

# Vom Bauchgefühl zum Algorithmus

## Warum KI-Verfahren in der Personalauswahl unverzichtbar werden

**Marcus Heidbrink, Florian Feltes** Zortify S.A., Luxemburg Published: 2025, zortify.com

### Abstract

Trotz jahrzehntelanger Kritik dominieren in der Personalauswahl weiterhin Verfahren wie unstrukturierte Interviews, Lebenslaufanalysen und insbesondere Assessment-Center, obwohl sie zentrale psychometrische Anforderungen oft nicht erfüllen und bei hohem Aufwand nur geringe Prognosekraft aufweisen. Angesichts der wachsenden Bedeutung evidenzbasierter, fairer und skalierbarer Auswahlprozesse rückt der Einsatz künstlicher Intelligenz (KI) zunehmend in den Fokus. Der vorliegende Beitrag beleuchtet zunächst die Defizite klassischer Auswahlverfahren auf Basis aktueller Meta-Analysen. Im Anschluss wird der wissenschaftliche Evidenzstand zu sprachbasierten KI-Verfahren in der Persönlichkeitsdiagnostik systematisch dargestellt. Am Beispiel des Verfahrens *Zortify* werden Ergebnisse zur Objektivität, Reliabilität und Validität präsentiert. Auf dieser Grundlage wird deutlich, dass KI-basierte Diagnostik klassische Verfahren nicht nur sinnvoll ergänzen, sondern in zentralen Aspekten wie Objektivität, langfristiger Messstabilität, prognostischer Validität und ressourcenschonender Skalierbarkeit, auch übertreffen kann. Abschließend werden Anforderungen an eine wissenschaftlich fundierte, ethisch vertretbare und datenschutzkonforme Implementierung diskutiert. Der Beitrag versteht sich als Plädoyer für eine evidenzbasierte, zukunftsfähige Personalauswahl. Angesichts des nachgewiesenen Potenzials ist der Verzicht auf KI-basierte Verfahren nicht nur rückständig, sondern zunehmend verantwortungslos.

**Schlagwörter:** Personalauswahl, künstliche Intelligenz, Assessment-Center, Gütekriterien, Sprachbasierte Persönlichkeitsmessung

#### Zentrale Erkenntnisse:

- Klassische Auswahlverfahren genügen häufig nicht den psychometrischen Mindeststandards.
- KI-gestützte Verfahren bieten einen diagnostischen Mehrwert, wenn sie evidenzbasiert sind.
- Das Assessment-Verfahren Zortify erfüllt zentrale psychometrische Anforderungen.
- Verfahren wie Zortify bieten die Chance, Personalauswahl gerechter, treffsicherer und gleichzeitig preiswerter zu gestalten.

## 1. Status quo der Personalauswahl: Grenzen klassischer Verfahren

Im Zeitalter des demografischen Wandels, verschärften Wettbewerbs um qualifizierte Arbeitskräfte und zunehmender Regulierung wie Datenschutzverordnungen und den europäischen AI Act stehen Organisationen mehr denn je unter Druck, Personalentscheidungen fundiert, effizient, rechtsicher und fair zu treffen. Entsprechend wird der Ruf nach objektiven, evidenzbasierten Auswahlverfahren lauter. Dennoch dominieren in der Unternehmenspraxis weiterhin klassische Verfahren wie unstrukturierte Interviews, die reine Analyse von Bewerbungsunterlagen oder Assessment-Center. Eine wachsende Zahl wissenschaftlicher Arbeiten weist jedoch darauf hin, dass diese Methoden deutliche Schwächen aufweisen, insbesondere im Hinblick auf psychometrische Gütekriterien wie Objektivität, Reliabilität und Validität. Dadurch sind sie maßgeblich mitverantwortlich für Fehlentscheidungen.

Drei aktuelle Übersichtsarbeiten bieten Orientierung. Erstens die umfassende internationale Review von Van Iddekinge, Lievens und Sackett (2023), die eine kritische Bestandsaufnahme klassischer und neuer Verfahren der Personalauswahl vornimmt mit Fokus auf Validität, Fairness, Candidate Experience und technologischen Entwicklungen. Zweitens die Studie von Arnoneit, Schuler und Hell (2020), die sich auf den deutschen Markt konzentriert und anhand einer Langzeiterhebung aufzeigt, wie sich die Nutzung psychologischer Verfahren in der Praxis entwickelt hat. Drittens die Meta-Analyse von Sackett et al. (2022), die konkrete Validitätskoeffizienten für unterschiedliche Auswahlverfahren zusammenfasst.

Die drei Studien offenbaren ein zentrales Paradox: Häufig eingesetzte Methoden in der Personalauswahl weisen nur begrenzte Gütekriterien auf, während innovativere, evidenzbasierte Ansätze bislang weitgehend ungenutzt bleiben. Die Verfahren selbst entwickeln sich nur langsam weiter. Das Potenzial moderner diagnostischer Methoden, etwa KI-gestützter Analysen, wird noch nicht ausgeschöpft, obwohl es substanziell zur Verbesserung von Fairness, Genauigkeit und Vorhersagekraft beitragen könnte.

## 1.1 Häufig eingesetzte Auswahlverfahren

Trotz der bekannten Defizite in Fairness, Validität, Reliabilität und Objektivität dominieren klassische Verfahren die Auswahlpraxis. Die Studie von Armoneit et al. (2020) belegt, dass deutsche Unternehmen nach wie vor primär auf folgende Methoden setzen:

1. die Analyse von Bewerbungsunterlagen (82 % der Unternehmen),
2. unstrukturierte Interviews (34 %), strukturierte Interviews (73 %),
3. Arbeitsproben (46 %)
4. und Assessment-Center (38 %).
5. Testbasierte Verfahren wie Persönlichkeitstests (19 %; online 24%)
6. oder Leistungstests (24 %; online 16%)

spielten in der Praxis bislang eine untergeordnete, wenn auch aufstrebende Rolle. Auch Van Iddekinge et al. (2023) berichten auf Basis internationaler Daten von einer ähnlichen Rangfolge und geringen Innovationsdynamik in der Auswahlmethodik.

## 1.2 Psychometrische Qualität klassischer Auswahlverfahren

Im Zentrum jeder personaldiagnostischen Methode steht ihre wissenschaftliche Qualität: ihre Objektivität, Reliabilität und Validität. Die Analysen von Armoneit et al. (2020), Sackett et al. (2022), sowie Van Iddekinge et al. (2023) verdeutlichen, dass gerade jene Verfahren, die am weitesten verbreitet sind, häufig nicht den psychometrischen Mindestanforderungen genügen. Unter den klassischen Verfahren nehmen Assessment-Center eine besondere Rolle ein: Sie gelten trotz zahlreicher empirischer Hinweise auf ihre begrenzte Validität, hohe Kosten und methodische Fehleranfälligkeit in vielen Organisationen als diagnostischer „Goldstandard“.

**Bewerbungsunterlagen** etwa werden in nahezu jedem Auswahlprozess herangezogen. Dennoch liegt ihre Validität mit Korrelationen zum Berufserfolg von  $r = .07$  für Berufserfahrung in Jahren und  $r = .22$ , und das für den Bestfall der systematisch ausgewertete Lebenslaufdaten (Sackett et al., 2022), deutlich unter dem akzeptablen Bereich. Die Bewertung erfolgt meist intuitiv, ohne einheitliche Kriterien, und unterliegt erheblichen Beobachtungsfehlern. Auch die Reliabilität ist entsprechend niedrig. Die Attraktivität dieser Methode ergibt sich vor allem aus ihrer Praktikabilität, nicht aus ihrer diagnostischen Aussagekraft.

**Interviews**, insbesondere wenn sie unstrukturiert durchgeführt werden, weisen eine geringe Objektivität und eine nur moderate Validität ( $r = .19$ ; Sackett et al., 2022) auf. Erst strukturierte Interviews erreichen etwas höhere Werte ( $r = .42$ ; Sackett et al., 2022), werden aber in der Praxis seltener eingesetzt. Die hohe Subjektivität und der subjektive Einfluss der Interviewer:innen, Formulierung der Fragen und fehlende Bewertungsschemata bleiben zentrale Schwächen unstrukturierter Interviews.

**Arbeitsproben** bieten eine moderate Validität ( $r = .33$ ; Sackett et al., 2022). Sie sind besonders wirksam in handlungsnahen Rollen und gelten als akzeptiert bei Bewerbenden. Ihr Einsatz ist jedoch oft mit erhöhtem organisatorischem Aufwand verbunden.

**Assessment-Center** gelten oft als „Goldstandard“ der Personalauswahl, obwohl sie seit Jahren hinter diesem Anspruch zurückbleiben. Die genannten Übersichtsarbeiten weisen darauf hin, dass sie diesem Ruf empirisch und praktisch nicht gerecht werden (Armoneit et al., 2022, Van Iddekinge et al., 2023; Sackett et al., 2022). Der Aufwand steht in vielen Fällen in keinem Verhältnis zur diagnostischen Aussagekraft. Studien weisen darauf hin, dass selbst hochstrukturierte Assessment-Center durch Beobachtungsfehler, geringe Kriteriumsvalidität ( $r = 0.29$ ; Sackett et al., 2022) und hohe Kosten nur bedingt geeignet sind, valide Aussagen über Berufserfolg zu treffen. Sie sind stark abhängig vom Design, der Schulung der Beobachter:innen und der Standardisierung der Übungen (Gaugler et al., 1987). Beobachtende unterschätzen regelmäßig die Verzerrungseffekte durch ihre Subjektivität. Die tatsächliche Aussagekraft dieser Verfahren ist limitiert.

**Persönlichkeitstests auf Selbstauskunftsbasis** erzielen in der Regel akzeptable Werte für Objektivität und Reliabilität. Ihre prognostische Validität für Berufserfolg ist ebenfalls als moderat einzustufen ( $r = .25 - .30$ ; Sackett et al., 2022). Studien zu Arbeitsleistung, kontraproduktivem Verhalten oder Teamfähigkeit (Barrick & Mount, 1991; Zell & Lesick, 2022) betonen allerdings die Relevanz von Persönlichkeitsmerkmalen für Berufserfolg. Persönlichkeitstests auf Selbstauskunftsbasis sind anfällig für sozial erwünschtes Antwortverhalten, insbesondere in Auswahlsettings (Birkeland et al., 2006; Kowalski et al., 2018), was zu eingeschränkten Werten der prognostischen Validität beitragen kann. Van Iddekinge et al. (2023) empfehlen daher, diese Tests mit beobachtungsbasierten zu kombinieren, um Verzerrungen zu minimieren und authentischere Persönlichkeitsprofile zu erfassen, auch durch den Einsatz validierter KI-basierter Verfahren.

### 1.3 Gründe für die Dominanz schwächerer Verfahren

Warum aber halten sich einige Verfahren trotz ihrer empirisch belegten Schwächen so hartnäckig? Die Antwort liegt in einer Kombination aus Praktikabilität, sozialer Akzeptanz und institutionellen Barrieren. So betonen Armoneit et al. (2020), dass unstrukturierte Interviews und Lebenslaufanalysen als besonders einfach, kosteneffizient und flexibel gelten und zugleich kulturell „normalisiert“ sind; Personaler:innen empfinden diese Methoden als intuitiv und interaktiv, was ihre Akzeptanz sowohl auf Seiten der Unternehmen als auch der Bewerbenden erhöht. Assessment-Center gelten häufig als besonders anschlussfähig, obwohl sie zu den aufwendigsten und teuersten Auswahlverfahren zählen, bei gleichzeitig nur moderater Validität.

Van Iddekinge et al. (2023) ergänzen, dass viele Entscheidungsträger die Aussagekraft subjektiver Verfahren überschätzen („illusorische Validität“) und evidenzbasierte Verfahren aufgrund ihrer Komplexität, wahrgenommenen Kälte oder rechtlicher Unsicherheit meiden. Ein Bericht von Rahe & Rahe (2017) belegt ebenfalls, dass lediglich 39 % der befragten Unternehmen „Wissenschaftlichkeit“ als zentrales Kriterium für diagnostische Verfahren nennen, während Aspekte wie Verständlichkeit (68 %) oder Praktikabilität (77 %) weit häufiger genannt werden.

Diese Befunde zeigen deutlich: Die Schwächen klassischer Auswahlverfahren sind bekannt. Assessment-Center und andere klassische Verfahren werden trotz hoher Kosten, logistischer Komplexität und moderater Validität weiterhin eingesetzt; häufig weniger aus ihrer diagnostischen Qualität als aus psychologischer Bequemlichkeit und institutioneller Trägheit.

### 1.4 Perspektiven KI-gestützter Diagnostik

Vor diesem Hintergrund rücken KI-gestützte Verfahren zunehmend in den Fokus der Forschung und Praxis. Van Iddekinge et al. (2023) sehen Potential im Einsatz von KI für die Personalauswahl. Diagnostische Verfahren können durch den Einsatz empirisch validierter KI-Verfahren objektiver, robuster und effizienter gestaltet werden, so dass Personalfehlentscheidungen seltener werden. Insbesondere Sprachdaten bieten einen neuen Zugang zu latenten Merkmalen wie Persönlichkeit, Motivation oder kognitiver Stil, der weniger anfällig für bewusste Verfälschung ist (z. B. Moreno et al., 2021; Yarkoni, 2010).

Zu den möglichen Potenzialen des Einsatzes von KI in der Personalauswahl gehören:

1. die Reduktion subjektiver Verzerrungen in Bewertungen,
2. der Zugang zu schwer messbaren Merkmalen,
3. die Analyse unstrukturierter Daten,
4. die Verarbeitung großer Datenmengen und
5. die Skalierbarkeit und Kosteneffizienz.

Aus den genannten Aspekten der Validität (1 und 2) sowie den analytischen Möglichkeiten moderner Datenverarbeitung (3 und 4) ergibt sich konsequent das Argument der Skalierbarkeit und Kosteneffizienz (5). Klassische Verfahren wie Assessment-Center verursachen pro Kandidat:in hohe Kosten in der Durchführung und schützen aufgrund ihrer Fehleranfälligkeit nicht vor teuren Personalfehlerscheidungen. Im Gegensatz dazu ermöglichen digitalisierte, evidenzbasierte KI-Verfahren sowohl eine ortsunabhängige, vollautomatisierte Durchführung mit minimalem Ressourceneinsatz als auch eine – bei erhöhter Prognosekraft insbesondere durch die Punkte 1 und 2 – höhere Trefferquote bei der Stellenbesetzung. Insbesondere im Hinblick auf Skalierbarkeit und langfristige Kostenreduktion bieten evidenzbasierte KI-Verfahren demzufolge eine nachhaltige Lösung, die klassischen Auswahlmethoden wie Assessment-Centern sowohl ökonomisch als auch diagnostisch überlegen sein können.

Voraussetzung für den erfolgreichen Einsatz ist jedoch, dass auch KI-gestützte Verfahren strenge wissenschaftliche Validierungsprozesse durchlaufen und dieselben psychometrischen Standards erfüllen wie klassische Instrumente. In diesem Sinne können sie nicht nur bestehende Schwächen klassischer Verfahren kompensieren, sondern die Personalauswahl in eine neue Ära führen; weg vom Bauchgefühl, hin zu datengestützter, fairer und skalierbarer Entscheidungsfindung.

## **2. Psychometrische Qualität KI-gestützter Verfahren: Eine Analyse aktueller Forschung**

Die oben dargestellten Ergebnisse machen deutlich, dass viele der aktuell in der Praxis dominierenden Auswahlverfahren, etwa unstrukturierte Interviews, Lebenslaufanalysen oder klassische Assessment Center, nur unzureichend validiert und anfällig für subjektive Verzerrungen sind. Gerade Assessment-Center zeigen sich in Studien als besonders anfällig für Beobachterverzerrungen, Gruppeneffekte und kontextuelle Einflussfaktoren. Die objektive Vergleichbarkeit zwischen Kandidat:innen

ist dadurch stark eingeschränkt. Trotz ihrer weiten Verbreitung weisen sie methodische Schwächen auf, die nachweislich zu systematischen Fehlentscheidungen führen können.

Vor diesem Hintergrund richtet sich der Blick zunehmend auf neue, technologiegestützte Ansätze, insbesondere auf Verfahren der künstlichen Intelligenz (KI). Sprachbasierte KI-Systeme, die natürliche Sprache (NLP-Modelle) aus Bewerbungsunterlagen, Interviews oder offenen Antworten analysieren, versprechen hier einen Paradigmenwechsel. Sie haben das Potenzial, standardisierte, transparente und skalierbare Diagnostikprozesse zu ermöglichen; mit dem Ziel, menschliche Urteilsverzerrungen zu minimieren, Persönlichkeitsmerkmale zuverlässiger zu erfassen und Personalfehlscheidungen zu minimieren.

Der folgende Abschnitt gibt einen systematischen Überblick über die empirische Evidenz zum Einsatz KI-basierter Verfahren in der Personalauswahl – mit Fokus auf Validität, Reliabilität, Objektivität und praktische Umsetzbarkeit. Grundlage sind aktuelle Meta-Analysen, systematische Reviews sowie vielversprechende empirische Einzelstudien der letzten Jahre.

## **2.1 Synthese aktueller Übersichtsarbeiten zur KI-basierten Persönlichkeitsdiagnostik**

Das Interesse an KI-basierten Verfahren zur Analyse von Persönlichkeit ist stark gestiegen. Zahlreiche Meta-Analysen und systematische Reviews haben untersucht, inwiefern KI-gestützte Verfahren zuverlässig Persönlichkeitsmerkmale erfassen können, insbesondere auf Basis von Textdaten. Im Mittelpunkt stehen dabei meist die Big-Five-Persönlichkeitsdimensionen, seltener auch die Dunkle Triade.

Die bisherige Studienlage zeigt ein insgesamt vielversprechendes Bild, auch wenn die Qualität der Verfahren stark variiert. Seit 2019 ist eine große Zahl an Überblicksarbeiten erschienen, die in diesem Kapitel systematisch zusammengefasst werden. Die analysierten Arbeiten bestehen aus zwei Meta-Analysen (Moreno et al., 2023; Naz et al., 2023) und drei Reviews (Bhandarkar et al., 2024; Hashemi-Motlagh et al., 2025; Koutsoumpis et al., 2024). Sie bewerten durchschnittliche Validitäten, vergleichen verschiedene Modellarchitekturen und benennen Herausforderungen wie mangelnde Standardisierung und Transparenz. Eine Übersicht der analysierten Arbeiten, ihrer Vorgehensweise und der zentralen Ergebnisse ist in Tabelle 1 abgebildet.

Tabelle 1

Übersicht der Überblicksarbeiten zur Qualität KI-basierter Verfahren zur Messung von Persönlichkeit auf Basis von Text

Autor:innen	Methode	Zentrale Ergebnisse
Moreno et al. (2021)	<ul style="list-style-type: none"> <li>• Meta-Analyse</li> <li>• 23 Primärstudien zu KI-basierter Persönlichkeitsdiagnostik via Text</li> <li>• Fokus: Zusammenhang zwischen Sprachdaten und Big Five</li> <li>• Moderatorenanalyse (z. B. Textquelle, Trait, Modelltyp)</li> <li>• Ziel: prädiktive Validität bestimmen</li> </ul>	<ul style="list-style-type: none"> <li>• Durchschnittliche prädiktive Validität: <math>r \approx .26-.30</math> für Big Five</li> <li>• Textquelle und Trait moderieren die Vorhersagegenauigkeit (z. B. höhere Werte für Gewissenhaftigkeit, geringere für Verträglichkeit)</li> <li>• Klassische ML-Modelle (z. B. SVM, Regression) zeigen moderate Leistungen</li> <li>• Validität verbessert sich mit längerem und thematisch breiterem Textinput</li> </ul>
Koutsoumpis et al. (2022)	<ul style="list-style-type: none"> <li>• Meta-Analyse</li> <li>• 31 unabhängige Stichproben (~85.000 Teilnehmende)</li> <li>• Fokus: Zusammenhänge zwischen KI-basierten-Kategorien und Big Five</li> <li>• Unterscheidung von Selbst- vs. Fremdbeschreibung</li> <li>• Effektgrößen als korrigierte Korrelationen (<math>\rho</math>)</li> <li>• Moderatoranalysen: Plattform, Sprache, Beschreibungstyp</li> </ul>	<ul style="list-style-type: none"> <li>• Verfahren zeigen geringe bis moderate Korrelationen mit Big Five (Selbstberichte: <math> \rho  \approx .08-.14</math>; Fremdberichte: <math> \rho  \approx .18-.39</math>)</li> <li>• Höhere Validität bei Fremdbeschreibungen als bei Selbstberichten</li> <li>• Kontextabhängigkeit entscheidend: Plattform, Sprache, Kommunikationsziel beeinflussen Ergebnisse</li> <li>• Schlussfolgerung: KI erfasst Aspekte von Persönlichkeit, aber keine vollständige Repräsentation</li> </ul>
Hashemi-Motlagh et al. (2025)	<ul style="list-style-type: none"> <li>• Review</li> <li>• Über 100 Studien zu Text-, Audio- und Videoanalyse</li> <li>• Fokus: Vergleich von Modalitäten und Modellarchitekturen</li> <li>• Kategorisierung nach Eingabetyp, Zielmerkmal, Methodik</li> <li>• Ziel: Überblick über Stand der Technik und Anwendungsbereiche</li> </ul>	<ul style="list-style-type: none"> <li>• Multimodale Verfahren (Text + Audio/Video) zeigen die höchsten Validitäten</li> <li>• Transformer-basierte Modelle schneiden konstant besser ab als klassische Modelle</li> <li>• Validierung häufig uneinheitlich; Empfehlung für mehr standardisierte Benchmarks</li> </ul>
Naz et al. (2025)	<ul style="list-style-type: none"> <li>• Review: methodisch-technischer Überblick</li> <li>• Fokus: Modelle SVM, CNN, RNN, Transformer</li> <li>• Vergleich von Verfahren, Datenarten und Feature-Engineering</li> <li>• Bewertung anhand Vorhersageleistung, Trainingsdaten, Evaluationsmetriken</li> </ul>	<ul style="list-style-type: none"> <li>• Transformer-Architekturen (BERT, GPT) liefern überdurchschnittliche Vorhersagegüte</li> <li>• Klassische Modelle (SVM, Random Forest) liefern inkonsistente Ergebnisse</li> <li>• Aufruf zu kontextsensitiven, datenschutzgerechten Anwendungen im Personalbereich</li> </ul>
Bhandarkar et al. (2024)	<ul style="list-style-type: none"> <li>• Review: Vergleich bestehender Verfahren</li> <li>• Bewertung anhand Gütekriterien (Objektivität, Validität, Fairness, Erklärbarkeit)</li> <li>• Diskussion ethischer Herausforderungen und Bias-Risiken</li> </ul>	<ul style="list-style-type: none"> <li>• Objektivität und Automatisierung als große Stärke</li> <li>• Kritik an fehlender Erklärbarkeit und psychologischer Fundierung</li> <li>• Plädoyer für ethische Mindeststandards und psychologische Fundierung</li> </ul>

## 2.2 Psychometrische Qualität KI-basierter Persönlichkeitsdiagnostik

Die Übersichtsarbeiten kommen zu dem Ergebnis, dass KI- und NLP-basierte Verfahren in vielen Fällen geeignet sind, um Persönlichkeitsmerkmale objektiv, reliabel und valide zu erfassen. Insbesondere durch ihre Unabhängigkeit von klassischen Selbstauskünften gelten sie als vielversprechend in Bezug auf ökologische Validität, Standardisierbarkeit und Automatisierung (Bhandarkar et al., 2024).

**Validität.** Die bisherige empirische Evidenz zur Validität sprachbasierter KI-Verfahren in der Persönlichkeitsdiagnostik zeigt ein insgesamt vielversprechendes, aber heterogenes Bild. Meta-Analysen und systematische Reviews kommen übereinstimmend zu dem Schluss, dass KI-basierte Modelle relevante Aspekte von Persönlichkeit erfassen können, die Validitäten jedoch moderat und stark kontextabhängig sind. Die Meta-Analyse von Moreno et al. (2021) berichtet für Big-Five-Merkmale durchschnittliche prädiktive Validitäten im Bereich von  $r \approx .26$ – $.30$ , wobei insbesondere längere und inhaltlich reichhaltige Texte bessere Vorhersagen ermöglichen. Koutsoumpis et al. (2022) ergänzt diese Befunde mit korrigierten Effektgrößen ( $\rho$ ) zwischen  $.08$  und  $.14$  (Selbstberichte) sowie  $.18$  bis  $.39$  (Fremdurteile). Die Ergebnisse verdeutlichen: KI-Verfahren sind in der Lage, relevante Persönlichkeitsmerkmale zu erfassen, auch wenn sie keine vollständige Abbildung klassischer Konstrukte leisten.

**Reliabilität.** Auch zur Reliabilität liegen erste Daten vor: In groß angelegten Studien mit offenen Texteingaben oder KI-basierten Chatbots werden Retest-Koeffizienten zwischen  $r = .48$  und  $r = .70$  berichtet, was auf eine moderate zeitliche Stabilität hinweist (Park et al., 2015; Fan et al., 2023; Hickman et al., 2022). Hybride Verfahren, die Fragebogendaten mit NLP-Analysen kombinieren, erreichen Retest-Reliabilitäten zwischen  $r = .30$  und  $r = .60$ , abhängig vom jeweiligen Persönlichkeitsmerkmal (Koutsoumpis et al., 2024; Zhang et al., 2024).

**Objektivität.** KI-basierte Verfahren zur Persönlichkeitsdiagnostik gelten in der Literatur als objektiv, da sie standardisiert, automatisiert und frei von menschlichen Urteilsverzerrungen ablaufen, wenn die Trainingdatenqualität hoch ist (Bhandarkar et al., 2024; Naz et al., 2025). Im Gegensatz zu klassischen Auswahlverfahren wie unstrukturierten Interviews oder Assessment-Centern, bei denen subjektive Einschätzungen der Beobachtenden einen erheblichen Einfluss auf das Ergebnis haben können, beruhen KI-gestützte Systeme auf konsistenten algorithmischen Auswertungen, die unabhängig von der Person der Bewerbenden oder Bewertenden sind. KI-basierte Verfahren nutzen ausschließlich freiwillig bereitgestellte

Textinformationen. Da keine Fotos, Namen oder Angaben zu Geschlecht und Alter erforderlich sind, wird nicht nur die Anfälligkeit für implizite Biases verringert, sondern zugleich ein hohes Maß an Datenschutz und Fairness gewährleistet. Diese Form der Standardisierung erhöht die Vergleichbarkeit über Personen, Kontexte und Zeitpunkte hinweg. Gleichzeitig verweisen mehrere Reviews auf notwendige Einschränkungen: Die tatsächliche Objektivität hängt maßgeblich von der Qualität und Verzerrungsfreiheit der Trainingsdaten sowie von der Transparenz der eingesetzten Modelle ab (Hashemi-Motlagh et al., 2025; Bhandarkar et al., 2024).

Insbesondere sogenannte Black-Box-Modelle bergen das Risiko, dass zwar die Auswertung automatisiert erfolgt, aber die zugrundeliegenden Entscheidungsregeln für Nutzende nicht nachvollziehbar sind. Objektivität im engeren psychometrischen Sinn setzt jedoch voraus, dass Ergebnisse nicht nur standardisiert erzeugt, sondern auch intersubjektiv nachvollziehbar sind. Der Einsatz erklärbarer KI (Explainable AI) wird daher zunehmend als zentrale Anforderung diskutiert.

### 2.3 Determinanten der Qualität KI-gestützter Verfahren

Die Qualität KI-basierter Persönlichkeitsdiagnostik wird durch eine Reihe technischer und psychometrischer Faktoren beeinflusst:

- **Psychologische Fundierung:** Verfahren, die auf validierten Persönlichkeitsmodellen (z. B. Big Five) basieren, schneiden konsistenter besser ab. Ihre theoretische Klarheit erhöht sowohl die Validität als auch die Akzeptanz in der Anwendung (Bhandarkar et al., 2024).
- **Textumfang und -qualität:** Die Länge und semantische Tiefe der analysierten Texte ist entscheidend. Längere, persönlichere Texte mit höherer inhaltlicher Vielfalt ermöglichen stabilere Persönlichkeitsvorhersagen (Moreno et al., 2021).
- **Sprachliche und kulturelle Robustheit:** Viele Verfahren sind primär auf englischsprachige Trainingsdatensätze ausgelegt. Die Übertragbarkeit auf andere Sprachen und kulturelle Kontexte ist oft eingeschränkt, was die internationale Anwendbarkeit limitiert (Koutsoumpis et al., 2022).
- **Modellarchitektur:** Der leistungsfähigste Prädiktor für Qualität ist die verwendete Modellstruktur. Transformer-basierte Modelle wie BERT, RoBERTa oder GPT zeigen deutlich bessere Vorhersageleistungen als klassische Ansätze (z. B. Random Forests, SVMs oder LIWC-basierte Modelle; Naz et al., 2025). Sie bieten zudem höhere Generalisierbarkeit und geringere Fehleranfälligkeit.
- **Multimodale Verfahren:** Die höchste Validität erreichen Ansätze, die Textdaten mit weiteren Kanälen kombinieren, z. B. Video- oder Stimmmerkmale. Solche Verfahren

eignen sich besonders für komplexe Auswahlkontexte wie Videointerviews (Hashemi-Motlagh et al., 2025).

- **Erklärbarkeit und Transparenz:** Die Nachvollziehbarkeit von KI-Entscheidungen ist ein zentraler Erfolgsfaktor, insbesondere in regulierten Bereichen wie der Personalauswahl. Erklärbare Modelle fördern das Vertrauen von Anwender:innen und erhöhen die Akzeptanz (Bhandarkar et al., 2024).

Mit Blick auf die zeitliche Entwicklung der durch die Überblicksarbeiten analysierten Einzelstudien ist eine deutliche Weiterentwicklung KI-gestützter Verfahren zur Persönlichkeitsdiagnostik zu beobachten, die sich in besseren Ergebnissen aktuellerer Arbeiten widerspiegeln. Insbesondere der Einsatz tiefer neuronaler Netze sowie multimodaler Architekturen, etwa durch die Kombination von Text-, Audio- oder Videodaten, hat die diagnostische Genauigkeit und Anwendbarkeit dieser Verfahren verbessert (Naz et al., 2025; Hashemi-Motlagh et al., 2025). Diese methodischen Fortschritte spiegeln sich nicht nur in den aggregierten Befunden von Meta-Analysen wider, sondern zeigen sich zunehmend auch in empirischen Einzelstudien, die moderne KI-gestützter Verfahren untersuchen.

## 2.4 Empirische Einzelbefunde: Evidenz für diagnostischen Mehrwert

Während systematische Übersichten wichtige aggregierte Erkenntnisse zur Leistungsfähigkeit KI-gestützter Verfahren liefern, bietet der Blick auf einzelne empirische Studien zusätzlichen Erkenntnisgewinn. Besonders aktuelle Arbeiten mit modernen Sprachmodellen und methodischen Ansätzen zeigen, dass KI-basierte Analysen in realen diagnostischen Situationen häufig valide, differenzierte und robuste Einschätzungen ermöglichen. Im Folgenden werden drei exemplarische Studien vorgestellt, die den Mehrwert solcher Verfahren unter spezifischen Bedingungen empirisch zeigen.

So zeigen Hickman et al. (2022), dass KI-basierte Scores aus offenen Interviewantworten eine mittlere bis hohe Übereinstimmung mit klassischen Selbst- ( $\bar{r} \approx .19$ ) und insbesondere mit Fremdratings ( $\bar{r} \approx .24$ ) aufweisen. Die höhere Konvergenz mit Expert:innen-Urteilen legt nahe, dass maschinelle Analysen soziale Wunschbilder oder Selbstverzerrungen besser umgehen können.

Sikström et al. (2025) demonstrieren, dass textbasierte KI-Klassifikationen der Big Five bei kurzen Texteingaben eine bis zu 10 % höhere Treffsicherheit erreichen als etablierte Fragebogenverfahren. Gleichzeitig berichten sie über eine verbesserte interne

Konsistenz der KI-basierten Skalen, insbesondere bei Traits wie Offenheit oder Neurotizismus.

Eine weitere Studie von Tu et al. (2024) zeigt am Beispiel von Narzissmus, dass GPT-basierte Sprachanalysen in offenen Antworten signifikant stärker mit Expert:innen-Einschätzungen korrelieren als traditionelle Selbstberichte. Die KI-basierten Einschätzungen erfassen dabei relevante Persönlichkeitsaspekte konsistenter und sensibler; ein Befund, der insbesondere bei verdeckten oder sozial unerwünschten Traits bedeutsam ist.

In der Zusammenschau zeigt sich, dass KI-gestützte Diagnostikverfahren nicht nur in der Lage sind, klassische psychometrische Anforderungen wie Objektivität, Reliabilität und Validität zu erfüllen, sondern in einigen Bereichen sogar überlegen sein können. Besonders hervorzuheben ist die hohe Standardisierbarkeit und Bewertungsobjektivität algorithmischer Auswertung, die im Vergleich zu menschlichen Beurteilungen deutlich robuster gegenüber Verzerrungen ist. Auch in Bezug auf prognostische Validität und Verhaltenserfassung, insbesondere bei schwer zugänglichen Konstrukten wie sozial erwünschtem Verhalten, subklinischen Persönlichkeitstendenzen oder impliziten Motiven, liefern KI-basierte Systeme zunehmend differenzierte und valide Ergebnisse. Damit können sie klassische Verfahren nicht nur sinnvoll ergänzen, sondern in zentralen Aspekten potenziell übertreffen. Dies legt nahe, dass KI-gestützte Instrumente in der modernen Personalauswahl nicht als technologische Spielerei, sondern als substantielle methodische Weiterentwicklung verstanden werden sollten.

Vor diesem Hintergrund lohnt sich ein genauerer Blick auf ein konkretes Anwendungsbeispiel aus der Praxis: das KI-gestützte Diagnostiktool *Zortify*. Im folgenden Kapitel wird untersucht, inwieweit dieses Verfahren den wissenschaftlichen Anforderungen genügt und welchen Beitrag es zur Professionalisierung moderner Personalauswahl leisten kann.

### **3. Zortify: Psychometrische Qualität eines KI-basierten Diagnostiktools**

Der Einsatz künstlicher Intelligenz in der Persönlichkeitsdiagnostik wirft berechtigte Fragen nach ihrer wissenschaftlichen Fundierung auf (vgl. van Iddekinge et al., 2023). Das Unternehmen Zortify hat daher ein mehrstufiges Verfahren entwickelt, das sowohl klassische Selbstauskünfte als auch NLP-basierte Textanalysen kombiniert. Ziel ist es,

valide, objektive und reliabel erhobene Persönlichkeitsprofile zu generieren, die sich sowohl theoretisch als auch empirisch bewähren. Im Folgenden wird dargestellt, wie Zortify die zentralen psychometrischen Gütekriterien erfüllt und welche empirischen Evidenzen für die Qualität des Verfahrens sprechen. Tabelle 2 fasst die Ergebnisse zusammen.

Tabelle 2  
Nachweise der psychometrischen Qualität von Zortify

Gütekriterium	Ergebnisse
Objektivität	Hohe Durchführungs-, Auswertungs- und Interpretationsobjektivität durch vollautomatisierte Abläufe
Interne Konsistenz	Cronbach's $\alpha = .59-.91$
Retest-Reliabilität	$r = .72 - .86$ (nach 2 Wochen)
	$r = .70 - .80$ (bis zu 410 Tage)
Konstruktvalidität	Faktorenanalyse: CFI = .98, RMSEA = .06
	Konvergente Validität anhand HEXACO ( $r = .44 - .80$ )
Kriteriumsvalidität	Geförderte Personen zeigen signifikant höhere Werte bei Optimismus & Unternehmerischem Kapital, niedrigere bei Psychopathie
	Vorhersage sprachlicher Muster gemäß wissenschaftlicher Literatur (z. B. „wir“ bei Extraversion, „Ich“ bei Neurotizismus)
Normierung	Normstichprobe mit >16.000 Personen aus Europa

### 3.1 Objektivität

Die Objektivität eines diagnostischen Verfahrens beschreibt, inwieweit die Ergebnisse unabhängig von der Person der Testleitung sind. Laut Testmanual erfüllt das Zortify-Verfahren diese Anforderung in hohem Maße. Die Durchführung erfolgt vollständig digitalisiert, die Instruktionen sind standardisiert, und die Auswertung erfolgt automatisiert über algorithmisch gesteuerte, KI-basierte Prozesse. Dadurch ist sichergestellt, dass keine manuelle oder subjektive Einflussnahme auf die Testergebnisse möglich ist. Insbesondere wird betont, dass bei Zortify sowohl Durchführungs-, Auswertungs- als auch Interpretationsobjektivität gegeben sind; ein wesentlicher Vorteil gegenüber klassischen, häufig durch menschliche Beurteilungen beeinflussten Verfahren wie unstrukturierten Interviews oder Assessment-Centern. Gerade im Vergleich zu Assessment-Centern, die trotz erheblichem Ressourceneinsatz mit Beobachtungsfehlern und Designsensitivität kämpfen, zeigen KI-Verfahren wie Zortify, dass valide Diagnostik auch ohne teure Präsenzformate realisierbar ist.

### 3.2 Reliabilität

Für die Persönlichkeitsdiagnostik ist eine hohe Reliabilität essenziell, da stabile Eigenschaften nur dann sinnvoll erfasst werden können, wenn die Erhebung selbst stabil operiert.

Die empirischen Befunde zur Reliabilität des Zortify-Verfahrens zeigen ein überzeugendes Bild. Die interne Konsistenz der Skalen, gemessen über Cronbach's Alpha, liegt je nach Dimension zwischen  $\alpha = .59$  und  $\alpha = .91$ . Damit erreicht das Verfahren Werte, die mit etablierten Fragebogenverfahren vergleichbar sind, und erfüllt die gängigen Anforderungen an die Konsistenz diagnostischer Skalen unter Berücksichtigung der Testlänge.

Auch die zeitliche Stabilität der Ergebnisse wurde umfassend untersucht. In einer Studie mit einer Testwiederholung nach zwei Wochen zeigten sich hohe Retest-Reliabilitäten zwischen  $r = .72$  und  $r = .86$ . Diese Werte sprechen für eine zuverlässige Erfassung stabiler Persönlichkeitsmerkmale über kurze Zeiträume hinweg.

Darüber hinaus wurde die langfristige Retest-Stabilität überprüft. Die Korrelationen über Zeiträume von bis zu 410 Tagen liegen zwischen  $r = .70$  und  $r = .80$ , was die dauerhafte Messpräzision des Verfahrens belegt. Im Vergleich mit anderen KI-gestützten Verfahren zeigt sich Zortify hinsichtlich der Reliabilität überlegen. Während frühere Studien zu textbasierten Persönlichkeitsanalysen, etwa von Park et al. (2015), Fan et al. (2023) oder Hickman et al. (2022), Retest-Koeffizienten im Bereich von  $r = .48$  bis  $r = .70$  berichten, liegt Zortify deutlich darüber. Dies lässt sich vermutlich auf die hybride Methodik zurückführen, die klassische Selbstauskunft mit KI-gestützter Sprachanalyse kombiniert.

Insgesamt weist das Verfahren eine robuste und zeitlich stabile Erfassung von Persönlichkeitsmerkmalen auf und erfüllt damit zentrale Anforderungen an ein psychometrisch tragfähiges Instrument.

### 3.3 Validität

Für ein Instrument zur Persönlichkeitsdiagnostik ist die Validität von zentraler Bedeutung, da nur so sichergestellt werden kann, dass die erfassten Merkmale inhaltlich zutreffend und in ihrer Bedeutung interpretierbar sind. Das Zortify-Verfahren wurde in mehreren empirischen Studien systematisch auf seine Konstrukt- und Kriteriumsvalidität hin überprüft.

### 3.4 Konstruktvalidität

Die Konstruktvalidität von Zortify wird durch eine Reihe psychometrischer Analysen gestützt. So bestätigten Faktorenanalysen die zugrunde liegende Struktur des Verfahrens. Die Modellgütekennwerte der konfirmatorischen Faktorenanalyse (CFI = .92; RMSEA = .06) sprechen für eine solide Modellpassung und stützen die Annahme, dass das Verfahren theoretisch fundierte Persönlichkeitsdimensionen repräsentativ abbildet.

Zur Überprüfung der konvergenten Validität wurde das Zortify-Verfahren mit dem etablierten HEXACO-Inventar verglichen (N = 169). Die dabei erzielten Korrelationen unterstreichen die inhaltliche Übereinstimmung beider Verfahren: Für Extraversion lag die Übereinstimmung bei  $r = .80$ , für Offenheit bei  $r = .73$ . Auch für Gewissenhaftigkeit ( $r = .57$ ), emotionale Stabilität ( $r = .43$ ) und Verträglichkeit ( $r = .44$ ) wurden substantielle Zusammenhänge festgestellt. Gleichzeitig konnte die diskriminante Validität nachgewiesen werden: Die Interkorrelationen zwischen Skalen, die unterschiedliche Konstrukte messen sollen, fielen niedrig aus, was für eine trennscharfe Erfassung der Merkmale spricht.

Zusätzlich wurde Messäquivalenz über kulturelle Kontexte hinweg nachgewiesen: Die Resultate zeigen, dass das Zortify-Verfahren auch über verschiedene europäische Länder hinweg vergleichbare Ergebnisse liefert. Dies spricht für seine interkulturelle Anwendbarkeit und unterstützt die Generalisierbarkeit der Ergebnisse im internationalen Personalkontext.

### 3.5 Kriteriumsvalidität

Die Kriteriumsvalidität wurde durch mehrere Studien belegt, die Zusammenhänge zwischen den durch Zortify erfassten Persönlichkeitsmerkmalen und externen Kriterien wie beruflicher Leistung und realen Auswahlentscheidungen untersuchten.

Eine besonders praxisnahe Überprüfung der prognostischen Validität lieferte die Studie von Kiesow-Berger et al. (2023). In diesem Feldexperiment bewarben sich 812 Gründer:innen um eine reale Förderung durch einen externen Investor. Alle Teilnehmenden absolvierten zuvor das Zortify-Assessment. Im Anschluss wurden Persönlichkeitsprofile jener Personen analysiert, die unabhängig vom Forschungsteam tatsächlich eine Förderung erhielten ( $n = 19$ ). Die Ergebnisse zeigen signifikante Unterschiede: Geförderte Personen wiesen signifikant höhere Werte in Optimismus und unternehmerischem Kapital insgesamt auf, sowie niedrigere Werte in subklinischer Psychopathie (Kiesow-Berger et al., 2023). Diese Resultate sprechen

dafür, dass die mit Zortify gemessenen Persönlichkeitsmerkmale relevante Prädiktoren für reale unternehmerische Förderentscheidungen darstellen.

Ein weiterer Nachweis der Kriteriumsvalidität ergibt sich aus der Analyse des Sprachverhaltens. In Übereinstimmung mit der wissenschaftlichen Literatur lassen sich charakteristische sprachliche Muster bestimmten Persönlichkeitsmerkmalen zuordnen – ein Ansatz, der insbesondere die Validität der NLP-basierten Komponenten des Verfahrens unterstützt. So verwenden extravertierte Personen nachweislich häufiger kollektive Pronomen wie „wir“ (Chen et al., 2020), ein Muster, das sich auch bei Zortify-Nutzer:innen mit hohen Extraversion-Scores zeigt. Ebenso wurde der erwartbare Zusammenhang (Edwards & Holtzman, 2017) zwischen hoher emotionaler Stabilität und geringerer Nutzung des Pronomens „Ich“ bestätigt. Offenheit für Erfahrungen korrelierte wie erwartet mit dem Gebrauch längerer Wörter. Auch für machiavellistische Tendenzen ergaben sich übereinstimmende Sprachmuster in Bezug auf Pronomengebrauch („wir“), was die diagnostische Validität der semantischen Analysekomponenten weiter unterstreicht.

Insgesamt zeigt sich damit ein konsistentes Bild: Das Zortify-Verfahren erreicht sowohl auf konzeptioneller als auch auf empirischer Ebene ein hohes Maß an Validität. Es misst etablierte Persönlichkeitsdimensionen theoriekonform, differenziert diese zuverlässig von anderen Konstrukten, zeigt bedeutsame Zusammenhänge mit externen Kriterien und bildet sprachliche Verhaltensmuster valide ab.

### **3.6 Normierung**

Für die individuelle Interpretation diagnostischer Ergebnisse ist eine angemessene Normierung unerlässlich. Zortify stützt sich auf eine umfangreiche europäische Normstichprobe mit über 16.000 Teilnehmenden, die aus unterschiedlichen Ländern und sozioökonomischen Kontexten stammen. Diese breite normative Datenbasis gewährleistet eine differenzierte Vergleichbarkeit der Ergebnisse innerhalb des europäischen Wirtschaftsraums; ein zentrales Kriterium für den praktischen Einsatz in international agierenden Organisationen.

### **3.7 Grenzen und Entwicklungsbedarf**

Trotz der überzeugenden Befunde zur psychometrischen Qualität von Zortify ist eine reflektierte Einordnung der methodischen Grundlagen und Limitationen wichtig, um das Potenzial des Verfahrens voll auszuschöpfen und zukünftige Entwicklungen gezielt voranzutreiben.

Ein Abgleich mit etablierten Richtlinien wie der DIN 33430 (Deutsches Institut für Normung, 2021) zeigt, dass Zortify eine solide empirische Basis liefert. So liegen zur Reliabilität Daten zur internen Konsistenz sowie zur zeitlichen Stabilität über kurze und lange Intervalle vor. Ergänzende Analysen können etwa zur Split-Half-Reliabilität oder zur Differenzierung nach Subgruppen die Evidenzbasis weiter stärken. Hinsichtlich der Validität wurden zentrale Aspekte wie Konstrukt-, diskriminante und prädiktive Validität sowie interkulturelle Messäquivalenz bereits überzeugend belegt. Ergänzend können Studien zur inkrementellen Validität und zur Vorhersage objektiver Leistungskennzahlen weitere Evidenz liefern.

Besonders relevant ist die methodologische Grundsatzfrage, wie KI-gestützte Diagnostikverfahren adäquat evaluiert werden können. In vielen Studien werden neue Verfahren primär anhand ihrer Korrelationen mit etablierten Selbstausskunftsinstrumenten validiert. Dieses Vorgehen kann jedoch problematisch sein: Wenn klassische Verfahren selbst methodische Schwächen aufweisen, etwa durch soziale Erwünschtheit oder limitierte Objektivität, besteht die Gefahr, dass innovative Verfahren wie Zortify systematisch unterschätzt werden. Diese Problematik wird auch in aktuellen Meta-Analysen (z. B. Naz et al., 2025; Bhandarkar et al., 2024) betont, die dafür plädieren, neue Validierungsstandards für KI-basierte Diagnostik zu entwickeln. Denkbar wären etwa stärker kriteriums- und verhaltensorientierte Validierungen oder der Aufbau konvergenter Evidenz über unterschiedliche Datenquellen hinweg (z. B. Sprache, Verhalten). Auch auf regulatorischer Ebene, etwa im Rahmen des EU AI Acts, könnte künftig noch stärker spezifiziert werden, wie psychologische KI-Systeme geprüft werden sollten, um sowohl wissenschaftlichen als auch ethischen Anforderungen zu genügen.

Insgesamt zeigt sich: Zortify steht für einen innovativen, empirisch fundierten und standardisierten Ansatz in der Persönlichkeitsdiagnostik. Die bisherigen Validierungen belegen eine hohe psychometrische Qualität. Zugleich ist der technologische und methodische Rahmen im Wandel, weshalb fortlaufende Forschung, Offenheit gegenüber Weiterentwicklungen und die Diskussion geeigneter Evaluationskriterien entscheidend sind, um das Potenzial KI-gestützter Diagnostik langfristig voll auszuschöpfen und verantwortungsvoll zu nutzen.

## **4. Zusammenfassung und Ausblick: Warum die Zeit reif ist für evidenzbasierte KI in der Personalauswahl**

Trotz der seit Jahren bekannten Schwächen klassischer Auswahlverfahren und der belegten Fehleranfälligkeit menschlicher Urteile zeigt sich die Personalauswahl in der Praxis weitgehend resistent gegenüber evidenzbasierten Verfahren und Innovationen. Verfahren mit unzureichenden psychometrischen Eigenschaften, etwa unstrukturierte Interviews oder die freie Interpretation von Lebensläufen, dominieren weiterhin das Auswahlgeschehen. Dieser Status quo besteht seltener aus Unkenntnis, häufiger aus institutioneller Trägheit, pragmatischem Kalkül und einer fortbestehenden Unterschätzung der Rolle systematischer Verzerrungen im menschlichen Urteil.

Gleichzeitig zeigt eine wachsende Zahl empirischer Studien, dass moderne, KI-gestützte Verfahren, insbesondere solche auf Basis natürlicher Sprachverarbeitung (NLP), das Potenzial haben, die Qualität von Auswahlentscheidungen substantiell zu verbessern. Sie bieten neue Möglichkeiten, Persönlichkeit differenzierter, valider und weniger verzerrt zu erfassen als klassische Fragebögen oder Interviews. So können sie besonders in Kontexten mit hohem Verfälschungspotenzial oder bei schwer zugänglichen Merkmalen eine wertvolle Ergänzung darstellen. Die psychometrische Prüfung des Zortify-Verfahrens liefert hierfür ein praxisnahes Beispiel: Reliabilität und Validität erreichen ein Niveau, das dem herkömmlicher Fragebogen-Instrumente der klassischen Testtheorie entspricht und zudem eine indirekte, nicht gezielt manipulierbare Verhaltenskomponente ergänzt. Zortify weist bessere Testgütekriterien auf als alle bislang veröffentlichten KI-Instrumente. Darüber hinaus ermöglichen evidenzbasierte KI-Verfahren wie Zortify eine ortsunabhängige, vollautomatisierte Durchführung mit minimalem Ressourceneinsatz und bieten insbesondere im Hinblick auf Skalierbarkeit und langfristige Kostenreduktion eine nachhaltige Alternative zu klassischen, personalintensiven Auswahlverfahren.

Gleichwohl gilt: Der Einsatz von KI in der Personalauswahl ist kein Selbstzweck. Der Grundsatz „es kommt darauf an“ bleibt zentral. Nicht jede KI-Anwendung ist sinnvoll, nicht jede valide. Die Forschung zeigt, dass insbesondere solche Verfahren erfolgreich sind, die psychologisch fundiert, empirisch überprüft und transparent in ihrer Funktionsweise sind. Ebenso müssen ethische Standards, Datenschutzrichtlinien und regulatorische Vorgaben konsequent mitgedacht und umgesetzt werden.

Ein zukunftsfähiger Umgang mit KI in der Personalauswahl verlangt daher zweierlei: Erstens die Bereitschaft, evidenzbasierte Verfahren ernst zu nehmen und tradierten Routinen den Rücken zu kehren. Assessment-Center stehen exemplarisch für die Kluft zwischen traditioneller Praxis und empirischer Evidenz: Sie sind kostenintensiv, schwer zu standardisieren und trotz ihres hohen Aufwands oft mit fraglicher diagnostischer Qualität verbunden. Dass sie dennoch weitverbreitet sind, zeigt die Trägheit diagnostischer Routinen. Zweitens erfordert es einen verantwortungsvollen Umgang mit technologischer Innovation: Auch KI-Verfahren dürfen nicht unkritisch übernommen werden, sondern müssen wissenschaftlich geprüft, kontextsensitiv eingesetzt und rechtlich verantwortungsvoll ausgestaltet werden.

KI-basierte Diagnostik bietet eine gerechtere, fundiertere und wirksamere Personalauswahl im Dienst von Fairness, Qualität und Effizienz. Angesichts der Verzerrungsanfälligkeit klassischer Verfahren und der zahlreichen Vorteile KI-gestützter Diagnostik ist es heute nicht nur klug, sondern geboten, moderne evidenzbasierte KI-Verfahren zu nutzen.

## Literaturverzeichnis

Armoneit, C., Schuler, H., & Hell, B. (2020). Nutzung, Validität, Praktikabilität und Akzeptanz psychologischer Personalauswahlverfahren in Deutschland 1985, 1993, 2007, 2020: Fortführung einer Trendstudie. *Zeitschrift für Arbeits- und Organisationspsychologie*, 64(2), 67–82. <https://doi.org/10.1026/0932-4089/a000311>

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1–26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, 14(4), 317–335. <https://doi.org/10.1111/j.1468-2389.2006.00354.x>

Bhandarkar, A., Wilson, R., Swarup, A., Webster, G. D., & Woodard, D. (2024). Bridging minds and machines: Unmasking the limits in text-based automatic personality recognition for enhanced psychology-AI synergy. *British Journal of Psychology*, 00, 1–23. <https://doi.org/10.1111/bjop.12755>

Chen, J., Qiu, L., & Ho, M.-H. R. (2020). A Meta-analysis of linguistic markers of extraversion: positive emotion and social process words. *Journal of Research in Personality*, 89, 104035. <https://doi.org/10.1016/j.jrp.2020.104035>

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72(3), 493–511. <https://doi.org/10.1037/0021-9010.72.3.493>

Hashemi-Motlagh, S. M., Rezvani, M. H., Khounsivash, M. (2025). AI methods for personality traits recognition: a systematic review. *Neurocomputing*, 640, 130301. <https://doi.org/10.1016/j.neucom.2025.130301>

Deutsches Institut für Normung. (2016). *DIN 33430: Eignungsdiagnostik*. Beuth Verlag.

Edwards, T., & Holtzmann, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68, 63–68. <http://dx.doi.org/10.1016/j.jrp.2017.02.005>

Fan, Y., Zhao, Y., Le, T. Q., Srikant, S. A., Fu, J., & Chen, Y. (2023). How well can an AI chatbot infer personality? A psychometric evaluation of language-based personality estimation. *Proceedings of the National Academy of Sciences*, 120(30), e2302337120. <https://doi.org/10.1037/apl0001082>

Hickman, L., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(2), 244–262. <https://dx.doi.org/10.1037/apl0000695>

Heidrink, M., & Feltes, F. (2025). *Vom Bauchgefühl zum Algorithmus: Warum KI-Verfahren in der Personalauswahl unverzichtbar werden* [Whitepaper]. Zortify. <https://zortify.com>

Kiesow-Berger, H., Fell, M., Philipp, F., & Steiner, E. (2023). KI-Unterstützung bei der digitalen Personalauswahl – Eine Fallstudie im Unternehmer-Kontext. In P. Stulle & R. T. Justenhoven (Hrsg.) *Personalauswahl 4.0* (pp. 163-177). Springer Gabler.

Koutsoumpis, G., Leicht-Deobald, U., & Li, C. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews (AVIs). *Journal of Applied Psychology*. Advance online publication. <https://doi.org/10.1016/j.chb.2023.108128>

Kowalski, C. M., Rogoza, R., Vernon, P. A., & Schermer, J. A. (2018). The Dark Triad and the self-presentation variables of socially desirable responding and self-monitoring. *Personality and Individual Differences*, 120, 234–237. <https://doi.org/10.1016/j.paid.2017.09.007>

Moreno, J. D., Martínez-Huertas, J. Á., Olmos, R., & García-Sancho, E. (2021). A meta-analysis of the validity of text-based personality detection using natural language processing. *Journal of Research in Personality*, 93, 104130. <https://doi.org/10.1016/j.paid.2021.110818>

Naz, A., Khan, H. U., Bukhari, A., Alshemaimri, B., Daud, A., & Ramzan, M. (2025). Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. *Artificial Intelligence Review*, 58, 239. <https://doi.org/10.1007/s10462-025-11245-3>

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934–952. <https://doi.org/10.1037/pspp0000020>

Rahe, C., & Rahe, T. M. (2017). *Nutzung von personaldiagnostischen Instrumenten bei der Auswahl von Führungskräften in deutschen Unternehmen*. Rahe Management Consultants.

Sackett, N., Zhang, P. R., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, 107(11), 2040–2068. <https://doi.org/10.1037/apl0000994>

Sikström, S., Valavičiūtė, I., & Kajonius, P. J. (2025). Personality in just a few words: Assessment using natural language processing. *Personality and Individual Differences*, 238, 113078. <https://doi.org/10.1016/j.paid.2025.113078>

Tu, Z., Zhang, Z., Zhang, W., Luo, F., & Bian, R. (2024). Using large language models to identify narcissism based on texts. *PsyArXiv*. <https://doi.org/10.2139/ssrn.4965442>

Van Iddekinge, C., Lievens, F., & Sackett, P. R. (2023). Personnel Selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology* 76(2), 651–686. <https://doi.org/10.1111/peps.12578>

Yarkoni, T. (2010). Personality in 100,000 words: a large scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44, 363–373. <https://doi.org/10.1016/j.jrp.2010.04.001>

Heidrink, M., & Feltes, F. (2025). *Vom Bauchgefühl zum Algorithmus: Warum KI-Verfahren in der Personalauswahl unverzichtbar werden* [Whitepaper]. Zortify. <https://zortify.com>

Zell, E., & Lesick, T. L. (2022). Big Five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, 90(4), 559–573. <https://doi.org/10.1111/jopy.12683>

Zhang, C., Lu, Z., & Wang, M. (2024). Can Large Language Models assess personality from asynchronous video interviews? A zero-shot evaluation of GPT-3.5 and GPT-4. *PsyArXiv Preprint*. <https://doi.org/10.1109/TAFFC.2024.3374875>

Heidrink, M., & Feltes, F. (2025). *Vom Bauchgefühl zum Algorithmus: Warum KI-Verfahren in der Personalauswahl unverzichtbar werden* [Whitepaper]. Zortify. <https://zortify.com>