Human Resource Management

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Can Legal and Professional Personnel Selection Principles be Met With Machine Learning (Artificial Intelligence)?

Michael A. Campion

Daniels School of Business, Purdue University, West Lafayette, Indiana, USA

Correspondence: Michael A. Campion (campionm@purdue.edu)

Received: 19 June 2025 | **Revised:** 23 June 2025 | **Accepted:** 10 July 2025

Funding: The author received no specific funding for this work.

Keywords: artificial intelligence | assessment | computer scoring | hiring | interviewing | machine learning | natural language processing | personnel selection | testing

ABSTRACT

The purpose of this article is primarily to evaluate whether machine learning (a form of artificial intelligence) can meet scientific, professional, and legal principles of personnel selection based on the rapidly accumulating research literature in Human Resource Management (HRM). It does so by addressing a series of questions in terms of what is known in the current research literature on these principles and related topics and by proposing a research agenda for what else needs to be studied. The review shows that there is enough current scientific evidence on the value of ML to support its use. ML tools are able to meet the basic principles of personnel selection and better meet many other very important principles. In addition, ML for personnel selection decisions is not a "black box" and can be understood and explained. It does not increase the legal risks from hiring, although it may require some additional steps in a few jurisdictions. There are uses of ML in selection that most organizations should be considering. Customized procedures will require good data, but generic vendor products may also serve some needs. Additional expertise may be required of HRM professionals, but not necessarily at a high level that would require new staffing or consultant expense. The availability of large language models (LLMs) may render unproctored remote assessments vulnerable to cheating, and all narrative candidate information provided in the future susceptible to AI-generated text, resulting in more questions than research answers at this stage. The article ends with practical and theoretical implications for the use of ML in selection.

1 | Introduction

Artificial intelligence includes machine learning (ML), which are algorithms that discover/learn relationships in data. The use of ML for personnel selection purposes has emerged as a major recent trend in Human Resource Management (HRM). Examples include scoring applications, interviews, and assessments (e.g., Campion et al. 2016; Campion, Campion, Johnson, et al. 2024; Liff et al. 2024; Sajjadiani et al. 2019; Studies 2–5 in Koenig et al. 2023), statistically combining scores to make predictions of job performance (e.g., Landers et al. 2023), and automating support activities like job analysis (e.g., Putka et al. 2023; Study 6 in Koenig et al. 2023). Many, if not most, testing vendors and contractors are advertising their use of ML and making promises of the benefits these days. Although some organizations have been early adopters, most are still waiting to decide. There are many questions to answer to determine adoption. The purpose of this article is specifically to evaluate whether ML can meet the professional and legal principles of personnel selection based on the scientific research literature in HRM. This research has only been available recently, but it is rapidly accumulating and is now sufficient to answer this question, along with some associated questions.

The initial literature in both research and practice publications was dominated by a fear of the unknown. Perhaps fueled by a highly publicized early court case involving Amazon (e.g., Dastin 2022; Kodiyan 2019), laypersons and many HRM professionals worried that the use of ML for hiring might result in perpetuating past discrimination by only hiring those who were like past employees demographically. Moreover, the lack of understanding as to how ML operated and the inability in some cases to examine exactly what it scored about candidates furthered the suspicion (e.g., Ajunwa 2020; Kelan 2024).

Considerable research has now accumulated to answer important questions by researchers and practitioners in HRM. The research is not yet definitive, but there is enough high-quality science to begin addressing many important issues and to identify the remaining questions. Specifically, there is considerable debate about whether ML is useful for personnel selection. The key focus of the article is to answer this question based on whether it can meet the scientific, professional, and legal requirements given the existing research. For example, there are technical challenges with respect to identifying the criteria used for validation, the algorithms to use, cross-validation, and changing data conditions that may reduce the validity over time (e.g., Allen et al. 2020; Goretzko and Israel 2022). Another key issue for both scientists who must justify the use of ML as well as HR managers who use it is whether ML can be understood or is a black box, which also bears on whether it can be legally defended. These are critical issues for HRM scientists and managers, given that ML will make important decisions about people, and hiring is dominated by legal as well as effectiveness issues. As stated by an early observer, ML "will only be truly accepted and trusted if explainability and transparency can be guaranteed" (Liem et al. 2018, 197). Other issues are what an HR researcher or manager should consider doing at this point, and are there important factors that should be determined or need more research. This includes whether there is enough evidence to adopt ML yet, what specific uses are best supported (e.g., scoring applications or interviews), and whether it depends on the specific context of the organization. A related concern is whether the availability of ML to candidates as well as hiring officials has an impact, such as if the widespread availability of Large Language Models like ChatGPT to the general public will influence hiring because candidates may use them to create application information or answer interview questions. On all topics, the article highlights the scientific issues that have yet to be resolved, thus providing a research agenda. The article is organized in terms of 10 questions to answer these issues.

1.1 | What Is the Current Scientific Evidence on ML in Personnel Selection?

The surge in recent studies was motivated in part to determine if ML is effective in terms of replacing a human rater and predicting important outcomes like job performance, but also largely due to its tremendous potential for improving efficiency by automating the hiring process. The result is that substantial evidence has been amassed demonstrating the potential benefits. Table 1 shows some of the most important findings to date based on the research published in the top scientific journals, such as *Journal of Applied Psychology* and *Personnel Psychology*, which have been the primary journals for reporting scientific discoveries on selection for the last 100 years and 70 years, respectively. The first study was published as recently as 2016, but nearly 20 have accumulated since. The table shows the focus or purpose of the studies and the key arguments made by the researcher as to why ML should work, followed by a brief description of the findings of example studies.

1.1.1 | Predictive Validity

The initial studies examined only the scoring of applications but found that ML could do as good a job as very highly trained human raters. This was shown by demonstrating that the ML scores correlated as highly with human ratings as human ratings correlate with each other (e.g., Campion et al. 2016). Soon, the research also explored scoring other candidate information such as answers to assessment and interview questions (e.g., Liff et al. 2024; Studies 2–5 in Koenig et al. 2023). This research also addressed whether ML scores can predict important outcomes like job performance and turnover, which are the usual criteria for validating hiring procedures.

1.1.2 | Adverse Impact

Another important outcome in personnel selection is whether there are gender and racial subgroup differences that might increase adverse impact (i.e., differences in passing rates between subgroups), potentially creating legal liability. Almost all the studies examined this and found that ML did not increase impact and might have reduced it slightly. One study found that using ML to score an applicant's past experience measured additional information on candidate skills that simultaneously increased both the prediction of job performance and reduced adverse impact (Campion, Campion, Johnson, et al. 2024). This is especially promising because increasing prediction (validity) and decreasing adverse impact have been historically viewed as not possible to obtain simultaneously and a necessary tradeoff, creating a dilemma for HRM (Ployhart and Holtz 2008).

1.1.3 | Measuring Personality

The ML applications above mostly measured candidate skills and experiences, but other research has explored their use for measuring personality. This research shows that ML scores can correlate with self-reported personality tests, but it has not yet been shown that these scores can predict job performance. Self-reported personality has a weak history of predicting job performance (Morgeson et al. 2007), but perhaps ML has the potential to improve prediction if it can avoid the distortions of self-report measures like faking and lack of self-awareness. Lastly, ML may be able to increase the efficiency of other time-consuming tasks related to selection, such as job analysis, by using the information in job descriptions or other existing job data to infer job requirements. It is important to observe that the most enabling aspect of ML for HRM purposes is its ability to score textual and narrative information from candidates and job postings.

The upshot is that ML is able to score applications, interviews, and other hiring information as well as a human recruiter, interviewer, or assessor. ML scores correlate as highly with humans

TABLE 1 | Current scientific evidence on machine learning for selection.

Research focus	Key arguments	Examples of studies
1. Scoring employment applications and resumes	• The use of Natural Language Processing (NLP), a type of ML for scoring the text in applications, is now feasible, which may allow the automation of a time-consuming and poorly conducted step in the hiring process	• Campion et al. (2016) scored application information and responses to questions on past accomplishments in the application to predict hiring ratings of professional candidates as well as a highly-trained human • Sajjadiani et al. (2019) scored application information of schoolteachers to predict job performance and turnover
2. Scoring answers to assessment questions	• Same argument as above, but to score open-ended written and oral answers to questions and assessment exercises • ML algorithms in addition to NLP may improve prediction	• Studies 2–5 in Koenig et al. (2023) scored various types of assessments across a range of hiring contexts, industries, and jobs to predict human hiring ratings, job performance, and turnover. Also, see Hickman et al. (2023)
3. Scoring answers to video interviews	• Same argument as above, but to score oral answers to remotely administered automated video interviews for prescreening • ML scores will have good psychometric (statistical) quality	• Liff et al. (2024) scored answers to automated video interviews to measure competencies across a range of jobs and organizations. The scores had good psychometric quality and predicted human hiring ratings and job performance
4. Not increasing diversity subgroup differences	• ML will not increase differences between gender and racial subgroups and may reduce them slightly, so it will not cause adverse impact (i.e., differences in passing rates between racial or gender subgroups), creating legal liability • Using ML algorithms to combine the hiring data into scores might help reduce subgroup differences	• None of the above studies found that machine learning increased adverse impact and sometimes it reduced impact slightly • Studies attempting to use ML algorithms to combine the data into scores to reduce subgroups differences have only reduced adverse impact slightly, but they also reduced validity slightly (e.g., Rottman et al. 2023; Studies 1–3 in Zhang et al. 2023) • However, Campion, Campion, Johnson, et al. (2024) found that using ML to score applicant past experience captured additional information on candidate skills that both increased the prediction of job performance and reduced adverse impact simultaneously
5. Measuring personality traits	• ML can also measure personality traits from candidate textual and other information, which may complement the measures of work experience and skills measured in previous studies	• Several studies have shown that ML can measure personality based on correlations with self-report personality tests (e.g., Fan et al. 2023; Hernandez and Nie 2023; Hickman et al. 2022, 2024). The studies did not examine job performance
6. Measuring job requirements	• ML can increase efficiency of other time-consuming tasks related to selection, specifically job analysis, by analyzing textual and other information to infer job requirements	• Several studies have identified the job requirements (knowledge, skills, and abilities) from job descriptions and other textual information on a wide range of jobs (e.g., Putka et al. 2023; Study 6 in Koenig et al. 2023)

as humans correlate with each other. Based on the studies conducted so far, it also predicts job performance as well as or better than existing hiring procedures, and it does not increase and may slightly decrease the adverse impact.

Despite these positive early findings of validity, ML may not be able to meet other important personnel selection principles (addressed in Questions 2 and 3 below), be difficult to understand (Question 4), create additional legal requirements (Question 5), not be applicable to all contexts (Questions 6 and 7), require some new expertise (Question 8), and still have many unknowns (Question 9). The implications for practice and theory are addressed at the end (Question 10).

2 | Can ML Meet the Basic Principles of Personnel Selection?

There are two foundational authorities that define the principles of personnel selection and are used in this article: the Uniform Guidelines on Employee Selection Procedures (published by the Equal Employment Opportunity Commission et al. 1978), and the Principles for the Validation and Use of Personnel Selection Procedures (published by the Society for Industrial and Organizational Psychology 2019).

Perhaps the most basic principles in the legal and professional hiring guidelines are whether the selection procedures are job related and reliably scored. Job relatedness is usually examined in one of two ways. One way to demonstrate criterion-related validity is to present statistical evidence, typically by showing a correlation between scores and a measure of job performance (e.g., supervisor ratings, productivity, or turnover reduction). Examples include predicting the probability of turnover among school teachers from work history (Sajjadiani et al. 2019), predicting supervisor ratings of job performance for managers and technical professionals from assessments (Studies 2–4 in Koenig et al. 2023), predicting turnover and job performance simultaneously for warehouse workers from assessments (Study 5 in Koenig et al. 2023), and predicting supervisor ratings of job performance of call center and maintenance employees from interviews (Liff et al. 2024). The magnitude of the prediction was comparable in size to that observed with mental ability employment tests, which is about 0.20 uncorrected and 0.40 corrected (Morgeson et al. 2007).

The other common way is to show content validity, usually by comparing what is scored and what is required on the job. Sometimes content validity is demonstrated by simulating the performance of job tasks. ML has proven to add unique value in this context by being able to score the narrative responses of candidates to open-ended questions in such simulations. Published examples include scoring constructed responses to assessment exercises (i.e., write-in answers to questions as opposed to multiple-choice questions) and scoring oral answers to questions in job simulations (Studies 2 and 4 in Koenig et al. 2023). Sometimes, content validity is demonstrated by linking what is scored with important prerequisite knowledge and skills required for the job. This requires knowing what the ML is measuring. Although ML is often criticized for being "black box" in that its inner-workings cannot be examined directly, this is not true of many types of ML, and even types that do not allow direct examination produce descriptions of the features scored that can be used for content validation (see next section for further information on topics related to validity).

The other most basic principle of personnel selection is reliable scoring. Based on the mathematics of psychometrics, if a measure is not reliable, it cannot predict job performance even if it is job related (e.g., Ghiselli et al. 1981). In plain language, if an assessment is not measuring something at least minimally well, it cannot predict anything else because the scores are mostly random errors. Reliability can be measured in many ways, with each providing somewhat different information. When examined, ML has been found to meet the same level of reliability as other hiring procedures like tests and interviews by being repeatable (test–retest reliability) and measuring a fairly homogeneous knowledge or skill (internal consistency reliability). As noted above, ML scores can also correlate with human raters as well as humans correlate with each other (interrater reliability of about 0.60; Campion et al. 2016). Moreover, ML is incapable of subjectivity and bias, which can plague human ratings, and thus, ML may be more reliable across candidates than human raters.

More research is needed on validity, especially whether ML might have greater validity than human raters when there is a large amount of candidate information or because of greater consistency. Some criteria might be better predicted by ML, such as when the criteria are categorical (e.g., when past decisions used for validation were hire, reject, hold, not considered, candidate withdrew.) or when you have many criteria to satisfy simultaneously (e.g., job performance, turnover, job satisfaction, time to hire).

3 | Can ML Better Meet Other Important Principles of Personnel Selection?

Aside from meeting the basic principles of job-relatedness and reliability as well as traditional hiring procedures, ML can better meet many of the other very important principles of personnel selection. Examples include the following topics, which are summarized in Table 2 along with some remaining future research questions.

3.1 | Making Better Statistical Predictions Based on the Same Data

Many forms of ML are more sophisticated prediction models, which can enhance the prediction possible compared to normal statistics, even with the same data. This is partly because it can find more nuanced relationships such as curvilinear effects. For example, an extremely high level of some personality traits may be as negative as extremely low levels (e.g., extraversion; Miller et al. 2001). Sometimes ML used to increase statistical prediction will be called "data mining" for its ability to seek out predictive relationships in the data by exploring every possible relationship. For example, some algorithms model the data as a series of decision trees wherein information is successively considered as decisions to incrementally improve the prediction of an outcome.

TABLE 2 | How machine learning can meet the principles of personnel selection.

Principle	Research evidence	Questions for future research
Q2. Ensuring validity (job relatedness)	• ML is able to score applications, interviews, and other hiring information as well as a human recruiter, interviewer, or assessor • ML predicts job performance as well or better than existing hiring procedures when examined based on the studies conducted to date • ML can demonstrate content validity by linking what is scored with important prerequisite knowledge and skills required by the jobs. The content it scores can either be directly examined or by feature variables it produces for interpretation • ML is especially useful for using job simulations to demonstrate content validity because it can score open-ended written or oral answers (i.e., constructed responses)	• Could ML score hiring information better than humans, such as when large amounts of candidate information must be considered or due to greater consistency and reliability? • Are some types of job performance criteria better predicted by AI, such as when the criteria are categorical (by using decision trees) or when predicting multiple criteria simultaneously (e.g., task performance and turnover as in Study 5 in Koenig et al. 2023)? • Should the content validity of ML be demonstrated in the same manner as other assessments, or are there better approaches that are more applicable or important to ML? • Are there aspects of simulation performance other than the content of the answers that can be better detected by AI, such as the pattern of responses or interactions?
Q2. Measuring reliably	• ML can usually meet the same level of reliability as other hiring procedures by being repeatable (test–retest reliability), measuring a homogeneous knowledge or skill (internal consistency reliability), and correlating with human raters as well as other humans (interrater reliability) based on current studies when examined	• Should reliability be measured in additional ways with ML (e.g., algorithmic generalizability; Study 1 in Koenig et al. 2023)? • Should the same level or types of reliability be expected given sparse data sets (i.e., many 0 values) from large numbers of predictors, which reduces internal consistency (Tausczik and Pennebaker 2010)? • What are the minimum requirements for reliable measurement with ML and are they the same as other assessments?
Q3.1. Improving statistical prediction	• Some types of ML are more sophisticated prediction models that can enhance prediction compared to normal statistics, even with the same data, by exploring every interaction and curvilinear effects ("data mining") • Decision tree models are especially useful when the predictors and criteria are dichotomies or categories • ML may also be advantageous when there is a large number of variables to score, but the sample size is small	• Will data mining lead to discovery or dustbowl empiricism (i.e., blind empiricism without theoretical understanding)? • Might some ML techniques have advantages over traditional statistical procedures in the context of selection (e.g., predicting decisions as opposed to scores or considering interactions between answers, skills, or credentials)?

(Continues)

TABLE 2 | (Continued)

Principle	Research evidence	Questions for future research
Q3.2. Not Increasing or Reducing adverse impact	• ML does not increase and may reduce adverse impact (i.e., differences in passing rates between gender and racial subgroups) • ML cannot make up for actual subgroup differences in knowledge, skills, or credentials. But if there is bias in that data that is not reflective of true subgroup differences, then the ML model will reflect that bias • ML algorithms to combine the data into scores to reduce subgroup differences and adjustments to the input datasets will have limited effect and will also reduce validity slightly	• How to determine if subgroup differences in past hiring data are due to true differences in skills or due to bias? • What new data can be measured by ML that can reduce subgroup differences (e.g., scoring work experiences as opposed to just test scores; Campion, Campion, Johnson, et al. 2024)? • Can ML models reduce adverse impact without reducing validity, giving more valid predictors less weight, or causing predictive bias (De Corte et al. 2007; Studies 1–3 in Zhang et al. 2023)?
Q3.3. Analyzing textual and narrative data	• A form of ML called Natural Language Processing (NLP) can analyze textual information or oral responses from candidates. A great amount of such data is collected in the hiring process, but often left unscored or are underutilized, that may reveal more selection information (e.g., applications, resumes, letters of reference) • ML can score interviews and answers to assessment questions as well as human interviewers	• What might the analysis of this previously unscored information reveal? For example, might letters of reference contain information that is useful for hiring (cf. Muchinsky 1979) based on what is said about the candidate as opposed to the recommendation? • Does interview structure (Campion et al. 1997) improve prediction with ML as it has done with human interviewers (Huffcutt and Arthur 1994), which was shown in one unpublished study recently (Campion, Campion, and Fogel 2024)? Can ML also improve validity with unstructured interviews? • Can "passively" scoring all selection information regardless of purpose for collecting the information predict hiring decisions and job performance (e.g., Campion, Campion, and Fogel 2024)?
Q3.4. Identifying new constructs	• Text analysis has helped identify and measure hundreds of new constructs in the past in other areas of research and could identify new KSAOs for selection • ML can automate text analysis (such as by using NLP), making it much more efficient than traditional qualitative text analysis approaches like manual content analysis or grounded theory	• What new KSAOs can ML identify that are relevant to selection? Will these new constructs be different from existing constructs and add value to prediction or understanding? • How do automated versus manual human text analysis results differ (e.g., Campion and Campion 2020, 2025)?

(Continues)

Principle	Research evidence	Questions for future research
Q3.5. Improving validity	• ML may be able to improve validity by measuring new constructs • ML may help resolve the validity-adverse impact tradeoff by increasing validity and reducing impact at the same time by measuring constructs possessed more equally among candidate subgroups • ML can facilitate job analysis to make it much easier to ensure validity (job relatedness)	• How much can validity be improved and how does it depend on the context? • Can ML be shown to resolve the validity-adverse impact dilemma in other situations than the single study in the literature (Campion, Campion, Johnson, et al. 2024)? • Can ML improve the quality of job analysis as well as enhance efficiency? • Will candidate use of AI-generated answers to assessment questions, interview answers, and other information in the application process reduce validity? • How can the use of AI-generated answers and information submitted by candidates be prevented?
Q3.6. Providing documentation	• ML is much more transparent than procedures that involve human judgment because you can control the information that it uses, and you can examine directly what it measures and how much weight it gives to each variable, unlike human judges	• What should be the essential documentation requirements for ML, and how do they differ from the documentation requirements for existing selection procedures (as required in the Uniform Guidelines and Principles)?
Q3.7. Increasing operational efficiency	• ML can score information to inform hiring decisions for a small fraction of the cost of using humans • ML can save time in the hiring process by enabling virtually instantiations scoring, allowing immediate employment offers • ML can make writing test and interview questions much easier, thus saving researcher time • ML can facilitate recruiting by finding candidates that match the job requirements for recruiters and suggesting possible jobs to candidates that apply for other jobs in the company, current employees, or candidates on the internet	• What are the demonstrable benefits of ML in terms of increasing financial and time efficiencies compared to the costs? • Are AI-generated test questions as valid as researcher-written questions? • What prompts should be used with LLMs to generate better test questions? • How valid is ML for matching candidates to jobs in terms of the possession of skills and the prediction of hiring decisions and job performance? • Can LLMs score assessments and other candidate hiring information as well as specially developed ML tools tailored to the organization, which would be much easier for those who do not have ML skills?
Q3.8. Ensuring consistency, objectivity, and fairness	• ML will score every candidate the exact same way and never fail to consider some information or interject other information (thus ensuring fair access and preventing measurement bias), and it is not capable of subjectivity (thus ensuring fair treatment)	• How can the benefits of ML in terms of consistency, objectivity, and fairness be communicated to increase trust among candidates, hiring managers, and society? • When is ML viewed as fairer? • When and why do candidates request to opt-out of ML scoring when required by law or allowed by the organization, and does it help their hiring probability? • Does ML scoring or administration help or hurt disabled candidates and why?

As a simplistic illustration in the prediction of turnover, a series of decisions might be (1) whether a candidate has a relevant type of experience, (2) whether that was only for one prior employer, and (3) whether the candidate left that company after a long period of time. Each yes answer decreases the likelihood of future turnover. Some of these models have many hundreds of decision trees with many branches each. The model learns the relevant information to consider from the pattern of prediction of turnover from data on past employees, and then uses that set of decision trees to make predictions about the potential turnover likelihood of future candidates. Decision tree models are common in machine learning, especially when the predictors and criteria are dichotomies or categories. Decision trees can also be used on continuous predictors or criteria as well, in which case they break the data into levels (e.g., high/low) to be examined individually in the search for optimal prediction. ML may also be advantageous when there is a large number of variables to score, but the sample size is small (Landers et al. 2023).

Although a great amount is already known about the value of ML for making better statistical predictions, future research should identify when there is an advantage in selection. For example, ML might be better than traditional statistical procedures because it can automatically explore for interactions (such as whether education can compensate for work experience) or curvilinear effects (such as whether there can be too little or too much education or work experience).

3.2 | Not Increasing and Possibly Reducing Adverse Impact

Adverse impact refers to differences in the percentage of candidates passing (passing rates) between gender and racial subgroups of candidates. The review of the research on the use of ML for hiring above showed that, despite some early negative publicity, the actual research evidence from more than a dozen studies published so far have shown that it does not increase adverse impact (i.e., lower hiring rates for women and minorities) and may decrease it slightly. However, an important caveat is that ML models will likely show the same subgroup differences in the data upon which it is trained. If there are subgroup differences in the skills or other job requirements, ML will not erase those. ML cannot make up for actual subgroup differences in knowledge, skills, or credentials. Those differences will show up using any valid hiring procedures. It will just not make them any larger, based on the literature to date. Moreover, ML is incapable of subjectivity and personal bias, unlike humans. As long as the data used for training a model does not contain bias, the model cannot either. But if there is bias in that data that is not reflective of true subgroup differences, then the ML model will reflect that bias. More research is needed to determine if subgroup differences in past data are due to true differences in skills or bias. The hypothesis to reject in that research should be that the past data do not have bias just because there are differences, as in the legal doctrine of "innocent until proven guilty," as opposed to assuming differences must be due to bias. Finally, algorithms can be used to weight combinations of tests to reduce subgroup differences in passing rates, but usually the reduction is small and it comes with a small cost of reduced validity. They also involve other tradeoffs such as giving the most weight to the less valid hiring procedures and the weighting may be sample-specific and unstable (e.g., De Corte et al. 2007; Song et al. 2017; Study 3 in Zhang et al. 2023). Moreover, all such procedures create some degree of prediction bias, meaning different statistical relationships between predictors and criteria across subgroups (Study 1 in Zhang et al. 2023).

There are many pressing questions for future research. Of most urgency is how to determine if subgroup differences in past hiring data are due to true differences in skills or due to bias because that data will be used to create the models. Less urgent, but the greatest opportunity for reducing impact while improving validity is if ML can help identify and score data that adds to the information currently considered but has smaller subgroup differences (e.g., types of work and other job-related live experience).

3.3 | Measuring Other Data, Especially Text and Candidate Narrative Responses

ML can analyze text data provided by candidates (in writing or orally) using a form of ML called Natural Language Processing (NLP). The hiring process in most organizations normally collects a great amount of narrative information from candidates, which is often left unscored but may reveal more selection information if systematically scored. Examples include applications, resumes, letters of interest from candidates, reference letters, and narrative answers to questions in applications. Other narrative information is usually scored, but at a high cost, such as interview answers and assessment exercise answers. The research reviewed above shows many examples of the value of ML for scoring all of these types of narrative information, including applications (e.g., Campion et al. 2016; Sajjadiani et al. 2019), assessment answers (e.g., Studies 2–5 in Koenig et al. 2023), and interviews (e.g., Liff et al. 2024).

Scoring narrative information may be the biggest opportunity afforded by ML in selection (Campion and Campion 2020, 2023) and a clear direction for future research. For example, could all the other narrative information collected on candidates that is left unscored or only considered informally increase prediction (e.g., letters of interest, letters of reference, personal statements, career objectives)? Another important research question is how we should collect narrative information from candidates to be scored by ML to improve prediction (e.g., should interview questions be structured, should applications be more detailed, should 1-page resumes be allowed?).

3.4 | Identifying New Constructs

The constructs of relevance to personnel selection are typically knowledge, skill, experience, or other human attributes that are related to job performance, turnover, or other outcomes of importance to organizations. One recent review of the literature showed that qualitative text analysis has helped identify and measure hundreds of new constructs in organizational behavior, strategy, and the broader management literature; historically, most of which are relevant to HRM (Campion and Campion 2025). A total of 99 articles in that review found new constructs specifically reflecting person characteristics that may be relevant to selection (e.g., knowledge, skills, abilities, or personality traits). Examples include skills (Nadkarni and Narayanan 2005), impression management in interviews (Waung et al. 2017), and leadership charisma (Bligh et al. 2004). Note that most of this prior research was not using AI, but instead was using qualitative (manual) coding of text data by the researcher. ML can automate text analysis (such as by using NLP) making it much more accessible and efficient than traditional qualitative text analysis approaches (e.g., content analysis or grounded theory). Future research on identifying new constructs should utilize ML for these reasons. On a side point, although ML can be used to measure facial features or paralinguistic cues, which would be new constructs measured in selection, the research literature to date suggests little value add in terms of improving prediction (e.g., Hickman et al. 2022, 2025), and the potential legal risks created by negative public reactions have resulted in few reputable vendors offering such products.

3.5 | Improving Validity

One important potential benefit of measuring new data and new constructs is the improved prediction of job outcomes such as performance and turnover. However, this is a provocative opportunity rather than a confirmed finding at this stage of the science. Most of the current studies in top scientific outlets in HRM have focused on whether ML works for personnel selection, how well, in what applications, and the other questions in this article, but not on this specific topic. Given how text analysis has been a primary method of discovering new constructs in the past management literature as reviewed in the previous section, this potential benefit seems very likely. In addition, ML has the potential to improve validity and reduce adverse impact at the same time by measuring constructs possessed more equally among minority subgroups as noted previously (Campion, Campion, Johnson, et al. 2024). Although this is the only existing study demonstrating that ML may help resolve the validity-impact tradeoff, the measurement of new constructs will likely be more effective than the many other techniques that are largely ineffective without some tradeoff (reviewed in Ployhart and Holtz 2008).

Demonstrating improved prediction by measuring new constructs is a ripe area for future research. A related area for future research is whether ML can improve the quality of job analysis aside from increasing efficiency, perhaps by better detecting the underlying skills required. A more troubling research question is whether the use of LLMs by candidates to answer assessment or interview questions or to describe work experiences will reduce validity and, if so, how can it be prevented or mitigated?

3.6 | Providing Documentation of Reasons for Hiring Decisions

ML creates much better transparency and documentation of reasons for hiring decisions than other selection procedures that involve human judgment because you can examine exactly what was measured. This is due to at least three reasons. First, the researcher can control the information that the ML uses to make predictions. You can only give it information that is job-related, such as the application, answers to questions, or other information collected as part of the hiring process. You can prevent giving it information that is not relevant, like non-job-related information. Examples include diversity-related information or visual appearance, personal impressions or knowledge of the candidates, or data the organization does not want it to know, such as where they went to school (e.g., Ivey league or minority-serving schools) or the name of the organizations they worked for as opposed to just what they did in their past jobs.

Second, most types of ML can allow you to examine directly what they measure, like the variables they are extracting with NLP from the language candidates use. This is even possible with so called “black-box” ML (that uses neural networks and transformers that have internal layers of analysis that cannot be observed directly) because they will produce an output of features they score in the form of descriptions (c.f. Chowdhury et al. 2023; Heidemann et al. 2024).

Third, they will indicate how much weight they are giving to each variable. All this may require more intensive study than some HRM managers are accustomed to doing, like reviewing many hundreds of variables, but it is possible and not beyond the capabilities of most professionals if they are dedicated to understanding. This is explained in more detail in the section below. An organization using ML should always insist on knowing what it measures, not only for content validation but also for due diligence and responsible use of ML (e.g., Campion et al. 2018; Alexander III et al. 2025). A weakness of current published research is not always reporting exactly what their ML is measuring in terms that can be interpreted by laypersons and the public.

3.7 | Increasing Operational Efficiency

This is the big advantage. ML can score information to inform hiring decisions at a small fraction of the cost of using humans. For example, one early study showed a saving of $200,000 annually in reducing assessor time in a limited use case (Campion et al. 2016), but it is easy to imagine much greater direct savings in other situations if used to replace more human effort or process larger numbers of candidates. The other big potential efficiency gain is in the time saving. ML can enable virtual instantiations scoring to make immediate employment offers, which is a major competitive advantage in tight labor markets. Future research should report a cost–benefit analysis as well as the validity and other outcomes to demonstrate the practical value of ML in terms of gaining efficiencies.

There are at least four future research questions that will help organizations decide if ML tools for personnel selection are cost-effective. Although these will depend on the organizational context and each organization must perform its own analysis, published research studies should describe how to conduct analyses and illustrate the potential savings. First, what are all the costs involved? These include the cost to develop or purchase the ML tool and the cost of recruiter time or other financial resources that might be saved. This should also include the cost to implement and maintain the system over time. Second, will the volumes of hiring make the cost savings greater than the cost of the ML tool? ML will likely be highly cost-effective for large-scale hiring, but what are the lower limits when the net saving goes from positive to negative? Third, are other outcomes of value to the organizations, such as faster decisions ? This may not depend as much on financial considerations as on other metrics like job offer acceptance rates and days to fill job openings. Fourth, will the quality of selection procedures be more valid, such that job performance is higher or turnover is lower, which may be difficult to measure financially but clearly has great value.

Future research should also provide guidance on the prompts to use with LLMs to generate better test questions and demonstrate that AI-generated test questions are as valid as researcher-written questions. In addition, future research should determine if LLMs can score candidate hiring information as well as specially developed ML tools, which would be much cheaper and easier to use by HRM staff. Finally, there is almost no objective, independent research in HRM evaluating how well ML can match candidates to jobs despite the widespread prevalence of these offerings by vendors.

3.8 | Promoting Consistency, Objectivity, and Fairness

Preventing discrimination in employment is the purpose of the Uniform Guidelines (Equal Employment Opportunity Commission et al. 1978, Section 1B), and the Principles describe the many definitions of fairness and bias because of their importance to good selection (Society for Industrial and Organizational Psychology 2019, 38–42). This benefit does not require a research study to demonstrate because it is true based on the nature of machine scoring. For example, it will score every candidate the exact same way and never fail to consider some information or interject other information (ensuring fair access and preventing measurement bias definitions of fairness), and it is not capable of subjectivity (ensuring fair treatment definition of fairness).

The topic of algorithmic bias has received a great amount of attention in the published literature. As noted at the start of this article, the initial reaction of the public and many researchers was distrust and skepticism. The most recent comprehensive review of the topic cited nearly 100 articles and other publications (Albaroudi et al. 2024). These studies show that bias is conceivable at many stages in the development of an ML selection procedure, ranging from the representativeness of the data collected, to the algorithm used, and to implementation (Albaroudi et al. 2024; Hunkenschroer and Luetge 2022; Kazim et al. 2021; Yanamala 2022). However, the literature is not conclusive as to the actual existence of algorithmic bias. Showing differences in outcomes, such as hiring rates, may be due to actual differences between subgroups on the predictors and criteria and not proof of bias by itself. Although subgroup differences may result in lower hiring rates, the use of a common regression line when there are subgroup differences will result in the overprediction of job performance (Society for Industrial and Organizational Psychology 2019, 40–41). One study in the ML literature observed that if bias exists, it may result in overestimation as well as underestimation of subgroup scores (Köchling and Wehner 2020). An important area for future research is a definitive answer to how often the phenomenon actually exists and the size and nature of the influence. Another important question is what to do if it exists. The solutions in the computer science literature focus mostly on ensuring the data are representative of the various subgroups of candidates and some analytical adjustments, but do not demonstrate that those solutions reflect bias or influence outcomes (Albaroudi et al. 2024). Of special note, most of the research on algorithmic bias has been in the recruiting uses of ML, not selection procedures, so it may be of less concern to the purpose of this article.

Future research should also explore how to communicate the benefits of ML to the public and other stakeholders to increase trust. It would also be of interest to know why candidates request to opt-out of ML scoring when this is allowed by law and does it help them. A related question is whether ML hurts disabled candidates because they may have uncommon backgrounds or helps them by making administration easier.

The important issue is whether there are any research questions that must be answered before an organization considers the use of ML for personnel selection purposes. The answer will depend on the specific context of the organization and the research question of most relevance, but there is enough supportive evidence that ML is worth considering for many organizational contexts.

4 | Can ML for Personnel Selection Decisions be Understood and Explained?

Concerns with explainability may be exaggerated and have solutions. As described briefly above, the content of what is scored can be examined with ML. How to see what is scored depends on the specific method of AI, but all have some way. Here are some common alternatives:

1. The ML algorithms that combine numeric data to make predictions can be examined just like any traditional statistical model because the researcher defines what variables are included, and virtually all techniques will show some metric of how much weight is given to each variable. Even those ML algorithms that are less familiar to selection researchers and professionals have indicators of the variables they measure and the weights they assign to the variables.
2. The variables scored with NLP can be examined directly. These are sequences of letters, single words, phrases, or combinations of words (called n-grams) that occur frequently together in the samples of text from candidates. Some forms of NLP will also identify combinations of words that commonly occur together in the same sentence or other unit of analysis, even if not adjacent. There may be a great number of variables scored with NLP, perhaps in the thousands, but they can be understood if the HR manager is willing to spend a couple of hours. This information can also be summarized by researchers or

subject matter experts so it can be easily communicated to others, such as hiring managers or candidates. For example, one study showed the words it scored to measure performance on a customer service assessment exercise (e.g., discount, apologize, delivery, resolve, contact, address; Study 2 in Koenig et al. 2023, Appendix B1), and another study showed how it measured leadership by the words and phrases it extracted reflecting descriptions of past accomplishments in leadership contexts (e.g., of a group, project, class, board, military; Campion et al. 2016, Table 2).

3. Even complex (so-called "black box") methods such as transformers that utilize neural networks produce indicators of the features they score in the form of descriptive labels and the statistical importance (weight) they receive. These features are mathematical composites of the variables scored that are created by the algorithm for the purpose of explaining what they are measuring. Some algorithms will also produce special visuals for depicting what is scored. A common example is SHAP Plots (Bordt and von Luxburg 2023), which are figures that simultaneously show the variables scored, the direction of their influence (positive or negative), and the strength of their influence on the total score or prediction probability. Many other visualization techniques for ML are available (for a recent review, see Chatzimparmpas et al. 2020). As an alternative, one hiring study showed how it measured reasons for past turnover by illustrating with many examples of candidate textual responses and the scores they received (Sajjadiani et al. 2019, Table 2).
4. Regardless of the ML method, nearly all will combine the variables into scores reflecting candidate quality, and often they will have multiple scores labeled for what they measure or predict, such as different competencies (e.g., adaptability, communication, developing others, problem solving; Liff et al. 2024). These scores are meant for interpretation as well as prediction.
5. Understanding can also be gained by examining relationships between ML scores and other variables. Such analyses include "construct validation" where meaning is inferred based on showing relationships with known measures of a construct. For example, the hiring studies in this article that attempted to measure personality with ML correlated the ML scores with self-report personality tests to determine whether the ML scores measured personality (e.g., Fan et al. 2023; Hickman et al. 2022, 2024). Relationships between ML scores and past hiring decisions by humans, ratings of candidate skills or quality, live interview scores, and job performance also bear on understanding by showing that the ML scores predict what they are intended to measure or reflect in some manner. All of the hiring studies cited in this article presented relationships with one or more of these variables.
6. The accuracy of ML scores also bears on explainability. ML often uses specialized metrics for reflecting accuracy, such as precision, recall/sensitivity, accuracy, and area under the curve (for a detailed explanation, see Campion and Campion 2023). They complement the traditional selection validity metrics, such as the correlation coefficient.

5 | Does ML Increase the Legal Risks From Hiring?

The legal risk of using ML is the same as any other selection procedures in terms of the need to demonstrate job-relatedness when adverse impact exists, according to federal law ever since the Civil Rights Act of 1964 (Equal Employment Opportunity Commission et al. 1978). However, some state and city laws have imposed new legal requirements, the most noteworthy of which are the Illinois Artificial Intelligence Video Interview Act (2020) and the New York City Automated Employment Decision Tools (AEDT) local law.

The Illinois law only applies to employers in Illinois who are using automated video interviews. It requires informing candidates that ML is used, providing information on what it measures, obtaining consent or opting out, not sharing the videos, and erasing them after a period of time. If the employer relies solely on the ML to make decisions on candidates, it must report demographics of passers and failures to the state government annually.

The NYC law only applies to employers in NYC and only when the tool is used to "substantially assist or replace discretionary decision-making processes." However, the definition of an automated employment decision tool is very broad. It requires posting information on what the tool measures, but only provides information on the data used if requested in writing. Importantly, the law requires a yearly bias audit performed by an independent auditor that reports the demographics of passers and failures. The recently passed Colorado Artificial Intelligence Act (CAIA), which will take effect in 2026, has similar applicability and requirements.

Other laws are pending or evolving at the time of this article, and legal advisors should be consulted, but the key concerns in the experience of the author can be addressed as follows. First, informing candidates that ML is used is not difficult and does not require great detail, and requests to opt out do not require more work than existing procedures or require that those candidates must be considered first. The review of those candidates will not be as fast because the purpose of the automated tool is often to speed up the process, which should be communicated to those candidates. Second, as in the NYC law, these laws may only apply when the ML makes totally automated decisions and may not apply when there is human overview, which is usually the case or can be done. Third, if the demographics of candidates hired must be reported, they should only show the influence of AI, not whether there are differences in the procedures being replaced, although legal views on this differ. If there are such differences, ensuring human oversight and avoiding the need to report may be the best solution.

Perhaps the biggest surprise is how few cases there have been with ML so far. In fact, future court cases may allege that an organization should have used ML as opposed to human judgment because ML cannot be subjective and intentionally biased.

6 | Are There Applications of ML That Most Organizations Should be Considering?

Described below are some of the applications with enough supporting evidence at this time, such that they should be considered. They are listed in Table 3 along with some implementation considerations.

6.1 | Scoring Employment Applications and Resumes

This was the first widely recognized use of ML for selection because it could score textual information, plus the large numbers of candidates at this stage of hiring and the cost of human review made it appealing to employers. The initial supportive research was also in this context (e.g., Campion et al. 2016; Campion, Campion, Johnson, et al. 2024; Sajjadiani et al. 2019). Most vendors include this in their products. This potential use of ML is applicable to most organizations.

6.2 | Scoring Automated Prescreening Interviews

This is also a widely recognized use of ML for selection, again because it can score textual information and the frequent use of prescreening interviews. The additional advantage of video interviews is that recruiters and hiring managers can view the recorded interviews of high-scoring candidates to evaluate themselves. The best supporting research evidence in HRM has been recent (e.g., Liff et al. 2024). Video interviews have been a commonly available vendor product for several years now, and they are increasingly offering automated scoring. This potential use of ML is applicable to many organizations.

6.3 | Automating the Administration of Other Assessments

This application of ML may only be of value to organizations that administer assessments to large numbers of candidates. As illustrated by some studies cited previously (e.g., Studies 2–4 in Koenig et al. 2023), AI's ability to score verbal answers to assessment questions, either written or spoken by candidates, can be done as well as a human scorer and much more efficiently by ML. This can be a great advantage in high-volume hiring contexts. A related advantage for facilitating administration is using ML to automate remote proctoring to reduce cheating. Such systems monitor for abnormalities in candidate behavior that may indicate cheating while taking tests remotely (e.g., eye movements off screen, talking to others, background noises, other people in the room, unusual keystrokes) that are ML-enabled additions to the other common features of remote proctoring that reduce cheating like candidate authentication and lockdown browsers (so they cannot go look up answers). Although no existing studies evaluating ML for this purpose have been published in HRM journals as of the time of this article, it has been discussed in the educational literature (e.g., Nigam et al. 2021).

6.4 | Combining Scores From Hiring Procedures to Increase Prediction and Not Increase Adverse Impact

ML can be used in the form of algorithms to augment or replace normal statistical procedures to create two potential advantages. First, it may improve prediction. The improvement may not be huge, but it may be worthwhile in some circumstances, as described above. Second, it will not likely increase adverse impact based on the many existing studies cited previously and may reduce impact by avoiding subjectivity, but this comes with potential caveats. Some methods modify the datasets to reduce the predicted differences in scores, and others adjust the weighting, but they have limitations: the reduction of impact is small, involves tradeoffs, reduces validity a small amount, and creates prediction bias (Studies 1–3 in Zhang et al. 2023). More critically, ML could be used to "adjust" scores, such as normalizing (equating) scores between subgroups. Such adjustments are explicitly illegal according to the Civil Rights Act of 1991, where it is called "within-group norming." It eliminates adverse impact by selecting the proper numbers of each subgroup to match the applicant flow or labor market, but it is illegal because it considers protected information, such as race or gender, in the employment decision. This mistake is more likely to be committed by information technology professionals because of their unawareness of employment discrimination laws. Moreover, they are more likely to be the vendors offering ML products. Organizations hiring ML vendors should ensure they have technically trained HRM professionals involved, either on the vendor's staff or the organization's staff.

6.5 | Creating Test and Interview Questions

The recent widespread availability of Large Language Models (LLMs) like ChatGPT is having a big impact on hiring at the time of writing this article. Because they can create text based on prompts, they are called generative AI. Due to their recency, few scientific articles in top HRM journals have been published, but they are soon to come because many researchers are working on this and presenting their studies at conferences, which usually precede journal publications. Their public availability for free or low cost and their ease of use by laypersons are certain to lead to many uses. At least two uses by organizations are likely. First, they will make creating tests and interview questions much easier, which should reduce costs and save time for employers. This is valuable for creating alternative forms of tests so repeat candidates will not get the same test, creating large numbers of questions for remote administration so they can be regularly replaced (in case candidates copy and share questions), and creating tests from job descriptions in a matter of hours rather than weeks of effort. In one of the few published studies, the researchers used an LLM to create personality test questions that demonstrated comparable psychometric properties (e.g., reliability) to the original test (Hernandez and Nie 2023). Second, researchers are currently exploring the use of LLMs for scoring assessments as an easier alternative than using more complex ML because

TABLE 3 | Applications of ML that most organizations should be considering.

Application	Implementation considerations
6.1. Scoring employment applications and resumes	• One of the most common uses • Will be limited by the quality and thoroughness of the information collected on the application. Reliance on resumes will be deficient because they are not comprehensive and only contain selected information • If the algorithm is scored based on past hiring decisions, the validity of past decisions must be adequate • The accuracy of vendor products not trained specifically for the organization should be evaluated and not assumed
6.2. Scoring automated prescreening interviews	• Requires some consistency in the interview questions used across candidates and time. • Must train to the interview questions used by the organization or use the questions provided by the vendor • If trained to the organization, the validity of past interview ratings must be adequate • The accuracy of generic vendor products not specifically trained for the organization should be evaluated and not assumed
6.3. Automating the administration of other assessments	• These uses of ML would be unique to the specific organization and probably limited to high-volume selection contexts to be worthwhile • Unproctored remotely administered assessments are threatened by applicant cheating and test insecurity, and thus not advised without precautions. ML products for remote proctoring are being developed, but not yet evaluated in the HRM research literature • Some assessments are more vulnerable to cheating, such as knowledge tests with correct answers that can be looked up, as opposed to personality tests where correct answers are more obvious and the main concern is limitations due to self-report measurement
6.4. Combining scores from hiring procedures to increase prediction and not increase adverse impact	• Would be customized to the organization's assessments and hiring context, but the cost is very low • Main limitation is the availability of adequate sample sizes to determine the optimal algorithm • Should expect ML not to increase adverse impact, but do not expect it to decrease impact. Be skeptical of promises to the contrary and learn exactly how decreases are achieved before adopting
6.5. Creating test and interview questions	• The quality of such questions should be confirmed by subject matter expert review and possibly also psychometric analyses • Companies should require lower cost for test development from vendors or in-house staff due to these efficiencies • Use of LLMs for scoring by laypersons is not recommended at this stage of the research knowledge

(Continues)

TABLE 3 | (Continued)

Application	Implementation considerations
6.4. Analyzing jobs and job requirements	• ML to facilitate job analysis information will ultimately make this burdensome task that is often not conducted much easier • Do not rely solely on such tools. Use as supplementary information as part of the job analysis • Do not assume the accuracy of the information produced without subject matter expert review • Local job analysis information and validated archival sources such as O*NET should still be used primarily
6.7. Assisting with recruiting	• One of the most common uses • There is a risk of overreliance on such tools because they are integrated into the Applicant Tracking Systems (ATS), thus the recruiter is required to use the process • The diversity of candidates sourced by ML outreach tools should be monitored and possibly supplemented with special targeted recruiting to ensure diversity when needed

they can be used by those with less technical skill. Whether they work as well as ML models developed for the specific purpose is unlikely but to be determined. Those studies are in progress or not yet published, as noted.

6.6 | Analyzing Jobs and Job Requirements

ML can also help with tasks that support selection. Job analysis is one necessary support activity because it identifies the job requirements. Job analysis in the future will probably always be easier with ML. Research has shown that specially developed ML can create in-depth identification of knowledge, skills, abilities, and other characteristics (KSAOs) based on job description information by using public archives of job analysis information such as the US Department of Labor's Occupational Information Network (O*NET, https://www.onetonline.org/O*NET; Putka et al. 2023) or an organization's unique job analysis information (Study 6 in Koenig et al. 2023). LLMs will even write a draft job description or assist in writing task statements and identify initial job requirements, thus making basic job analysis activities much easier.

6.7 | Assisting With Recruiting

Another support activity is recruiting. ML can potentially assist by identifying candidates that best match the job requirements to make prescreening easier. They use NLP to analyze both the job descriptions and the candidates' applications or resumes, and then they compare the features extracted using a matching algorithm. ML can also be used to find potential candidates and encourage them to apply, such as applicants who applied for other jobs in the organization who match the job, current employees in the organization, or candidates on the web who have not applied to the organization. These uses of ML are far superior to former methods that relied on simple keyword matching because they employ more advanced ML techniques. A large number of studies on this topic have accumulated that evaluated the effectiveness and refined the techniques in the ML literature (e.g., Garg et al. 2022; Jirjees et al. 2025; Pessach et al. 2020; Reddy et al. 2020; Surya et al. 2024), while there is currently only basic awareness of the potential value in the HRM literature (e.g., Laumer et al. 2022). However, many vendors of HRM information technology and applicant tracking systems offer these services. There are potential legal and fairness-related concerns, even though this use of ML is for recruiting rather than selection. For example, if the use of the tool results in an adverse impact, the organization would be responsible for justifying the job-relatedness, and validation studies are rarely conducted on such tools. This is why most vendors recommend that the scores or matching process only be used to identify candidates who might be a better fit to review first, and not used to make any selection decisions. That reduces the likelihood that the tool will be singled out as the source of the impact, but the organization may still have to defend the entire process, of which the tool is a part. This concern applies to both candidates who apply and also to candidates for other jobs and employees who are solicited to apply, thus creating greater liability should discrimination be demonstrated. There is also a large literature on the potential for algorithmic bias, which refers to models that identify candidates more from some subgroups due to being trained on inaccurate or biased data. As reviewed in Section 3.8 above, this literature is dominated much more by speculation than data, and the procedures to reduce the differences have uncertain effectiveness. Also, as noted previously with reference to ML for selection procedures, ML for recruiting cannot erase true subgroup differences in the possession of the underlying job requirements.

7 | Are There Contingencies or Common Difficulties With Applying AI?

There are some common difficulties. First, there may be inadequate data available to train the ML model. There are at least two primary problems here. For one, fairly large samples are needed to train ML models. Many hundreds at a minimum, but many thousands are necessary for accurate measurement. Some studies found a minimum of 500 (e.g., Campion et al. 2016; Ramineni and Williamson 2013), but it depends on the number of features scored in the model (Landers et al. 2023). For another reason, depending on the intended use, the organization may not have collected the data needed to train the model. If ML is to be used to score applications, many organizations may only have collected brief information on application forms or only resumes. Such information lacks the details needed to evaluate candidate skills (e.g., descriptions of past job responsibilities, classes taken) or to evaluate the entire work history (e.g., not reporting all jobs). Resumes tend to be short (e.g., 1-page) with only employment highlights and not thorough documentation of work history. Preferably, fairly detailed applications should be required, as opposed to resumes only, that ask the candidate to describe past jobs, duties, education, and other credentials in some detail, perhaps including descriptions of accomplishments, unique skills, and other job-relevant experiences and activities. If ML is used for interview scoring, most organizations have not collected recordings of past interview answers to their questions. Therefore, the use of ML will require collecting that information for a period of time to train the model, or the organization will have to use the questions that the vendor used to train its model.

Second, and a related difficulty, is that only poor-quality data may be available to train the ML model. If an organization is using ML to reduce the HR resources needed to score the information in its applications or interviews, the scores or selection decisions on past candidates will usually be used to serve as the criteria against which the model is trained. However, such past decisions may be of very low quality. They may not accurately reflect the skills of candidates. For example, applications or resumes of past candidates have often been evaluated poorly, such as skimming 100 applications to pull out a handful of the best to interview. Although the resumes of those interviewed may be studied closely, the other 95 have not, and the models are trained based on all the candidates. In addition, often recruiters or hiring managers do not agree with each other on the best candidates, and thus, the data on past selections has low reliability. That statistically limits whether those scores can even be predicted by an ML model, as noted earlier (Ghiselli et al. 1981). In such contexts, an organization might collect better data for a time by using a more structured and thorough evaluation process to obtain the data to develop the model.

As an alternative, an organization could instead use a model based on the content of what it scores rather than its statistical prediction. One easy method for doing this is to use dictionaries to measure words that reflect the attributes to be measured (e.g., skills, personality traits, motivations), which is a very simple type of NLP. The advantage is that these dictionaries can be developed deductively without the need for data on past candidates, and they require only normal statistical skills and not advanced ML knowledge. Emerging research shows they can have a surprising level of validity, approaching that of more complex ML procedures (Campion, Campion, and Fogel 2024). Finally, other problems with the quality of past data also include no measures of job performance or poor-quality measures (like typical appraisal ratings), which limit an organization's ability to show that the ML model predicts actual job success, and there may be unknown bias in the past hiring decisions, which cannot be detected by simple subgroup differences because they may reflect real differences in skills.

Third, there may be too many complexities in implementation. These are all the operational issues with making a new process work. The first issue is deciding when and how the ML scores will be used in the hiring process. This includes the stage of the hiring process (e.g., prescreening or final selection), use of scores (e.g., cutting scores or rank order), and amount of input to the decisions (e.g., minimum hurdle, combine with other scores, or factor to consider as needed). There is also a myriad of information technology issues, such as capturing the data electronically, moving the data from the point of collection to the ML model to score, sending the scores to decision makers, and systems integration with vendors when used. Finally, there are data security, data privacy, data retention, and other common issues. All these complexities in implementation could mount up to making the expected efficiency gain not worthwhile.

Fourth, a related concern is the cost–benefit for small employers or limited applications. The potential improved efficiency may not be worth the increased complexity in small-use situations. There is a risk that ML may become a fad and be adopted based on the perceived pressure to modernize or impress senior management with HRM's innovativeness when it may not be worthwhile. The determination should be based on a consideration of both direct and indirect costs compared to the benefits of increased efficiency and quality.

These caveats are not meant to be discouraging. Most organizations have adequate data, even if it is not perfect. Organizations may also be able to collect better data in a short period of time, such as a few months or after one hiring wave. It is important to note that some vendor products may not require having your own data to train an ML model for your organization. For example, they may have an application/resume scoring product, an interview scoring product, or a product for facilitating recruiting that would fit your organization's needs adequately. The early uses of ML were primarily by organizations that developed their own custom systems, but vendors are coming out with many generic products because of the high demand in the market. Using a vendor product will not predict as well as a model developed for your organization, so it will create a tradeoff between the quality of prediction and the cost of a turnkey solution, but that may be a reasonable decision in many contexts, such as smaller organizations or fewer jobs or candidates.

8 | What Expertise Does ML Require?

The unbridled use of ML by amateurs is not recommended. Large Language Models (LLMs) such as ChatGPT are widely available and can provide some assistance in personnel selection that can be used by almost everyone. For example, a recruiter could use ChatGPT to score an inputted interview answer. It will provide a score, but will probably not be accurate enough to replace human judgment. As such, despite the promise, there is still much to be determined before the laypersons can be recommended for use.

There is an entire science around prompt engineering, and the value of LLMs for scoring is one where the prompts matter (e.g., Korzynski et al. 2023; Marvin et al. 2023). That is currently an area where much research is being conducted.

Nevertheless, hiring new staff or expensive consultants may not be necessary. If you have selection researchers or similar technical professionals in your organization now, they may be able to learn to utilize some ML techniques with their skills. New hires from graduate college programs may have these skills because such training is increasingly being incorporated into curricula.

The critical skills for developing and implementing ML in HRM include the following, and perhaps others:

First, a basic level of understanding of ML is necessary. A high level may not be needed, and this is not necessarily the most important skill. It depends on the role of the staff and the needs of the context. If the development of ML tools is internal, obviously, more skill is needed. However, technically or statistically oriented staff may be able to readily learn how to use basic ML programs. The most commonly used software (Python) is publicly available, with many online resources (e.g., code for various purposes) and forums for asking other users for help. LLMs like ChatGPT will also provide the Python code when prompted correctly. Some types of AI, such as dictionaries, can be used without any ML skills beyond statistics, and some software programs are very user-friendly and familiar to selection researchers (e.g., IBM SPSS Modeler. https://www.ibm.com/docs/en/spss-modeler/saas?topic=guide-about-spss-modeler, and for the text analysis feature, see: https://www.ibm.com/docs/en/spss-modeler/saas?topic=help-about-spss-modeler-text-analytics).

Organizations will commonly use vendors to provide ML services. In those situations, the level of skill needed of internal staff is only being able to ask the right questions and understand the answers at a basic level. Many organizations will hire a consultant to review the vendor's product and the application to provide an outside verification of the quality of the product and whether it is job-related, and also sometimes help translate the technology into explanations that can be understood by management and legal advisors. This should not be a large financial commitment because the necessary work can be conducted quickly. The tasks only include a review of vendor documentation, the technical details of the ML tool sufficiently to know what it measures, the job requirements, and a report.

Second, personnel selection and HRM knowledge are more important. Although many small organizations use only very basic selection principles and techniques (e.g., an application and interview), large and progressive organizations use more sophisticated concepts and techniques. Regardless of the tools used, this requires a range of skills such as the psychological science of individual differences, the measurement of human attributes, psychometrics (statistical quality analysis), the research literature in selection, complex statistics, and the practice of selection in organizations. It is important to recognize that vendors of ML may be primarily or only information technology experts and not selection or HRM experts. They may not have this knowledge beyond basic awareness. The organization is ultimately responsible for the selection procedures it uses. An organization cannot rely on the assurances of vendors about the job-relatedness (validity) of its products should they be challenged legally, as explicitly stated in the government employment selection guidelines (Equal Employment Opportunity Commission et al. 1978, Section 9A). Even if a vendor claims to have these skills on its staff, an organization is wise to have its own internal expertise or use an outside consultant briefly for this purpose. Experts in the science of selection are most commonly Industrial and Organizational Psychologists.

Third, basic employment law knowledge remains important. Employment decisions are litigated more than any other topic in HRM. Those involved in selection must be well-versed in discrimination laws, maintaining compliance, avoiding problems, how litigation occurs in selection, and especially the legal defense of job-relatedness (validation). Selection experts will usually have this knowledge because it is fundamental to the practice. However, vendors usually will not have in-depth knowledge of this area, and the organization is responsible for its own compliance.

In summary, some new expertise is required, but that may not require hiring new employees or engaging in long-term consulting engagements. Most of these knowledge areas are probably already possessed by the current selection staff or can be easily acquired. An organization can rely partially but not totally on the vendors for this knowledge and must assume responsibility for evaluating the quality of the ML tools used. The necessary knowledge and skills should not be an impediment to the adoption of ML for selection.

9 | What Is Going to be the Effect of Large Language Models (LLMs) on Personnel Selection?

At the time of writing this article, an important question is the effects of the potential use of AI-generated answers by candidates on the accuracy of information provided on applications, of answers to unproctored interviews (e.g., automated video interviews), and of other textual information collected at the time of hire. In particular, ChatGPT can create credible resumes, information in applications (e.g., descriptions of work experience or skills, letters of interest, or almost any other textual information), answers to test questions that require a narrative response, and answers to interview questions. This creates a critical vulnerability to the veracity of unproctored assessments. At the time of this writing, there are many more questions than answers. For example, will remote unproctored assessment be usable in the future, or will remote proctoring or other procedures be needed? What procedures are effective in preventing cheating (e.g., lockdown browsers, video monitoring services)? Are warnings not to use LLMs to create answers to application questions effective? Is detection software accurate enough to be used, and should such candidates be rejected if they use AI-generated answers? Will candidates simply modify their answers sufficiently to escape detection? Are LLMs no worse than other writing tools like spelling and grammar checkers, and candidates will only use them to suggest ideas, to create first drafts that they will modify, or to improve their writing? Do LLMs increase the chances of getting hired?

In the experience of the author, those using AI-generated answers are weaker candidates who do not score well enough to get hired. Their answers tend to be generic and lack the nuanced content to get high scores. However, that was for highly skilled jobs with low selection ratios. For low-skill jobs, LLM answers may be sufficient and could mask major weaknesses like a lack of basic language skills. Because LLMs work in part by statistically estimating the next most probable word in a sentence or phrase, they read very smoothly, often better than real human writing, and thus cannot be detected by reading the answers. Finally, will LLMs increase the likelihood of actual lying about credentials and skills, given that they may suggest credible false information the candidate may adopt, making follow-up verification more necessary? Research finds that more capable candidates are less likely to fake, but when they do, higher mental ability makes them more effective at faking (Levashina et al. 2009).

10 | What Are the Practical and Theoretical Implications for the Use of ML in Selection?

The answers to the questions posed in this article lead to several practical implications for HRM professionals who are considering the adoption of ML tools for selection.

1. There is enough scientific evidence of the benefits to demonstrate its value. ML can make decisions as well as humans and has many capabilities we never had before, such as analyzing text to enhance selection. The practical implication is that there is enough evidence to consider using ML.
2. ML can meet the basic principles of personnel selection, such as job relatedness and reliability. It can also better meet some other very important principles such as better statistical predictions, not increasing or reducing adverse impact, measuring other data, especially text and candidate narrative responses, identifying new constructs (e.g., knowledge, skill, experience, or other human attributes), improving validity by measuring new constructs, providing transparency and documentation of reasons for hiring decisions, increasing operational efficiency, and promoting consistency, objectivity, and fairness. The practical implication is that ML will better meet some of the most important principles of personnel selection than current hiring procedures.
3. ML is not necessarily a "black box" and what it measures and how can be readily understood, although it may not be obvious with only a cursory examination. The practical implication is that it might take a little more effort and thought on the part of the HRM manager or professional to understand what a specific ML tool does.
4. Its primary legal risks are the same as those of all other selection procedures, such as the need to demonstrate job relatedness if there is adverse impact. The current evidence suggests that it will not exacerbate the adverse impact and may well reduce it by measuring new constructs and avoiding subjectivity. There are new laws in some jurisdictions, and more are likely to come, but they can be met and are not unduly burdensome. The practical implication is that the new laws should not be viewed as an impediment to adoption.
5. There are some applications of ML that most organizations should consider such as scoring applications, scoring automated interviews, facilitating administration, creating assessment questions, perhaps improving prediction, conducting job analysis, and recruiting. The practical implication is that there are many uses for common purposes that most organizations would find valuable.
6. If you want to develop your own ML tool tailored to your organization's needs, there may be some difficulties to overcome, such as the adequacy and quality of data available to train the ML model. The practical implication is that there are some potential difficulties with developing an ML tool internally, but most organizations have or can collect such information, and the benefits of a custom tool might be much greater. If not, vendor offerings may be an alternative. Generic vendor products are rapidly emerging that may meet your needs.
7. Some level of knowledge of ML will be needed, but the level depends on the situation, with higher levels if you intend to develop your own tools and lower levels if you intend to use a vendor, in which case you only need enough knowledge to ask the right questions. Selection and HRM knowledge are more important. Vendors are usually only information technology experts and not selection or HRM experts, and an organization cannot rely on their assurances about the job-relatedness (validity) of the products. Employment law knowledge is also important, which vendors usually will not have at an in-depth level, and the organization is responsible for compliance. The practical implication is that new expertise is necessary, but that may not require hiring new employees or engaging in long term consulting engagements.
8. The widespread availability and increased use of LLMs may render remotely administered unproctored assessments too vulnerable to candidate cheating to be used in the future without some additional preventative procedures such as remote proctoring. This is especially the case with tests that have correct answers that depend on knowledge or skill and perhaps automated interview answers. Even normal application information is vulnerable to AI-generated answers, including resumes, descriptions of experience and skills, letters of interest or reference, and any other textual or narrative information. At this stage, there are many questions and few answers, and future research will be needed. The practical implication is that unproctored remote assessments should not be used in the future without preventative procedures, and all candidate information is vulnerable and should be checked for AI-generated text.

There are many potential theoretical implications on whether ML can meet the scientific principles of personnel selection based on the research discovered to date and the remaining specific research questions in Table 2. Here are several of the most obvious. First, the construct domain relevant to personnel selection is likely to expand based on the ability of ML to analyze text, which has been the source of discovering many new constructs in management research historically, and on the ability of ML to discover new relationships in data by exploring interactions and curvilinear effects, fitting models better to the data, and examining all possible relationships (data mining). Understanding more of the construct domain relevant to selection would have theoretical implications for our understanding of individual differences, such as the models of job performance (e.g., Campbell and Wiernik 2015) and the domain of intelligence (e.g., Ackerman 2023), as well as improved prediction of job performance and other organizationally relevant outcomes. There may also be other types of data relevant to selection that could be analyzed by ML, such as paraverbal and visual information (e.g., Hickman et al. 2022), but societal acceptance of considering such information in hiring may be low because of perceptions of discrimination.

Second, psychometrics theory may evolve due to the unique characteristics of ML algorithms, which may not fit the classic True Score Model underlying traditional testing and assessment methods (e.g., Lord 1965). The True Score Model says the observed score is equal to the true score, plus two kinds of error (random error and systematic or constant error). Random errors can either increase or decrease the observed score, but they are assumed to cancel each other out, so they have no consistent influence on the observed scores. Examples when taking a multiple-choice test might include guessing, luck, swings in alertness and attention, and other "noise" in the measurement. Systematic errors are those that might influence all observed scores in the same direction, such as a high reading level of the exam, confusing questions, instructions, and other factors that would influence all scores either positively or negatively. ML scoring models will certainly have these types of errors, but the examples will be different depending on the data used (e.g., candidate not reporting some information on a resume or luck in having a good experience to describe in response to a past behavior interview question). More importantly, there may be other categories of error with ML scoring models. As illustrated in the discussion of algorithmic bias, which is effectively systematic error, bias might occur in the representativeness of the data used to develop the model and in the computations used in the algorithm, which may require expanding the types of error in the True Score Model. Other implications for psychometric theory include that ML algorithms do not test significance, create confidence intervals, or make inferences to future datasets, nor do they require any statistical assumptions, and the most appropriate manner to assess reliability and the minimum levels to expect may be different for ML models that score thousands of variables, many of which will have missing values, and scores are created by complex weighting of variables rather than simple summing like most traditional tests.

Third, validation theory may also need to expand to consider the differences between validating ML models compared to traditional procedures. Data scientists rely almost entirely on criterion validity with cross validation to ensure that the models will hold up in future samples. That requires large samples given the large number of variables and associated weights in ML models, which are often not available in selection. Thus, how should ML models be validated in such contexts? Part of the solution may lie in determining the sample sizes required compared to the number of variables (e.g., Landers et al. 2023) or the use of algorithms that have fewer variables, less complex weighting, or that are simplified so as to require smaller samples.

Although not yet addressed in the literature, is content validation appropriate for ML models? If so, how should it be performed? The traditional HRM approaches, such as using job analysis, linkages to assessment items, and judgments by subject matter experts may not be appropriate for ML models because they are developed based on a dataset and not a job analysis, have thousands or more variables in many cases, and subject matter experts on the job do not understand the ML models well enough to make judgments about them.

Another possible difference with ML models is whether the process of developing the model requires validation. Data scientists will usually explore a range of models and use the best, which likely causes some capitalization on chance that requires validation. This step often occurs before the step of cross-validating the model on subsamples of the dataset, thus is not explicitly validated. As a final comment, how should models that learn/evolve continuously be validated? Their parameters are adjusted on an on-going basis as data are accumulated by the use of the model. Thus, the models and the underlying dataset are not static enough to allow normal validation procedures.

Fourth, our theoretical understanding of fairness and bias of ML models will evolve as research accumulates. One change is likely to be in the dimensions of fairness and bias of importance when it comes to selection procedures involving ML. The past focus has been on uncertainty (black box) and distrust as to what it measures, reflecting primarily the procedural justice dimensions of job relatedness, information, communication, and propriety of questions (Bauer et al. 2001; Gilliland 1993). Concern over those justice dimensions should subside as understanding increases and ML becomes more commonplace and accepted, along with greater communication from organizations as to what is measured. Candidates and the public will likely then begin to appreciate the other procedural dimensions that are actually improved by ML, such as consistency of administration, having the exact same chances as other candidates, transparency as to what is measured, and the total lack of subjectivity or intentional bias on the part of decision makers.

11 | Conclusion

The primary question addressed in this article is: Can the legal and professional personnel selection principles be met with machine learning? The answer appears to be "yes." Enough high-quality research has accumulated to support its use. ML models developed for personnel selection purposes are able to meet the basic principles of personnel selection, such as job relatedness and reliability, and perhaps better meet many other principles than existing procedures. ML models can be understood and explained, and they should not increase the legal risks from hiring. There are some uses of ML in selection that may apply to most organizations and are worth considering, but there are contingencies to consider. Some additional expertise may be required of HRM professionals depending on their role, but that is not an inhibiting factor. The use of large language models (LLMs) by candidates is certain to have a major impact on the information they provide, but too little is known to make definitive statements at this time other than unproctored remote assessments should probably not be used in the future.

It is certain that ML will play a greater role in selection in the future. It has too many known and emerging advantages to ignore in terms of improving hiring procedures through measuring additional information and better prediction, improving fairness through consistency and objectivity, and improving efficiency through saving costs and time. Acceptance by management, candidates, and society also appears to be increasing as people begin to trust ML and recognize its benefits. The use of ML is probably inevitable, especially in large organizations, so the question is a matter of timing rather than potential unknown risks or waiting for additional research evidence. Many important scientific questions remain, defining a research agenda for the future. We are still at the beginning of what is likely to be a tidal wave of research on the topic in the near future. These are exciting times in the evolution of personnel selection.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

Ackerman, P. L. 2023. "Intelligence… Moving Beyond the Lowest Common Denominator." *American Psychologist* 78, no. 3: 283–297. https://doi.org/10.1037/amp0001057.

Ajunwa, I. 2020. "The 'Black Box' at Work." *Big Data & Society* 7, no. 2: 2053951720966181. https://doi.org/10.1177/2053951720938093.

Albaroudi, E., T. Mansouri, and A. Alameer. 2024. "A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring." *Ai* 5, no. 1: 383–404. https://doi.org/10.3390/ai5010019.

Alexander, L., III, Q. C. Song, L. Hickman, and H. J. Shin. 2025. "Sourcing Algorithms: Rethinking Fairness in Hiring in the Era of Algorithmic Recruitment." *International Journal of Selection and Assessment* 33, no. 1: e12499. https://doi.org/10.1111/ijsa.12499.

Allen, K. S., M. Affourtit, and C. M. Reddock. 2020. "The Machines Aren't Taking Over (Yet): An Empirical Comparison of Traditional, Profiling, and Machine Learning Approaches to Criterion-Related Validation." *Personnel Assessment and Decisions* 6, no. 3: 4–12. https://doi.org/10.25035/pad.2020.03.002.

Artificial Intelligence Video Interview Act. 2020. "Illinois General Assembly." https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68.

Bauer, T. N., D. M. Truxillo, R. J. Sanchez, J. M. Craig, P. Ferrara, and M. A. Campion. 2001. "Applicant Reactions to Selection: Development of the Selection Procedural Justice Scale (SPJS)." *Personnel Psychology* 54: 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x.

Bligh, M. C., J. C. Kohles, and J. R. Meindl. 2004. "Charisma Under Crisis: Presidential Leadership, Rhetoric, and Media Responses Before and After the September 11th Terrorist Attacks." *Leadership Quarterly* 15, no. 2: 211–239. https://doi.org/10.1016/j.leaqua.2004.02.005.

Bordt, S., and U. von Luxburg. 2023. "International Conference on Artificial Intelligence and Statistics." In *From Shapley Values to Generalized Additive Models and Back*, vol. 206, 709–745. PMLR.

Campbell, J. P., and B. M. Wiernik. 2015. "The Modeling and Assessment of Work Performance." *Annual Review of Organizational Psychology and Organizational Behavior* 2, no. 1: 47–74. https://doi.org/10.1146/annurev-orgpsych-032414-111427.

Campion, E. D., and M. A. Campion. 2020. "Using Computer-Assisted Text Analysis (CATA) to Inform Employment Decisions: Approaches, Software, and Findings." *Research in Personnel and Human Resources Management* 38: 287–327. https://doi.org/10.1108/S0742-730120200000038010.

Campion, E. D., and M. A. Campion. 2025. "A Review of Text Analysis in Human Resource Management Research: Methodological Diversity, Constructs Identified, and Validation Best Practices." *Human Resource Management Review* 35: 101078. https://doi.org/10.1016/j.hrmr.2025.101078.

Campion, E. D., M. A. Campion, J. Johnson, et al. 2024. "Using Natural Language Processing to Increase Prediction and Reduce Subgroup Differences in Personnel Selection Decisions." *Journal of Applied Psychology* 109, no. 3: 307–338. https://doi.org/10.1037/apl0001144.

Campion, M. A., and E. D. Campion. 2023. "Machine Learning Applications to Personnel Selection: Current Illustrations, Lessons Learned, and Future Research." *Personnel Psychology* 76: 993–1009. https://doi.org/10.1111/peps.12621.

Campion, M. A., E. D. Campion, and M. H. Fogel. 2024. "Natural Language Processing Dictionaries of Competencies and Personality Traits for Personnel Selection." Society for Industrial and Organizational Psychology Annual Conference, Chicago, IL, USA.

Campion, M. A., D. K. Palmer, and J. E. Campion. 1997. "A Review of Structure in the Selection Interview." *Personnel Psychology* 50: 655–702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x.

Campion, M. C., E. D. Campion, and M. A. Campion. 2018. "Big Data Techniques and Talent Management: Recommendations for Organizations and a Research Agenda for I-O Psychologists." *Industrial and Organizational Psychology: Perspectives on Science and Practice* 11, no. 2: 250–257. https://doi.org/10.1017/iop.2018.14.

Campion, M. C., M. A. Campion, E. D. Campion, and M. H. Reider. 2016. "Initial Investigation Into Computer Scoring of Candidate Essays for Personnel Selection." *Journal of Applied Psychology* 101: 958–975. https://doi.org/10.1037/apl0000108.

Chatzimparmpas, A., R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. 2020. "The State of the Art in Enhancing Trust in Machine Learning Models With the Use of Visualizations." *Computer Graphics Forum* 39, no. 3: 713–756. https://doi.org/10.1111/cgf.14034.

Chowdhury, S., S. Joel-Edgar, P. K. Dey, S. Bhattacharya, and A. Kharlamov. 2023. "Embedding Transparency in Artificial Intelligence Machine Learning Models: Managerial Implications on Predicting and Explaining Employee Turnover." *International Journal of Human Resource Management* 34, no. 14: 2732–2764. https://doi.org/10.1080/09585192.2022.2066981.

Dastin, J. 2022. "Amazon Scraps Secret ML Recruiting Tool That Showed Bias Against Women." In *Ethics of Data and Analytics*, 296–299. Auerbach Publications.

De Corte, W., F. Lievens, and P. R. Sackett. 2007. "Combining Predictors to Achieve Optimal Trade-Offs Between Selection Quality and Adverse Impact." *Journal of Applied Psychology* 92, no. 5: 1380–1393. https://doi.org/10.1037/0021-9010.92.5.1380.

Equal Employment Opportunity Commission, Civil Service Commission, Department of labor, and Department of Justice. 1978. "Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. Part 1607 et Seq." 43 Federal Register (August 25, 1978).

Fan, J., T. Sun, J. Liu, et al. 2023. "How Well Can an ML Chatbot Infer Personality? Examining Psychometric Properties of Machine-Inferred Personality Scores." *Journal of Applied Psychology* 108, no. 8: 1277–1299. https://doi.org/10.1037/apl0001082.

Garg, S., S. Sinha, A. K. Kar, and M. Mani. 2022. "A Review of Machine Learning Applications in Human Resource Management." *International Journal of Productivity and Performance Management* 71, no. 5: 1590–1610. https://doi.org/10.1108/IJPPM-08-2020-0427.

Ghiselli, E. E., J. P. Campbell, and S. Zedeck. 1981. *Measurement Theory for the Behavioral Sciences*. WH Freeman.

Gilliland, S. W. 1993. "The Perceived Fairness of Selection Systems: An Organizational Justice Perspective." *Academy of Management Review* 18, no. 4: 694–734. https://doi.org/10.5465/amr.1993.9402210155.

Goretzko, D., and L. S. F. Israel. 2022. "Pitfalls of Machine Learning-Based Personnel Selection." *Journal of Personnel Psychology* 21, no. 1: 37. https://doi.org/10.1027/1866-5888/a000287.

Heidemann, A., S. M. Hülter, and M. Tekieli. 2024. "Machine Learning With Real-World HR Data: Mitigating the Trade-Off Between Predictive Performance and Transparency." *International Journal of Human Resource Management* 35, no. 14: 2343–2366. https://doi.org/10.1080/09585192.2024.2335515.

Hernandez, I., and W. Nie. 2023. "The AI-IP: Minimizing the Guesswork of Personality Scale Item Development Through Artificial Intelligence." *Personnel Psychology* 76: 1011–1035. https://doi.org/10.1111/peps.12543.

Hickman, L., N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo. 2022. "Automated Video Interview Personality Assessments: Reliability, Validity, and Generalizability Investigations." *Journal of Applied Psychology* 107: 1323–1351. https://doi.org/10.1037/apl0000695.

Hickman, L., C. N. Herde, F. Lievens, and L. Tay. 2023. "Automatic Scoring of Speeded Interpersonal Assessment Center Exercises via Machine Learning: Initial Psychometric Evidence and Practical Guidelines." *International Journal of Selection and Assessment* 31, no. 2: 225–239. https://doi.org/10.1111/ijsa.12418.

Hickman, L., R. Saef, V. Ng, S. E. Woo, L. Tay, and N. Bosch. 2024. "Developing and Evaluating Language-Based Machine Learning Algorithms for Inferring Applicant Personality in Video Interviews." *Human Resource Management Journal* 34, no. 2: 255–274. https://doi.org/10.1111/1748-8583.12356.

Hickman, L., L. Tay, and S. E. Woo. 2025. "Are Automated Video Interviews Smart Enough? Behavioral Modes, Reliability, Validity, and Bias of Machine Learning Cognitive Ability Assessments." *Journal of Applied Psychology* 110, no. 3: 314–335. https://doi.org/10.1037/apl0001236.

Huffcutt, A. I., and W. Arthur. 1994. "Hunter and Hunter (1984) Revisited: Interview Validity for Entry-Level Jobs." *Journal of Applied Psychology* 79, no. 2: 184–190. https://doi.org/10.1037/0021-9010.79.2.184.

Hunkenschroer, A. L., and C. Luetge. 2022. "Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda." *Journal of Business Ethics* 178, no. 4: 977–1007. https://doi.org/10.1007/s10551-022-05049-6.

Jirjees, A. K., A. M. Ahmed, A. A. Abdulla, et al. 2025. "Machine Learning for Recruitment: Analyzing Job-Matching Algorithms." *Machine Learning* 27: 1.

Kazim, E., A. S. Koshiyama, A. Hilliard, and R. Polle. 2021. "Systematizing Audit in Algorithmic Recruitment." *Journal of Intelligence* 9, no. 3: 46. https://doi.org/10.3390/jintelligence9030046.

Kelan, E. K. 2024. "Algorithmic Inclusion: Shaping the Predictive Algorithms of Artificial Intelligence in Hiring." *Human Resource Management Journal* 34, no. 3: 694–707. https://doi.org/10.1111/1748-8583.12511.

Köchling, A., and M. C. Wehner. 2020. "Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development." *Business Research* 13, no. 3: 795–848. https://doi.org/10.1007/s40685-020-00134-w.

Kodiyan, A. A. 2019. "An Overview of Ethical Issues in Using ML Systems in Hiring With a Case Study of Amazon's ML Based Hiring Tool." Researchgate Preprint, 12, 1–9.

Koenig, N., S. Tonidandel, I. Thompson, et al. 2023. "Improving Measurement and Prediction in Personnel Selection Through the Application of Machine Learning." *Personnel Psychology* 76: 1061–1123. https://doi.org/10.1111/peps.12608.

Korzynski, P., G. Mazurek, P. Krzypkowska, and A. Kurasinski. 2023. "Artificial Intelligence Prompt Engineering as a New Digital Competence: Analysis of Generative ML Technologies Such as ChatGPT." *Entrepreneurial Business and Economics Review* 11, no. 3: 25–37. https://doi.org/10.15678/EBER.2023.110302.

Landers, R. N., E. M. Auer, L. Dunk, M. Langer, and K. N. Tran. 2023. "A Simulation of the Impacts of Machine Learning to Combine Psychometric Employee Selection System Predictors on Performance Prediction, Adverse Impact, and Number of Dropped Predictors." *Personnel Psychology* 76: 1037–1060. https://doi.org/10.1111/peps.12587.

Laumer, S., C. Maier, and T. Weitzel. 2022. "HR Machine Learning in Recruiting." In *Handbook of Research on Artificial Intelligence in Human Resource Management*, 105–126. Edward Elgar Publishing. https://doi.org/10.4337/9781839107535.

Levashina, J., F. P. Morgeson, and M. A. Campion. 2009. "They Don't Do It Often, but They Do It Well: Exploring the Relationship Between Applicant Mental Abilities and Faking on Biodata Measures." *International Journal of Selection and Assessment* 17: 271–281. https://doi.org/10.1111/j.1468-2389.2009.00469.x.

Liem, C. C., M. Langer, A. Demetriou, et al. 2018. "Explainable and Interpretable Models in Computer Vision and Machine Learning." In *Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening*, 197–253. https://doi.org/10.1007/978-3-319-98131-4_9.

Liff, J., N. Mondragon, C. Gardner, C. J. Hartwell, and A. Bradshaw. 2024. "Psychometric Properties of Automated Video Interview Competency Assessments." *Journal of Applied Psychology* 109, no. 6: 921–948. https://doi.org/10.1037/apl0001173.

Lord, F. M. 1965. "A Strong True-Score Theory, With Applications." *Psychometrika* 30, no. 3: 239–270. https://doi.org/10.1007/BF02289490.

Marvin, G., N. Hellen, D. Jjingo, and J. Nakatumba-Nabende. 2023. "Prompt Engineering in Large Language Models." In *International Conference on Data Intelligence and Cognitive Informatics*, 387–402. Springer Nature. https://doi.org/10.1007/978-981-99-7962-2_30.

Miller, J. D., D. R. Lyman, T. A. Widiger, and C. Leukefeld. 2001. "Personality Disorders as Extreme Variants of Common Personality Dimensions: Can the Five Factor Model Adequately Represent Psychopathy?" *Journal of Personality* 69, no. 2: 253–276. https://doi.org/10.1111/1467-6494.00144.

Morgeson, F. P., M. A. Campion, R. L. Dipboye, J. R. Hollenbeck, K. Murphy, and N. Schmitt. 2007. "Reconsidering the Use of Personality Tests in Personnel Selection Contexts." *Personnel Psychology* 60: 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x.

Muchinsky, P. M. 1979. "The Use of Reference Reports in Personnel Selection: A Review and Evaluation." *Journal of Occupational Psychology* 52, no. 4: 287–297. https://doi.org/10.1111/j.2044-8325.1979.tb00463.x.

Nadkarni, S., and V. K. Narayanan. 2005. "Validity of the Structural Properties of Text-Based Causal Maps: An Empirical Assessment." *Organizational Research Methods* 8, no. 1: 9–40. https://doi.org/10.1177/1094428104271999.

Nigam, A., R. Pasricha, T. Singh, and P. Churi. 2021. "A Systematic Review on AI-Based Proctoring Systems: Past, Present and Future." *Education and Information Technologies* 26, no. 5: 6421–6445. https://doi.org/10.1007/s10639-021-10597-x.

Pessach, D., G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal. 2020. "Employees Recruitment: A Prescriptive Analytics Approach via Machine Learning and Mathematical Programming." *Decision Support Systems* 134: 113290. https://doi.org/10.1016/j.dss.2020.113290.

Ployhart, R. E., and B. C. Holtz. 2008. "The Diversity–Validity Dilemma: Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact in Selection." *Personnel Psychology* 61: 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x.

Putka, D. J., F. L. Oswald, R. N. Landers, A. S. Beatty, R. A. McCloy, and M. C. Yu. 2023. "Evaluating a Natural Language Processing Approach to Estimating KSA and Interest Job Analysis Ratings." *Journal of Business and Psychology* 38, no. 2: 385–410. https://doi.org/10.1007/s10869-022-09824-0.

Ramineni, C., and D. M. Williamson. 2013. "Automated Essay Scoring: Psychometric Guidelines and Practices." *Assessing Writing* 18: 25–39. https://doi.org/10.1016/j.asw.2012.10.004.

Reddy, D. J. M., S. Regella, and S. R. Seelam. 2020. "Recruitment Prediction Using Machine Learning." In *2020 5th International Conference on Computing, Communication and Security* (*ICCCS*) (1–4). IEEE. https://doi.org/10.1109/ICCCS49678.2020.9276955.

Rottman, C., C. Gardner, J. Liff, N. Mondragon, and L. Zuloaga. 2023. "New Strategies for Addressing the Diversity–Validity Dilemma With Big Data." *Journal of Applied Psychology* 108: 1425–1444. https://doi.org/10.1037/apl0001084.

Sajjadiani, S., A. J. Sojourner, J. D. Kammeyer-Mueller, and E. Mykerezi. 2019. "Using Machine Learning to Translate Applicant Work History Into Predictors of Performance and Turnover." *Journal of Applied Psychology* 104: 1207–1225. https://doi.org/10.1037/apl0000405.

Society for Industrial and Organizational Psychology. 2019. *Principles for the Validation and Use of Personnel Selection Procedures*. 5th ed. Author.

Song, Q. C., S. Wee, and D. A. Newman. 2017. "Diversity Shrinkage: Cross-Validating Pareto-Optimal Weights to Enhance Diversity via Hiring Practices." *Journal of Applied Psychology* 102: 1636–1657. https://doi.org/10.1037/apl0000240.

Surya, C. S., G. Kiriti, K. S. Saketh, P. S. Charan, and A. Anjali. 2024. "Optimizing Job Matching Through Automated Resume Screening With NLP and Machine Learning." In *2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies* (*INSPECT*) (1–6). IEEE. https://doi.org/10.1109/INSPECT63485.2024.10896014.

Tausczik, Y. R., and J. W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29, no. 1: 24–54. https://doi.org/10.1177/0261927X09351676.

Waung, M., P. McAuslan, J. M. DiMambro, and N. Mięgoć. 2017. "Impression Management Use in Resumes and Cover Letters." *Journal of Business and Psychology* 32: 727–746. https://doi.org/10.1007/s10869-016-9470-9.

Yanamala, K. K. R. 2022. "Dynamic Bias Mitigation for Multimodal AI in Recruitment Ensuring Fairness and Equity in Hiring Practices." *Journal of Artificial Intelligence and Machine Learning in Management* 6, no. 2: 51–61. https://www.researchgate.net/publication/387937583.

Zhang, N., M. Wang, H. Xu, et al. 2023. "Reducing Subgroup Differences in Personnel Selection Through the Application of Machine Learning." *Personnel Psychology* 76: 1125–1159. https://doi.org/10.1111/peps.12593.