die Arbeitsmarktlage, und die relevanten Vergleichszeiträume der meisten aktuellen Fallstudien überspannen den Wandel von akuter Arbeitskräfteknappheit hin zu deutlich entspannteren Bedingungen. Eine Organisation, die nichts verändert hätte, würde über diesen Zeitraum hinweg ebenfalls eine Verbesserung zeigen. Die Bewegung einem währenddessen eingeführten System zuzuschreiben, erfordert ein Kontrafaktum, und keine Fallstudie dieses Genres enthält eines.

Auch die Einschränkung verdient Beachtung. Eine für Schlüsselpositionen statt für alle Positionen berichtete Verbesserung weist auf eine Teilmenge hin, und wo diese Teilmenge erst nach Bekanntwerden der Ergebnisse ausgewählt wurde, beschreibt die Zahl die Auswahl, nicht das System.

Behauptung zwei: Reduzierung manueller Verwaltungstätigkeiten

Diese Behauptung erfordert eine Ausgangsmessung der manuellen Tätigkeit, und fast keine Personalabteilung verfügt über eine solche.

Wo eine Zahl vorliegt, wurde sie in der Regel auf eine von drei Arten hergeleitet: aus dem Implementierungsmodell eines Anbieters, aus einer Umfrage, in der Mitarbeitende ihre eigene Zeitaufteilung schätzten, oder aus einer Zählung von Tickets oder Vorgängen, deren Definition sich mit der Einführung des neuen Systems geändert hat. Keine davon trägt einen Prozentvergleich über die Zeit, weil sich das Messinstrument im selben Moment änderte wie die Intervention.

Es gibt zudem eine Umfangsfrage, die die Kennzahl verschleiert. Transaktionale Arbeit, die aus der Personalabteilung verschwindet, taucht häufig anderswo wieder auf, meist bei Mitarbeitenden und Führungskräften über Self-Service-Systeme. Eine Reduzierung der von der Funktion bearbeiteten Transaktionen um sechzig Prozent kann einen echten Effizienzgewinn darstellen, eine Verlagerung derselben Arbeit auf Personen, deren Zeit nicht erfasst wird, oder eine Kombination aus beidem. Um dies zu unterscheiden, müsste der gesamte organisatorische Aufwand gemessen werden, was



niemand tut, weil die Kennzahl dazu dient, die Investition der Funktion zu rechtfertigen, nicht die Kosten der Organisation zu beschreiben.

Behauptung drei: Steigerung der Bindung von High Potentials

Dies ist die interessanteste der vier Behauptungen, weil der Mechanismus, der die Zahl erzeugt, meist die Einführung selbst ist.

High Potential ist kein beobachtetes Merkmal. Es ist eine Zuschreibung, die die Organisation vornimmt, und ein KI-gestütztes Talentsystem verändert typischerweise, wie diese Zuschreibung erfolgt. Das ist der erklärte Zweck. Wenn das System verändert, welche Mitarbeitenden als High Potential eingestuft werden, dann ist die nach der Einführung gemessene Gruppe eine andere Population als die zuvor gemessene, und ein Bindungsvergleich zwischen ihnen vergleicht nicht dasselbe.

Die Richtung der resultierenden Verzerrung ist vorhersehbar. Ein Modell, das High Potential anhand von Leistungshistorie, Engagement-Signalen und Karriereverlauf identifiziert, wird bevorzugt Mitarbeitende auswählen, die engagiert sind, kürzlich befördert wurden und sich weiterentwickeln – alles Merkmale, die zu den stärksten verfügbaren Prädiktoren für Verbleib zählen. Das System wählt Menschen aus, die ohnehin geblieben wären, und die Organisation verbucht deren Verbleib als Wirkung des Systems.

Die Kohortengröße verstärkt dies noch. In den meisten Organisationen ist die Gruppe der High Potentials so klein, dass eine Veränderung um achtzehn Prozent nur eine Handvoll Personen betrifft, was gut innerhalb der Schwankungen liegt, die jede kleine Kohorte von Jahr zu Jahr aufweist. Eine solche Zahl sollte mit der Kohortengröße und dem Hinweis versehen sein, ob sich die Definition geändert hat. Das geschieht so gut wie nie.

Behauptung vier: Bias-Minderung

Die vierte Behauptung unterscheidet sich der Art nach von den anderen, denn sie ist nicht bloß unbelegt. Sie ist eine rechtliche Aussage.

Ihre Untermauerung erfordert eine Messung von Bias vor und nach der Einführung, an definierten Gruppen, mit Zellengrößen, die den Vergleich tragfähig machen. Sie er-

fordert außerdem, dass die Gruppen bekannt sind – genau daran scheitert die Behauptung meist. Behinderung lässt sich auf diese Weise überhaupt nicht messen: Arbeitgeber wissen vor einem bedingten Angebot nicht, welche Bewerber behindert sind, und dürfen dies rechtmäßig auch nicht erfragen, sodass kein Instrument zur Messung von Auswirkungsverhältnissen Diskriminierung aufgrund von Behinderung erfassen kann. Eine Organisation, die eine Bias-Minderung bei der Nachfolgeplanung behauptet, hat in der Regel Ethnie und Geschlecht gemessen, sofern sie überhaupt etwas gemessen hat, und nichts über die Population gesagt, die von leistungsdatengetriebenen Modellen am stärksten betroffen ist.

Die Nachfolgeplanung birgt eine weitere Schwierigkeit. Ein Modell, das Bereitschaft oder Potenzial anhand von Betriebszugehörigkeit, Karrierestufe, Zeit in der Position, Aktualität von Qualifikationen und Tempo früherer Beförderungen ermittelt, stützt sich auf Signale, die stark mit dem Alter korrelieren. Eine Bias-Minderung für ein solches System zu berichten, ohne dessen Altersverteilung untersucht zu haben, behauptet das Gegenteil dessen, was die Eingabedaten vorhersagen.

Und die Behauptung selbst schafft Risiko. Title VII, in der Fassung von 1991, erklärt es im Zusammenhang mit der Auswahl oder Vermittlung von Kandidaten für Beschäftigung oder Beförderung für unrechtmäßig, Punktzahlen anzupassen, unterschiedliche Grenzwerte anzuwenden oder die Ergebnisse arbeitsbezogener Tests aufgrund von Ethnie, Hautfarbe, Religion, Geschlecht oder nationaler Herkunft anderweitig zu verändern. Nichts hier deutet darauf hin, dass diese Systeme dies tun. Der Punkt ist: Eine öffentliche Aussage, die ein automatisiertes Beförderungswerkzeug mit demografischen Ergebnissen verknüpft, ist ein Dokument, das die Organisation selbst verfasst hat, das sie nicht zurücknehmen kann und zu dem sie möglicherweise Rechenschaft ablegen muss.

Was allen vieren fehlt

Jedes der vorangegangenen Versäumnisse ist eine Variante derselben Leerstelle. Keine dieser Fallstudien enthält ein Kontrafaktum, das heißt eine Beschreibung dessen, was ohne die Intervention geschehen wäre.

Die Organisation führte ein System in einem Zeitraum ein, in dem sich der Arbeitsmarkt verschob, die Belegschaftsgröße sich änderte, Führungskräfte wechselten, Vergütungen angepasst wurden und andere Initiativen parallel liefen. Sie schrieb dann die Bewegung in vier Kennzahlen einer dieser Veränderungen zu. Dies ist kein kleiner

methodischer Einwand. Es ist die gesamte Fragestellung, und ihr Fehlen ist es, was das Genre unwiderlegbar macht.

Die veröffentlichte Evidenz zu benachbarten Behauptungen sollte jeden Käufer nachdenklich stimmen. Als die Harvard Business School und das Burning Glass Institute untersuchten, was tatsächlich bei Unternehmen geschah, die öffentlich kompetenzbasierte Einstellungsverfahren einführten, stellten sie fest, dass der Verzicht auf Abschlussanforderungen die Einstellung von Kandidaten ohne Abschluss um lediglich rund dreieinhalb Prozentpunkte verschob, was weniger als einer von siebenhundert Einstellungen entspricht, wobei nahezu die Hälfte der Unternehmen trotz ihrer Ankündigungen überhaupt keine Verhaltensänderung zeigte. Die Ankündigungen waren aufrichtig gemeint. Die Praxis wich vollständig davon ab, und die Organisationen, die sie machten, wussten es nicht.

Der naheliegende Schluss ist nicht, dass Menschen lügen. Er ist, dass Organisationen zuverlässig nicht wissen, ob ihre eigenen Initiativen gewirkt haben, und dass eine Fallstudie von der Partei verfasst wird, die am wenigsten in der Lage ist, dies herauszufinden.

Warum der Anspruchsteller es meist selbst glaubt

Es lohnt sich, dies unumwunden zu sagen, denn die alternative Lesart ist wohlwollend und in der Regel falsch.

Diese Zahlen sind selten erfunden. Sie entstehen in einem Messumfeld, in dem keine Ausgangsbasis festgelegt wurde, sich die Definitionen verschoben haben, der Anbieter den analytischen Rahmen geliefert hat und diejenige Person, die die Ergebnisse zusammenstellt, dieselbe ist, die den Kauf befürwortet hat. Jede dieser Bedingungen verzerrt das Ergebnis in dieselbe Richtung, und keine davon erfordert böse Absicht. Die Zahl ist insofern real, als sie jemand berechnet hat. Sie bedeutet nur nicht das, was der Satz darum herum behauptet.

Das ist auch der Grund, warum sich das regulatorische Umfeld zu bewegen begonnen hat. Die Federal Trade Commission ist gegen unbelegte Behauptungen zu Fähigkeiten künstlicher Intelligenz aktiv vorgegangen, darunter eine Anordnung vom Januar 2025, die einen Anbieter von Barrierefreiheitssoftware zur Zahlung von einer Million Dollar verpflichtete, weil dieser behauptet hatte, sein automatisiertes Produkt könne jede Website in Konformität mit Barrierefreiheitsstandards bringen. Das zugrunde

liegende Prinzip beschränkt sich nicht auf diese Produktkategorie. Leistungsbehauptungen zu künstlicher Intelligenz bedürfen einer Substantiierung, und eine Organisation, die Erfolgsprozentsätze in einem Marketingdokument veröffentlicht, stellt genau eine solche Behauptung auf.

Was ein Käufer fragen sollte

Vier Fragen, die etwa zehn Minuten in Anspruch nehmen und die die große Mehrheit der Fallstudien nicht überstehen.

Was war die Ausgangsbasis, wann wurde sie gemessen und von wem. Wurde die Ausgangsbasis erst nach der Einführung erstellt oder vom Anbieter geliefert, ist der Vergleich keiner.

Hat sich die Definition geändert. Bei jeder Kennzahl, die eine bestimmte Personengruppe betrifft, etwa Talente mit hohem Potenzial oder Schlüsselpositionen, sollte man fragen, ob die Zuordnungskriterien vorher und nachher dieselben waren. Wenn das System sie geändert hat, unterscheiden sich die Gruppen, und der Vergleich ist ungültig.

Was sich sonst noch im betreffenden Zeitraum verändert hat. Arbeitsmarkt, Mitarbeiterzahl, Vergütung, Führung, parallele Initiativen. Eine ehrliche Antwort wird eine Liste enthalten. Wer keine hat, hat nicht genau hingesehen.

Wie lautet der Nenner, und wie viele Personen sind erfasst. Prozentangaben ohne Kohortengrößen sind nicht interpretierbar, und kleine Kohorten erzeugen aus geringfügigen Veränderungen große Prozentsätze.

> **EINE FÜNFTE FRAGE**
>
> Eine fünfte Frage lohnt sich bei jeder Behauptung zum Thema Verzerrung: Welche Gruppen wurden gemessen, bei welchen Zellgrößen, und was wurde mit den Gruppen unternommen, die nicht gemessen werden können. Die Antwort auf den letzten Teil lautet fast immer: nichts, und genau das ist die nützlichste Erkenntnis für den Käufer.

Was eine glaubwürdige Fallstudie enthalten müsste

Das Genre ist nicht unrettbar, und eine Organisation, die bereit wäre, Folgendes zu veröffentlichen, würde sich sofort in einem Feld hervorheben, in dem das niemand tut.

Nennen Sie die Ausgangsbasis und deren Datum. Nennen Sie die Definitionen und vermerken Sie, welche sich geändert haben. Berichten Sie Kohortengrößen zusammen mit Prozentangaben. Benennen Sie die parallelen Veränderungen, die die Entwicklung erklären könnten, und sagen Sie, welche Sie nicht ausschließen können. Beschreiben Sie, was nicht funktioniert hat, denn eine Einführung, bei der sich alles verbessert hat, ist eine Einführung, bei der nichts gemessen wurde. Und wo eine Verzerrungsbehauptung aufgestellt wird, nennen Sie die Gruppen, die Methode, die Zellgrößen und was ungemessen bleibt.

Dieses Dokument ist weniger beeindruckend als die derzeitige Form und erheblich nützlicher. Es ist auch die einzige Version, die es übersteht, von jemandem gelesen zu werden, der ein Interesse daran hat, ob sie wahr ist.

Fazit

Die Fallstudie ist das primäre Beweisstück beim Kauf von Personaltechnologie in Unternehmen, und sie ist die schwächste Form von Evidenz im routinemäßigen professionellen Gebrauch. Sie stammt von einer einzigen Organisation, ist unkontrolliert, beruht auf Selbstauskunft, wird nachträglich definiert und von einer Partei veröffentlicht, die ein Interesse am Ergebnis hat.

Nichts davon macht sie wertlos. Es macht sie zu einer Hypothese. Die Organisationen, die tatsächlich Erträge aus diesen Systemen erzielen werden, sind diejenigen, die die Zahlen anderer als Anlass zur Untersuchung und nicht als Kaufgrund behandeln, und die ihre eigenen Zahlen an dem Maßstab messen, den sie auch bei einem Anbieter anlegen würden. Die Disziplin ist weder schwierig noch teuer. Sie besteht aus einer Ausgangsbasis, einer stabilen Definition, einer Kohortengröße und einer ehrlichen Darstellung dessen, was sonst noch zu jener Zeit geschah.

Quellen

- Federal Trade Commission, *In the Matter of accessiBe, Inc.*, Aktenzeichen 2223156, Beschwerde und vorgeschlagene Anordnung bekannt gegeben am 3. Januar 2025.
- Fuller, Joseph, et al. *Skills-Based Hiring: The Long Road from Pronouncements to Practice*. Harvard Business School Project on Managing the Future of Work und The Burning Glass Institute, Februar 2024.
- Title VII of the Civil Rights Act of 1964, in der geänderten Fassung, 42 U.S.C. § 2000e-2, einschließlich § 2000e-2(l).
- Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. Part 1607, anwendbar auf Auswahlverfahren, die Einstellung, Beförderung und andere Beschäftigungsmöglichkeiten betreffen.

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Zitierweise

Grizzle, H. M. (2025, May 23). Vier Zahlen und kein Kontrafaktum: Die Fallstudie zur Künstlichen Intelligenz im Personalwesen lesen. Novara Consulting Group. https://www.novaracg.com/de/2025/05/23/four-numbers-and-no-counterfactual-reading-the-artificial-intelligence-case-study-in-human-resources/