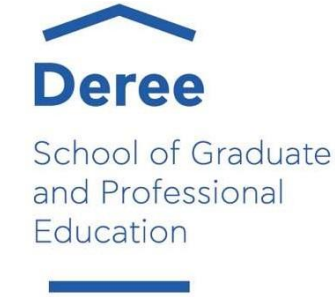

HUMAN VS. MACHINE: AN EMPIRICAL STUDY OF HRM PROFESSIONALS' PERCEPTIONS OF BIAS AND FAIRNESS ISSUES IN AI-DRIVEN EVALUATIONS

by

IOANNIS BARRETT

A thesis submitted in partial fulfillment of the

requirements for the degree of MASTER OF SCIENCE

in

ORGANIZATIONAL PSYCHOLOGY

The American College of Greece

2025

An Abstract of the Thesis of

Ioannis Barrett for the Thesis of Science

In Organizational Psychology to be awarded in

June 2025

Title: HUMAN VS. MACHINE: AN EMPIRICAL STUDY OF HRM PROFESSIONALS' PERCEPTIONS OF BIAS AND FAIRNESS ISSUES IN AI-DRIVEN EVALUATIONS

Abstract

Recent studies suggest that while Artificial Intelligence (AI) can enhance structure and efficiency in performance management, human biases embedded in AI systems can have pernicious effects on the process. However, a perceptible gap remains in understanding the interplay between attitudes toward AI-based decision-making and the effects of algorithmic biases on performance management and other key Human Resource Management (HRM) functions. To this end, the present study aimed to investigate issues of bias and fairness in performance evaluations. Fifty HRM professionals evaluated a biased promotion recommendation made by either an AI or a human agent, rating its rationality, objectivity, and fairness to gauge their implicit attitudes toward AI in promotion decisions. A questionnaire was also employed to examine the relationship between their implicit and explicit attitudes regarding AI's perceived superiority in these qualities. As hypothesized, findings revealed that participants in the AI condition showed a significantly greater implicit endorsement of the biased recommendation, with implicit favoring of AI also predicting its explicit endorsement as superior in promotion contexts. These findings contribute to the existing literature by elucidating potential biases and discrimination arising

from algorithmic decision-making in relation to perceptions and attitudes toward it in HRM. Theoretical and practical implications, along with recommendations for future research, are discussed.

Keywords: human resource management, artificial intelligence in performance management, algorithmic decision-making, perceived fairness, gender bias

Approved: _____

Dr. Olivia Kyriakidou,

Thesis Advisor

Table of Contents

Human vs. Machine: An Empirical Study of HRM Professionals' Perceptions of Bias and Fairness Issues in AI-Driven Evaluations

Artificial Intelligence (AI) is a broad term encompassing various methods, models, and techniques aimed at simulating human intelligence, primarily for collecting, processing, and acting on data. AI includes research areas such as *machine learning* (ML), *natural language processing* (NLP), and *speech and image recognition* (Kaplan & Haenlein, 2019; Paschen et al., 2020). *Generative artificial intelligence* (GenAI) refers to algorithms like ChatGPT—arguably the most widely known chatbot and virtual assistant today—that can create new content, including audio, images, videos, code, and simulations. Since ChatGPT's emergence in November 2022, GenAI has made significant strides, with new tools, regulations, and technological advancements being launched almost every month.

While many have reacted to ChatGPT and AI in general with fear, it is nonetheless evident that ML can be used for good (e.g., societal benefits, economic growth, etc.). Organizations across industries have rushed to integrate GenAI tools into their business models, aiming to capture a substantial market share (McKinsey & Company, 2024). Moreover, research suggests that GenAI applications could contribute up to $4.4 trillion to the global economy annually (Chui et al., 2023). Within the next three years, any technology, telecommunications, or media entity not connected to AI will likely be viewed as ineffective or obsolete (Atluri et al., 2024).

In 2024, organizations have significantly advanced their use of GenAI, transitioning from mere exploration to deriving tangible business value (Singla et al., 2024). According to McKinsey's latest Global Survey on AI, 65% of respondents reported regular use of GenAI in their organizations, nearly doubling the previous year's figures (Singla et al., 2024). Expectations

for GenAI's impact remain high, with 75% of respondents anticipating significant industry changes (Singla et al., 2024). The benefits include cost reductions and revenue increases in business units utilizing GenAI (Singla et al., 2024). Overall, AI adoption has surged to 72%, a substantial rise from the steady 50% seen over the past six years, indicating global expansion (Singla et al., 2024). What is more, professional services show the largest adoption increase, with GenAI most commonly applied in functions like IT, product development, sales, and marketing, where it adds the most value (Singla et al., 2024).

Automation, driven by increasingly sophisticated algorithms, is transforming organizational decision-making processes. These computerized systems, often powered by AI developed to achieve specific goals, are gradually replacing humans (Brynjolfsson & McAfee, 2014; Frey & Osborne, 2017). Additionally, these systems frequently outperform humans while being more cost-effective (Ford, 2015). Algorithms have demonstrated their superiority over human experts across various domains, leading to their rapid implementation in business, legal, and social contexts (Grove et al., 2000; Stone et al., 2013).

In recent years, *algorithmic decision-making* has also become increasingly prevalent in Human Resource Management (HRM), with its significance expected to grow as organizations undergo rapid digital transformation. The process of algorithmic decision-making can be defined as the standardization of routine workplace decisions and the automation and remote control of decision-making processes (Möhlmann & Zalmanson, 2017). Algorithms essentially serve as the foundation for various AI decision-making tools (Möhlmann & Zalmanson, 2017). The gradual replacement of human decision-making by algorithms carries substantial individual and societal implications for organizational optimization (Lee, 2018; Lindebaum et al., 2020). This shift to algorithmic decision-making simplifies the identification of hidden talent within organizations

and enables the automatic review of large volumes of applications (Carey & Smith, 2016; Savage & Bales, 2017).

Several commercial providers, including Google, Microsoft, IBM, and SAP, offer algorithmic platforms and systems that enhance human resource (HR) practices (Walker, 2012). AI is transforming the HR field, particularly in recruitment and management (Brown, 2024). Large companies such as Vodafone, Intel, Unilever, and IKEA have integrated AI technologies into their HR functions (Daugherty & Wilson, 2018). AI offers various tools to streamline HR processes, from automated résumé screening to predictive analytics for employee performance (Brown, 2024). The key drivers of algorithmic decision-making include increasing decision-making certainty, enhancing productivity, minimizing risks, and reducing time and costs (McColl & Michelotti, 2019; Suen et al., 2019; Woods et al., 2020). Beyond these economic motivations, organizations also use algorithmic decision-making to reduce human biases (e.g., personal beliefs, prejudices, etc.) and improve the *fairness*, *objectivity*, and *consistency* of HR processes (Langer et al., 2019; Raghavan et al., 2020). However, *bias*, *unfairness*, and *discrimination* remain potential risks when relying on algorithmic decision-making (Lindebaum et al., 2020; Simbeck, 2019). While AI can improve HR efficiency, it is crucial for HR professionals to address AI-related biases to foster a fair and inclusive workplace (Brown, 2024).

Generally speaking, discrimination can be defined as the unequal treatment of different groups based on arbitrary attributes such as ethnicity, age, or gender rather than on qualitative differences like individual performance (Arrow, 1973). Algorithms can produce biased outcomes and perpetuate discrimination if they are trained on biased (Barocas & Selbst, 2016), inaccurate (Kim, 2017), or unrepresentative data (Suresh & Guttag, 2019). As a result, algorithms are prone to generating or reinforcing biased decisions when their input data—also known as “training

data"—are flawed (Chander, 2017). Thus, challenges related to algorithmic decision-making include a lack of *transparency* and *accountability* concerning the training data, the algorithm itself, and other factors that might influence algorithmic outcomes as well (Pasquale, 2015).

Human raters often process information inconsistently, leading to decisions that may be insufficient, contradictory, or not evidence-based (Woods et al., 2020). This is why many argue that algorithmic decision-making can, by and large, enhance the standardization of procedures, making decisions more objective and reducing the likelihood of errors (Kaibel et al., 2019). The argument is that such systems can improve two key dimensions of fairness: *distributive fairness*, which concerns the perceived fairness of outcomes, and *procedural fairness*, which focuses on the perceived fairness of the processes used to make decisions—both of which benefit from standardization. However, *interactional fairness*—the quality of interpersonal treatment people receive during decision-making processes—remains difficult to achieve with algorithms due to the absence of human interaction. Especially in employee evaluations, fairness is not only about the procedure or its outcomes but also about how those involved perceive the fairness of the entire process (Köchling & Wehner, 2020). Since algorithms lack personal interaction, fulfilling all three dimensions of fairness, particularly interactional fairness, remains challenging (Köchling & Wehner, 2020).

Literature Review

Bias in AI Systems

Over the past few years, society has contended with the profound ramifications of human biases permeating AI systems (Manyika et al., 2019). AI-based technologies are increasingly responsible for decisions traditionally made by humans, such as determining who is hired or

fired, who is granted a loan, or the sentence someone will serve in prison (Citron & Pasquale, 2014; O'Neil, 2016). It is widely accepted that AI algorithms are rapidly infiltrating all areas of life. Today, businesses, governments, and other organizations extensively deploy these algorithms to make decisions that significantly affect individuals and society at large. While these decisions can provide solutions to everyday problems across various fields, they also carry risks, such as being denied employment or medical treatment. Even AI technologies not specifically designed for high-stakes tasks can still be integrated into pipelines that perform such tasks (Buolamwini & Gebru, 2018). It is estimated that nearly 100% of organizations will be using AI by 2025, with the AI software market projected to reach $37 billion by the same year (Gualtieri, 2021). Therefore, in the absence of thorough testing and diverse teams, unconscious biases can easily seep into ML models, causing AI systems to automate and perpetuate these biases (Marr, 2022).

Moreover, AI has demonstrated surprising effectiveness across a broad range of tasks traditionally associated with human intelligence. Nevertheless, it can still perpetuate inherent human biases. Ample evidence indicates that AI models can incorporate and scale human and societal biases (Silberg & Manyika, 2019). What is more, extensive research highlights the significant impact of these biases infiltrating AI systems and their detrimental consequences for society (Manyika et al., 2019). In many cases, however, AI can reduce human subjectivity in data interpretation, as ML algorithms are trained to focus solely on variables that enhance predictive accuracy based on the provided training data. It is therefore essential to recognize that as more enterprises seek to integrate AI into various aspects of their operations, acute awareness and mitigation of these risks emerge as urgent priorities (Manyika et al., 2019).

Biases in human decision-making processes are well documented, ranging from empirical evidence obtained in field experiments (e.g., Bertrand & Mullainathan, 2004) to unconscious biases uncovered in implicit association tests (e.g., Greenwald et al., 1998). Human bias can manifest in various ways, such as outright discrimination (e.g., Dovidio & Gaertner, 2000) or by considering demographic characteristics instead of merit to guide rewards like promotions, leading to unfair decisions (e.g., Goldman et al., 2006). Given the substantial evidence of both conscious and unconscious discrimination (e.g., Kite & Whitley, 2016), the increasing prevalence of algorithms initially inspired a great deal of hope. People were inclined to believe that algorithmic agents could perhaps generate fairer evaluations by avoiding human agents' reliance on biased *heuristic assumptions*, ultimately minimizing—or even eliminating—the discrimination perpetuated by humans (Jago & Laurin, 2022). Although we are well aware of the flaws in human decision-making across various domains—shaped by individual and societal biases often operating at an unconscious level—a critical question arises: will decisions made by AI exhibit less bias than those made by humans, or will AI exacerbate these issues?

Discrimination, which is closely related to unfairness, can be defined as "the unequal treatment of different groups" (Arrow, 1973). Discriminatory categories can be strongly correlated with non-discriminatory ones; for example, age (a discriminatory category) might correlate with years of work experience (a non-discriminatory category) (Persson, 2016). Additionally, there is a distinction to be made between *implicit discrimination*, which stems from unconscious attitudes or stereotypes, and *explicit discrimination*, which involves a conscious aversion to certain groups (Bertrand et al., 2005). In the context of fairness and justice within organizations, discrimination can erode fairness and justice, which are typically understood through three core dimensions: *distributive*, *procedural*, and *interactional* (Gilliland, 1993).

Distributive justice concerns the fairness of individuals' outcomes, addressing who receives what (Cropanzano et al., 2007). Factors that promote distributive justice include *need* (allocating resources based on urgency), *equality* (providing equal amounts to everyone), and *equity* (distributing based on individual contributions) (Cropanzano et al., 2007).

Humans tend to simplify their decisions using *heuristic thinking*, whereas algorithms can analyze vast amounts of data—derived from numerous variables—to generate more accurate, data-driven predictions (James et al., 2013). However, if an algorithm is trained on a biased dataset, such as one where women received lower performance evaluations than men, it may use gender as a factor in predicting future performance, leading to significant problems. Furthermore, if these evaluations were biased (e.g., due to managers' biases) and there is no actual link between gender and performance in objective reality, the algorithm would likely fail to detect this and could perpetuate the unfair bias, thereby contributing to gender discrimination in the workplace. Additionally, this bias would be replicated invisibly, as modern algorithms are often opaque about the weights they assign to different predictor variables (Kleinberg et al., 2019).

Sources and Examples of Bias in AI

Bias can creep into algorithms in various ways. AI systems develop their decision-making capabilities from the training data used to “feed” ML algorithms (Manyika et al., 2019). Often, this data includes prejudiced human judgments or reflects societal and historical inequities (Manyika et al., 2019). Research has uncovered pernicious biases not only in the underlying training data but also in organizational structures, such as the predominance of male developers (Demetis & Lee, 2018; Leicht-Deobald et al., 2019). For example, research has revealed the

adverse effects of algorithms on women's career opportunities, such as in the delivery of STEM (science, technology, engineering, and math) advertisements (Lambrecht & Tucker, 2019).

A notable example is Amazon.com, Inc.'s attempt to automate its hiring process with ML, which faced significant challenges due to gender bias (Dastin, 2018). Amazon's ML experts discovered a significant issue of gender discrimination in their automated hiring process, leading the company to cease using this hiring algorithm (Dastin, 2018). The system favored male applicants due to historical résumé data from the predominantly male-dominated tech industry (Dastin, 2018). Moreover, this hiring algorithm was designed to evaluate job candidates in a similar manner to how products are rated on Amazon (Dastin, 2018). Interestingly, it was revealed that the system favored applicants with terms more prevalent in men's résumés, such as "executed" or "captured" (Dastin, 2018). Despite efforts to neutralize biased terms, this HR algorithm continued to produce discriminatory results, prompting Amazon to disband the project (Dastin, 2018). This case underscores the limitations of ML-enabled systems in recruitment and the ongoing struggle to ensure fairness and transparency in AI-driven processes. It serves as a cautionary tale about the dangers of overreliance on AI technology in hiring and other HRM functions, fueling growing public concern regarding AI's impact on our lives.

In the following days, several articles on Amazon's "sexist AI" appeared in major news outlets, including the BBC, The Guardian, and The Wall Street Journal. In addition to that, just a few days later, the Public Employment Service Austria published the specifications for an algorithm used to classify unemployed citizens based on their chances of success in the labor market. What stood out was that the algorithm predicted lower employment chances for unemployed women compared to men with the exact same characteristics. This sparked public

outrage, with Austrian media publishing caustic headlines such as “Computer says no: Algorithm gives women fewer chances” (Wimmer, 2018), which persisted for several weeks (Reiter, 2019).

This problem, however, goes back a long way. Back in 1988, a British medical school was found culpable of discrimination by the UK Commission for Racial Equality (Lowry & Macpherson, 1988). Their computerized selection system—intended to emulate human admissions decisions—was biased against women and applicants with non-European names (Lowry & Macpherson, 1988). Despite achieving an accuracy rate of 90-95%, the program perpetuated biases present among human decision-makers (Lowry & Macpherson, 1988). Thus, reverting to the previous method would not resolve the issue, as biased decision-making would continue to persist (Lowry & Macpherson, 1988). Interestingly enough, the school had enrolled a higher proportion of non-European students at the time than other medical schools in London (Lowry & Macpherson, 1988). Over three decades later, we still find ourselves grappling with the same challenges, despite the increased complexity of today’s algorithms (Manyika et al., 2019). The key issue here is that while AI can aid in identifying and mitigating human biases, it can also exacerbate the problem by embedding and deploying biases extensively in sensitive application areas (Manyika et al., 2019).

Training Data as a Source of Bias in AI

Much of the current discussion surrounding bias in AI systems is often oversimplified to terms like "racist algorithms." However, it is important to bear in mind that the problem is not the algorithms per se, but rather the data that research teams feed them. In most cases, the primary source of bias stems from the underlying data rather than the algorithm itself. For example, the collection of *historical data* (i.e., data from the past) is a common starting point for

data science projects. The problem, though, is that historical data is more often than not biased in ways we do not wish to perpetuate into the future (Civin, 2018).

Consider, for instance, a company building a model to decide which job applicants to invite for interviews. If the model is trained on résumés of applicants previously invited for similar positions, and if the company's HR staff have historically rejected applications from former stay-at-home parents attempting to return to the workforce—which, regrettably, is a common practice (Weisshaar, 2018)—the training algorithm could produce a model that excludes applicants with long employment gaps. This would then also disproportionately reject women, who still make up the majority of stay-at-home parents (Varathan, 2017), even if gender is not explicitly included in the training dataset. Consequently, the model would give rise to gender discrimination by amplifying existing human biases.

Research by Sweeney (2013) on racial differences in online ad targeting found that searches for African-American-identifying names yielded more ads featuring the word "arrest," compared to searches for White-identifying names. Even if ads with and without the term "arrest" were initially shown equally, user interactions might cause the algorithm to display the "arrest" ads more frequently based on the search patterns (Sweeney, 2013). Similarly, Datta et al. (2015) found that Google's Ads tool for targeted advertising exhibited gender bias by serving significantly fewer ads for high-paid jobs to women than to men.

Models are often trained with data that reflects human decisions or the indirect consequences of societal and historical inequalities. For example, word embeddings—a set of NLP techniques—trained on news articles can exhibit societal *gender stereotypes* (Packer et al., 2018). Bias can also creep into data based on how they are collected or selected for use. In criminal justice models, for instance, oversampling certain neighborhoods due to over-policing

can result in recording more crime in these areas, which then leads to even more policing (Lum & Isaac, 2016). Similarly, user-generated data can create feedback loops that perpetuate bias. Due to the limited transparency of these algorithms, even to their developers (Rai, 2020), researchers in business ethics emphasize the need to hold companies accountable for the algorithms they develop and/ or use, in order to mitigate negative impacts on underrepresented groups, including women and minorities (Buhmann et al., 2020; Martin, 2019).

Furthermore, in the context of HRM, consider a recruitment algorithm trained on historical employment data as an example. If this data contains an *implicit bias* favoring White men over Hispanic men, the algorithm could perpetuate this bias even without having been fed explicit data on ethnicity or gender. It could thus identify patterns that inadvertently reveal an applicant's membership in a protected group, which has historically been less likely to receive job interview invitations (Köchling & Wehner, 2020). As a result, certain groups might face a *systematic disadvantage,* even if the algorithm designer did not intend to marginalize individuals based on these categories or if the algorithm was not directly fed this information (Barocas & Selbst, 2016). For instance, *representation bias* could occur if women are underrepresented in the training data compared to men, leading to a preference for the overrepresented group (i.e., men) and potentially resulting in discriminatory outcomes.

The Opaque Nature of Algorithmic Discrimination

Although many expect that algorithmic evaluation will become ubiquitous and bias-free in the future (Kleinberg et al., 2019), numerous scholars argue that, in their current form, algorithms not only perpetuate but also scale discrimination significantly. More concerning is that due to their complexity, algorithms often discriminate in an opaque manner (e.g., Larson et al., 2016; O'Neil, 2016). This suggests that as algorithms proliferate throughout business and

society, people's assumptions and perceptions may lead them to unwisely seek algorithmic evaluations in contexts where these technologies actually perpetuate—or even amplify—discrimination while simultaneously obscuring it (Jago & Laurin, 2022).

Legal and Ethical Implications of AI Bias

Moreover, many AI systems, such as facial recognition software, rely on ML algorithms trained with labeled data. Recent studies have shown that algorithms trained on biased data can lead to algorithmic discrimination (e.g., Bolukbasi et al., 2016; Caliskan et al., 2017). For example, research by Buolamwini and Gebru (2018) revealed significant biases in commercial facial analysis algorithms. Their study found notably high misclassification rates for darker-skinned females (up to 34.7%) compared to lighter-skinned males (0.8%). These results highlight the importance of inclusive datasets and rigorous performance reporting to ensure fairness, transparency, and accountability in AI systems (Buolamwini & Gebru, 2018). Addressing these issues is crucial to mitigating bias and improving the overall effectiveness and equity of ML technologies.

In 2020, a significant incident involving facial recognition technology occurred when Detroit police arrested Robert Williams—a Black man living in a Detroit suburb—on his front lawn in front of his wife and two young daughters and was detained for nearly 30 hours (Burton-Harris & Mayor, 2020). Williams was misidentified as a suspect in a watch theft case by facial recognition software, despite the fact that the only similarity between him and the actual suspect was that they were both large-framed Black men (Burton-Harris & Mayor, 2020). This case underscores the flaws and biases in facial recognition technology, which has been repeatedly found to be particularly unreliable in identifying Black individuals (Burton-Harris & Mayor,

2020). Williams' ordeal highlights the inherent dangers of using such technology in law enforcement.

Another instance that illustrates this issue is that of a “racist” criminal justice algorithm used to predict recidivism in Broward County, Florida. This algorithm was found to disproportionately categorize Black defendants as "high risk," providing a case in point (Angwin et al., 2016). The investigative news site ProPublica discovered that this criminal justice algorithm erroneously labeled Black defendants as being at a heightened risk for recidivism nearly twice as often as it mislabeled White defendants with similar backgrounds (Angwin et al., 2016). Notably, this was true despite the system not being explicitly fed any data regarding the defendants’ race (Angwin et al., 2016).

On the other hand, there are also data indicating that defendants deemed risky by algorithms do, in fact, often commit many crimes; they miss court appearances at a rate of 56.3%, commit new crimes at a rate of 62.7%, and perpetrate serious offenses such as murder, rape, and robbery at a rate of 4.8% (Kleinberg et al., 2018). There is, in other words, evidence to suggest that algorithms can improve human decision-making and make it fairer in the process (Kleinberg et al., 2018). For example, Kleinberg et al. (2018) demonstrated that algorithms can assist in mitigating racial disparities in the criminal justice system.

Moreover, there may be instances where an ML algorithm detects statistical correlations that are deemed socially unacceptable or even illegal (The Royal Society, 2017). For example, suppose a mortgage lending model finds that older individuals are more likely to default, and consequently reduces lending based on age. In that case, legal institutions and society at large may view this as illegal age discrimination (The Royal Society, 2017). This presents us with the conundrum of how to codify definitions of fairness.

Definitions of Fairness in AI

A significant portion of the discussion surrounding definitions of *fairness* has centered on two primary aspects: *individual fairness*, which emphasizes treating similar individuals similarly, and *group fairness*, which aims to ensure that the model's predictions or outcomes are fair and equitable across various groups, particularly those that may be more vulnerable (Zemel et al., 2013). Attempts to define fairness however have revealed possible trade-offs among different definitions as well as between fairness and other objectives. For example, research has shown that a model cannot simultaneously adhere to more than a few group fairness metrics, except under highly specific circumstances (Chouldechova, 2016; Kleinberg et al., 2017).

First, it is crucial to note that there is a lack of consensus among experts regarding the optimal approach to resolving these trade-offs. Some propose that the best balance might be achieved by implementing different decision thresholds for different groups (e.g., the predicted score required to obtain a loan), especially if there is suspicion of bias in some of the model's underlying variables (Kleinberg et al., 2018). On the other hand, some argue that it would be fairer to maintain a single threshold for all groups (Corbett-Davies et al., 2023). Given these complexities, achieving a singular, universally applicable definition of fairness—and a corresponding metric to measure it—is unlikely (Silberg & Manyika, 2019). Nevertheless, different standards and metrics may be more appropriate depending on the specific circumstances and use case (Silberg & Manyika, 2019).

Approaches to Bias Mitigation

Approaches to *bias mitigation* in AI systems, which aim to reduce discrimination and ensure fairness in ML models, are broadly categorized into three main methods: *preprocessing*,

in-processing, and *post-processing* techniques. Each of these techniques focuses on different stages of the ML pipeline, as outlined in detail below.

Preprocessing Approaches

Preprocessing methods address the primary source of bias at the data level by creating a balanced dataset before it is fed into the learning algorithm (Calmon et al., 2017). The rationale behind this approach is that fairer training data result in less discriminatory models (Calmon et al., 2017). Techniques in this category include "altering class labels" by modifying the labels of instances close to the decision boundary (Luong et al., 2011) and "instance weighting" by assigning different weights to instances based on group membership (Calders et al., 2009).

In-Processing Approaches

In-processing methods involve directly integrating fairness constraints into the ML algorithm. These methods reformulate the classification problem by incorporating discrimination behavior into the objective function through training on latent target labels, constraints, or regularization. Key techniques in this vein include Constraint-based methods: Imposing constraints to minimize disparate mistreatment in models like logistic regression and SVMs (Zafar et al., 2017); Regularization: Reducing indirect prejudice by integrating regularizers that minimize the mutual information between sensitive features and class labels (Kamishima et al., 2012); Modification of decision criteria: Adjusting the splitting criterion in decision trees to consider the impact on protected attributes (Kamiran et al., 2010).

Post-Processing Approaches

Post-processing methods achieve fairness by modifying the model or its predictions after training. These methods include Black-box approaches: Adjusting predictions to maintain

proportionality between protected and unprotected groups, either by altering predictions near the decision boundary (Kamiran et al., 2018) or by wrapping a fair classifier around a base model (Agarwal et al., 2018); and White-box approaches: Altering the internals of the model, such as correcting classification rule confidences or adjusting class labels at decision tree leaves (Calders & Verwer, 2010).

In conclusion, effective bias mitigation in AI requires a comprehensive approach that addresses various stages of the ML process. Preprocessing ensures fair data, in-processing embeds fairness into algorithms, and post-processing adjusts outcomes to meet fairness constraints. The bottom line is that continuous advancements and innovations are crucial for addressing the dynamic and complex nature of bias in AI systems.

Explainable AI (XAI)

The increasing role of ML in decision-making systems, from banking to bail, presents both the opportunity to build better, less biased systems and the risk of reinforcing existing biases (Civin, 2018). To address this concern, the General Data Protection Regulation (GDPR) stipulates that all citizens must be granted a "right to explanation," allowing them to demand an explanation for any "legal or similarly significant" decisions made by machines that could affect them (Kuang, 2017). In other words, the GDPR's right to explanation provides individuals with legal recourse, ensuring that algorithmic decisions affecting their lives are not just made in a *black box* but are open to scrutiny and challenge. Consequently, the effects of such biases could be mitigated by giving victims of "discrimination-by-algorithm" the right to explanation and recourse to human authorities. However, generating these explanations to create *explainable AI* (XAI) is a complex task.

Even when such explanations are available, critics argue that it is unclear whether they effectively counter bias or merely mask it (Civin, 2018). This is where XAI could prove useful. By having human operators review the "reasoning" behind algorithms used in decision-making for high-risk groups, they might be able to address algorithmic bias before it leads to serious consequences (Civin, 2018).

Unlike traditional computer programs where humans explicitly write code, ML systems operate based on the data used to train them. This means that, although humans can measure the accuracy of ML systems, our understanding of how these systems make decisions remains limited. In other words, we are unable to pinpoint the exact decision processes within ML systems (Civin, 2018). To address this, XAI aims to make the decisions of ML algorithms more interpretable (Civin, 2018). For instance, in 2016, researchers from the University of Washington developed an explanation technique called "LIME," which they tested on Google's Inception Network—an image classification neural net (Ribeiro et al., 2016). LIME investigates the image itself instead of analyzing the network's internal operations by modifying parts of the image and observing the changes that most significantly disrupt the algorithm's classification (Ribeiro et al., 2016). In this way, LIME links the algorithm's decisions to specific features of an image (Ribeiro et al., 2016). For example, it was found that obscuring certain parts of a tree frog's face hindered the network's ability to identify the frog, suggesting that the face was crucial to the classification (Ribeiro et al., 2016).

Although methods like LIME cannot fully explain why an algorithm's decisions are not universally effective across all ML models, they remain particularly beneficial for image classification (Civin, 2018). Furthermore, there have been instances of controversy over bias in image classification, such as a racially offensive error made by Google Photos' AI software back

in 2015, which mislabeled two Black individuals as “gorillas” (Civin, 2018). Explanation techniques therefore can help mitigate such biases by enabling human operators to override questionable classification decisions and adjust algorithms accordingly (Civin, 2018).

AI in Human Resource Management

The rapid innovations in computing technology have led to the automation of the vast majority of HR functions and business environments, to varying degrees, intending to enhance both effectiveness and efficiency across numerous activities, such as performance appraisals, 360-degree assessments, and management and organizational development, to name a few (Hendrickson, 2003). Therefore, while simple answers to these emerging challenges may be elusive, it is essential to understand how technology impacts HRM professionals’ attitudes toward its implementation in HR activities and their roles within the profession (Roehling et al., 2005). Recent research by Mantzaris and Myloni (2023) found that HRM professionals believe technology is more effective than humans in addressing HRM challenges. In addition to that, this perception of technology's superiority was found to not differ significantly across cultures (Mantzaris & Myloni, 2023).

Moreover, as ML algorithms are increasingly used in HRM to predict factors like turnover intentions, employee satisfaction, and personality traits, understanding how these algorithms operate becomes essential. This includes insights into how they process data, weigh criteria, and the rationale behind their decisions (Köchling & Wehner, 2020). Just as biased image classification can lead to harmful misclassifications, biased algorithmic decisions in HRM can similarly affect hiring, promotion, and employee treatment in much the same way.

In the context of HRM, ensuring the reliability as well as responsibility of AI and ML applications requires three core elements: transparency, explainability, and interpretability, each

playing a crucial role (Roscher et al., 2020). Transparency refers to the ML approach itself, while interpretability addresses how the ML model interacts with the data so as to clarify its outcomes (Roscher et al., 2020). Meanwhile, explainability involves not only the model and data but also the human involvement in understanding these outcomes (Roscher et al., 2020). While transparency refers to understanding how the ML model functions, explainability helps stakeholders comprehend the rationale behind specific decisions, and interpretability ensures that the results are clearly tied to the input data, providing a holistic view of the algorithm's operation. This distinction is particularly important in HRM, where contextual information and HR expertise are essential for interpreting results and drawing actionable conclusions from algorithm outputs (Roscher et al., 2020).

The rapid innovations in AI applications are expected to transform the structure of HR departments radically and significantly influence the workplaces of tomorrow. This underscores the urgency of mastering human-machine collaboration (Mantzaris & Myloni, 2023). Furthermore, the role of HR managers and their practices are becoming increasingly crucial, as globalization and emerging HR challenges have transformed HR into a key strategic partner within organizations (Nasir, 2017). In view of the fact that intelligent robots and a combination of AI and ML techniques are becoming a priority for top engineers globally, and leading industrial sectors are becoming increasingly automated, HRM must quickly adapt to these challenges and unprecedented transformations.

From an organizational perspective, biases in AI-based decision-making can result in critical misjudgments, erroneous financial assumptions, damage to organizational reputation, and a lack of transparency (Garg et al., 2021). The successful development and implementation of AI-powered solutions in HRM require five core principles, namely, beneficence, non-

maleficence, autonomy, explicability, and justice (Floridi et al., 2018). Of particular interest to the present study is the principle of justice. According to this principle, the adoption of AI in HRM activities and functions should promote fair and just outcomes, such as eliminating bias and fostering diversity (Floridi et al., 2018). Therefore, the integration of AI-driven solutions in HR departments must align with an organization's ability to predict, detect, and mitigate potential biases in these systems to ensure fairness (Akter et al., 2021). This can be achieved by auditing how underlying algorithms behave using empirically sound methodologies and diverse perspectives (Tuffaha, 2023). Additionally, it is essential to build a heterogeneous and well-educated workforce that can collaboratively scrutinize, detect, and address issues of bias and fairness, minimizing the risk of harmful effects (Rozado, 2020).

AI in Performance Management

Employees with high levels of participation, productivity, and on-the-job efficacy add significant value to an organization (Howard, 2019; Mahmoud et al., 2019). That being said, assessing these factors can be challenging for companies that rely on traditional success metrics, as these are often too broad. In this case, AI can improve the precision of *performance evaluations* conducted by HR administrators, allowing them to assess performance over more specific and narrower timeframes (Bhardwaj et al., 2020; Mahmoud et al., 2019).

AI can bring structure to the performance management process, particularly by improving performance feedback (Garr & Jackson, 2019). The use of AI can also automate manual performance evaluations, for example, by clustering employees into distinct groups based on factors such as job satisfaction or performance levels (Aktepe & Ersoz, 2012). Additionally, AI can serve as a powerful tool for gathering and reviewing employee performance data. Some argue that this can actually lead to fairer evaluations (e.g., Budhwar et al., 2022), as it flags

potential biases while also providing managers with insights on how to deliver more meaningful feedback (Garr & Jackson, 2019).

Moreover, AI can optimize training resource allocation (Beane, 2019) and assess training effectiveness (Budhwar et al., 2022). The use of AI can also automate the analysis of organizational skill needs and training, offering relevant training recommendations based on collected data (Garg et al., 2021). Advanced algorithms have been reported to predict employees' performance levels based on their background, performance characteristics, or occupational level at various stages in their careers (Garg et al., 2021). Notably, AI applications aim to detect potential subjectivity in performance appraisals (Garg et al., 2021), evaluate employee expertise, assess the impact of financial incentives on performance (Massrur et al., 2014), and develop personalized incentive programs (Petruzzellis et al., 2006).

AI can also help managers improve the accuracy of the data they use for employee performance assessments (Williams, 2019). Instead of only comparing performance with targets at the beginning and end of specified weekly, monthly, quarterly, or yearly periods, AI enables this process to be ongoing and real-time (Sakka et al., 2022). For instance, an effective way to optimize work performance is to create a work schedule that outlines specific targets and establishes intervals for appraising results. AI can assist in this process by swiftly comparing performance outcomes with initial goals (Radonjic, 2019; Rastgoo, 2016). In addition, enhanced efficiency in performance appraisal, when combined with appropriate rewards, can foster more effective motivational strategies (Anderson et al., 2018).

At present, organizations routinely maintain records of pay and compensation data. AI has the potential to enhance the efficiency of querying this information, particularly in large organizations (Semmler & Rose, 2017). Furthermore, improved processing of remuneration data

facilitated by AI would enhance the perception of fairness within the organization, which can lead to increased organizational performance (Sakka et al., 2020). However, others contend that such approaches could reduce employee performance to mere numbers and threaten *autonomy* and *privacy* through tracking and surveillance (Giermindl et al., 2021), thereby altering the trust relationship between employees and organizations (Appio et al., 2024).

Furthermore, pertinent literature suggests a pressing need to develop a deeper understanding of how biases infiltrating AI systems can affect core HRM functions like performance management, as AI-driven solutions increasingly play a central role in predicting, managing, analyzing, and evaluating employee performance (Ozkazanc-Pan, 2021; Qamar et al., 2021). Previous research has shown that AI-related biases in performance management emphasize the need for data generated by AI-powered solutions to support decision-makers by providing key insights into strategic performance and developing metrics around key success factors (Raffoni et al., 2018). However, these benefits can be compromised by an organization's inability to eliminate biased data generated by these AI solutions (Akter et al., 2021).

The extant literature highlights several detrimental consequences of biases in AI-powered performance management systems, including heightened doubts about AI's ability to accurately evaluate applicants' capabilities (e.g., Akter et al., 2021); reduced trust in feedback quality, and increased concerns about job displacement (e.g., Tong et al., 2021); diminished reliability of the evaluation process (e.g., Minbaeva, 2021); negative effects on fairness, trustworthiness, and organizational effectiveness (e.g., Zhang & Yencha, 2022); and a lack of holistic, heuristic evaluations (e.g., Kim & Heo, 2021). Nonetheless, there is still a perceptible gap in the in-depth understanding of how AI-related biases can impact the performance management process.

Theoretical Background

Algorithmic Discrimination and Public Perception

While the debate over how algorithmic decision-making compares to human decision-making in terms of discrimination is ongoing, it is nonetheless clear that algorithms can, and often do, discriminate. However, the beliefs people hold about algorithmic discrimination, as well as the decisions they make based on those beliefs, may not always align with objective truth and reality (O'Neil, 2016). Psychological factors—such as *stereotypes* and *intuitions* about machines—can shape these beliefs (Jago & Laurin, 2022). As a result, people might understandably conclude that algorithms are less likely to discriminate compared to humans (Jago & Laurin, 2022). Assumptions about AI's *accuracy* also play a role; people often perceive algorithms as exceptionally accurate, believing they can reliably detect patterns and accurately predict "real" values (James et al., 2013). Consequently, machines are mentally represented as relatively agentic entities with a strong capacity for calculation (Gray et al., 2007).

Gender Bias and Algorithmic Evaluation

In a recent series of studies, Jago and Laurin (2022) found that people perceive algorithms as less capable of discrimination than humans and tend to prefer being evaluated by algorithms, especially when they expect discrimination from a human evaluator. The authors argued that groups, organizations, and social systems—such as governments and courtrooms—can appear fair by emphasizing their use of technology, even while potentially operating in discriminatory ways (Jago & Laurin, 2022). It seems that systems can hide behind technology to project an image of *impartiality* (Jago & Laurin, 2022). Their findings suggest that people may feel more comfortable with, and less reactive to, discriminatory behaviors when they believe biased decisions were made—or heavily influenced—by machines (Jago & Laurin, 2022).

Moreover, research has shown that societal stigma influences women's receptivity to algorithmic decision-making, which closely relates to the concept of *mechanical objectivity* (Pethig & Kroenung, 2022). This concept suggests that algorithms are perceived as less biased than human decisions, which are often seen as prone to frailty, irrationality, and discrimination (Martin, 2019). Algorithms are viewed as reducing human involvement (Christin, 2016) and are therefore considered to have greater cognitive abilities while being less influenced by emotion (Castelo et al., 2019). This perception reinforces the belief that algorithms are more competent at performing objective tasks than subjective ones (Pethig & Kroenung, 2022). Consequently, women concerned that their gender identity might affect their evaluations may prefer being assessed by an algorithm rather than a human (Pethig & Kroenung, 2022). From their perspective, evidence of gender bias in algorithms is unlikely to outweigh their everyday experiences of bias and discrimination from colleagues, hiring managers, and others (Bohnet et al., 2016; Moss-Racusin et al., 2012).

In a recent series of studies, Pethig and Kroenung (2022) explored how women perceive algorithmic evaluations, particularly in situations where they feel disadvantaged due to their gender. The research revealed that women in the workplace preferred evaluations conducted by an algorithm over those carried out by a male HR manager. In other words, when given the choice between an algorithmic and a male evaluator, women anticipated less bias from the algorithm. Interestingly, this preference did not extend to evaluations performed by female HR managers (Pethig & Kroenung, 2022). Women viewed algorithms as more objective and less prone to gender bias than male evaluators, which reinforced their preference for algorithmic assessments (Pethig & Kroenung, 2022). In contrast, men did not display a preference for either

type of evaluator, likely due to a lesser concern about gender bias in evaluations (Pethig & Kroenung, 2022).

Emotional Neutrality and Mechanical Objectivity in Algorithmic Perception

People often rightly recognize that algorithms can detect patterns beyond human capability, which is a major driver behind the automation of decision-making processes (Ford, 2015). This perception contrasts with human evaluators, who are more prone to errors, partly because their judgments frequently rely on extraneous factors and mental shortcuts. These mental shortcuts are susceptible to biased stereotypes, which can influence decision-making. As a result, algorithmic decision-making, which appears more consistent and accurate than human decision-making, is likely to be perceived as less discriminatory. Additionally, the perception of algorithmic impartiality is reinforced by the fact that algorithms, unlike humans, cannot experience emotions (Shank & DeSanti, 2018).

While algorithms may seem capable of “thinking,” they are viewed as particularly low in their ability to feel emotions (Shank & DeSanti, 2018). In contrast, human decision-making is often clouded by emotion. Humans may harbor implicit (Dasgupta et al., 2000) or explicit (Jones et al., 2016) negative emotions toward certain demographic groups, such as ethnicities or genders, which can skew their judgments. Therefore, because algorithms are seen as *emotionally neutral,* people assume they are free from the biases and discriminatory attitudes that often drive human decision-making.

To further explore this perception of algorithmic impartiality, it is helpful to consider the concept of objectivity in the context of science and technology. Philosophers of science often define objectivity in terms of "objective knowledge," which is regarded as reliable because it is detached from the subjective perspective of the person(s) producing it (Gunton et al., 2021).

Historically, the belief that objectivity offers an unbiased depiction of reality has granted machines an advantage over humans in producing reliable knowledge. For instance, technological advancements like X-rays and photographs are seen as providing "an unmediated representation of natural phenomena" (Christin, 2016). This belief is captured in the concept of mechanical objectivity, introduced by Daston and Galison (1992). Mechanical objectivity reflects the idea that technology, free from human subjectivity, can transcend personal biases, perspectives, values, and interests (Sprenger & Reiss, 2020). Essentially, it portrays technology as capable of providing an unbiased, faithful representation of reality, as if "nature could speak for itself" (Daston & Galison, 1992).

The appeal of mechanical objectivity is particularly relevant to algorithms, which are often viewed as neutral arbiters in decision-making. The notion that machines, especially algorithms, offer an objective and impartial representation of situations is intuitively appealing and aligns with common perceptions of technology (Castelo et al., 2019). In contexts where "faithfulness to reality" is a central concern, such as employee promotion decisions, the concept of mechanical objectivity provides a useful framework for understanding how algorithms are contrasted with human decision-makers (Gunton et al., 2021).

Automation Bias in AI Perception

AI can be defined as a machine agent capable of thinking and acting intelligently (Russell & Norvig, 2009). Interactions with AI are increasingly designed to mimic human-to-human communication, as AI agents are developed to be more embodied, proximate, intimate, and ultimately human-like (Guzman, 2020; Westerman et al., 2020). However, certain characteristics—such as specific beliefs about AI such as *machine heuristics*—can lead people to perceive it as distinct from human agents, resulting in different evaluations of decisions made by

AI compared to those made by humans (Jones-Jang & Park, 2023). Although AI possesses both human-like and machine-like features, people often focus more on its machine-like characteristics. This tendency can be attributed to a lack of prior direct experiences with AI and the technology's inherent lack of transparency (Jones-Jang & Park, 2023). Consequently, people's beliefs and perceptions of AI are frequently shaped by its representation in the media (Banks, 2020).

Generally speaking, people tend to view AI and machines in general as operating consistently in a pre-programmed and objective manner (Sundar & Kim, 2019). Therefore, it is reasonable to assume that cueing the machine heuristic would lead people to view AI-based decisions as more accurate, consistent, objective, and so forth. Research on machine heuristics indicates that the process by which people judge AI follows an if-then-therefore logic (Bellur & Sundar, 2014). Specifically, if AI is recognized as a source cue, then machine heuristics—entailing perceived characteristics such as consistent performance and being programmed—are triggered, therefore influencing users' judgments of their AI experience (Cloudy et al., 2021). Moreover, research has demonstrated that automation bias arises from the heuristics-based evaluation mechanisms related to AI (Logg et al., 2019).

Automation bias occurs when users overestimate AI's accuracy and performance, driven by overly positive expectations of its programmed capabilities, which creates a perfection schema for AI's performance (Lee, 2018). It refers to people's tendency to favor suggestions made by automated decision-making systems while disregarding contradictory information from human agents (Hoffman, 2024). This bias is particularly common in environments where computer-based systems are used for decision-making, such as financial forecasting and medical

diagnosis. As more decision-making tasks are delegated to AI, automation bias is becoming increasingly prevalent across society (Hoffman, 2024).

Empirical evidence has shown that people are more likely to take advice from AI in areas like forecasting and music recommendations compared to advice from humans (Logg et al., 2019). Additionally, research has found that AI-authored stories and fact-checking messages can reduce partisan biases, leading users to rate these AI-generated messages as more credible than those written by humans (Cloudy et al., 2021; Moon et al., 2022; Wojcieszak et al., 2021). However, as previously discussed, AI-driven content is not immune to bias due to the inherent limitations and biases in the data that inform AI decisions (Hsu, 2020; Noble, 2018)—a fact that users often overlook or underestimate.

Furthermore, it is reasonable to assume that most AI users are unfamiliar with well-documented implicit biases in AI, leading them to perceive AI as a *neutral agent*. A key difference in how people generally perceive AI agents compared to human agents is that they often harbor overly high expectations of AI's pre-programmed and consistent performance, which can then result in greater disappointment when AI produces unsatisfactory outcomes (Alvarado-Valencia & Barrero, 2014).

Purpose of the Present Study

A thorough review of the relevant literature revealed a notable gap in empirical research regarding HRM professionals' perceptions and attitudes toward AI-based decision-making systems, particularly in their application to performance evaluations and the identification of leadership potential. This indicates a significant lack of comprehensive exploration in this area. Furthermore, research in this domain in general is still in its early stages, with the preponderance of existing studies focusing on fairness issues in algorithmic decision-making during recruitment

processes (Appio et al., 2024; Kaushal et al., 2021). This is therefore a crucial under-researched area of ethical decision-making that warrants greater attention due to its ethical implications, as well as the legal and reputational risks it poses for organizations.

The issues of bias and fairness in AI-based decision-making systems within HR functions are a timely topic that is gaining increasing importance. Companies may face reputational and legal risks if their HR methods are found to be discriminatory, and applicants or employees may perceive algorithmic processes as unfair (Brown, 2024). Therefore, it is crucial for companies to recognize the potential for unfairness, discrimination, and employee dissatisfaction that can arise from algorithmic decision-making in HRM (Köchling & Wehner, 2020). Although the existing computer science literature has addressed issues related to biases, research on the potential downsides of algorithmic decision-making due to inherent biases in HRM is still in its infancy despite its growing importance with the increasing digitization and automation in the field (Köchling & Wehner, 2020).

To address this gap, the present study aimed to explore how discrimination may emerge from the implementation of AI-based decision-making and how these issues might be exacerbated by HRM professionals' perceptions of AI-related biases and their explicit attitudes toward such systems. This work represents one of the first attempts to address these issues. Additionally, the study sought to provide practical guidance on preventing discrimination and unfairness in AI-driven performance management while offering directions for future research, particularly in the HRM field. Although women are often disadvantaged by both algorithms and human decision-makers (e.g., Dastin, 2018; Shellenbarger, 2019), this study draws upon literature from various disciplines—including Management Studies, HRM, Organizational Psychology, Cognitive Psychology, Social Psychology, Business Ethics, and Ethics in AI. This study also aimed to

highlight the risks of discrimination and unfairness in algorithmic decision-making, especially within performance management and leadership development, and draw special attention to the impact these practices can have on women's career advancement.

Building on the research findings, concepts, and theories discussed above, the present study sought to investigate issues of bias and fairness in AI-based decision-making systems within HR functions, with a specific focus on performance evaluations and their perceived effectiveness in identifying leadership candidates. To achieve this, the research aimed to address *two key areas*: (a) the perception of AI-related biases among HR professionals and people managers, and (b) their comparative evaluation of AI systems versus human decision-making processes in terms of perceived qualities such as rationality, objectivity, and fairness.

Theories about the concept of leadership have evolved and been refined over time, with none being entirely irrelevant, as their usefulness largely depends on the context in which they are applied (Khan et al., 2016). It is important to bear in mind that contexts, work environments, situations, and organizational complexities can significantly impact the leadership concept and make it adaptable to changing organizational dynamics (Amabile et al., 2004). Among contemporary leadership styles, *transformational leadership* has emerged as one of the most prominent and stands out from other theories by focusing on fostering personal growth among followers within the organization (Khan et al., 2016). These leaders are often described as visionaries who guide their followers toward higher, more universal needs while elevating their and their teams' motivation and morality (House & Shamir, 1993; MacGregor Burns, 2003).

To accomplish the present study’s objectives, a scenario-based experimental design was employed in which participants were asked to assess promotion recommendations for a leadership role. These recommendations were generated either by a human agent or an AI agent

and featured candidate profiles differing in demographic characteristics along the dimension of gender. Specifically, the majority-group candidate (a man) was favored over the minority-status candidate (a woman) despite having a less favorable profile, thereby illustrating potential gender bias. Participants had to evaluate these recommendations based on several criteria, including rationality, scientific soundness, objectivity, impartiality, fairness, and trustworthiness. Additionally, the study included a post-scenario questionnaire to explore participants' explicit beliefs about AI's advantages concerning these qualities compared to human decision-makers.

The dual approach employed in this study aimed to provide comprehensive insights into how both implicit and explicit attitudes toward AI might influence promotion decisions and potentially contribute to workplace discrimination against historically marginalized groups. To this end, the study sought to address the following research questions:

1. How do HRM professionals and people managers perceive the rationality, scientific soundness, objectivity, impartiality, fairness, and trustworthiness of AI-based technologies in predicting leadership potential compared to human decision-making?
2. To what extent do HRM professionals and people managers recognize and acknowledge biases in AI systems that may affect promotion outcomes for leadership positions?
3. What are the explicit beliefs and attitudes of HRM professionals and people managers toward the use of AI in employee promotion processes, particularly regarding rationality, scientific soundness, objectivity, impartiality, fairness, and trustworthiness?

In addition, based on the research findings, concepts, and theories discussed above, the study developed and tested the following hypotheses:

- **Hypothesis 1 (H1):** HRM professionals will perceive promotion decisions made by an AI agent as more rational, scientifically sound, objective, impartial, fair, and trustworthy than identical decisions made by a human agent.
- **Hypothesis 2 (H2):** HRM professionals who perceive promotion decisions made by an AI agent as more rational, scientifically sound, objective, impartial, fair, and trustworthy than identical decisions made by a human agent will be more likely to endorse AI-based decision-making as superior in promotion contexts.

Materials and Methods

Sample

The sample consisted of 50 HRM professionals and people managers, with ages ranging from 21 to 60 (M = 37.16, SD = 10.11). Of the participants, 19 identified as men (38%) and 31 as women (62%). Regarding their level of education, one of them reported having a high school diploma or equivalent, two had further education (A-levels, BTEC, etc.), 17 held a Bachelor's degree, 25 had a Master's degree, two held a Professional degree (M.D., J.D., etc.), and three had a Doctoral degree (Ph.D., Ed.D., etc.). In terms of managerial experience, six participants reported having approximately one year of experience, five had one year, 11 had two years, 10 had five years, seven had approximately ten years, and 11 had over ten years of managerial experience.

The study aimed to include a diverse range of participants in terms of demographic characteristics such as age, gender, years of experience, level of education, and industry to ensure the obtained findings reflected a broad spectrum of views and experiences within HR and management roles. In addition, to be eligible, participants needed to be currently working in an

HR-related role or as a people manager, preferably with direct experience in employee evaluations, promotions, or leadership development processes. Moreover, primary interactions with participants occurred through online channels, including email and social media messaging. The study employed a combination of convenience sampling, whereby participants were recruited through online platforms and personal networks based on their availability and willingness to participate; purposive sampling, thereby targeting HRM professionals with specific roles and experience; and snowball sampling, with participants being recruited via word of mouth, thereby extending the participant pool through their networks.

More specifically, participants were recruited through word of mouth, personal emails, social media platforms like Facebook, professional networking sites like LinkedIn, and social news forums like Reddit. Invitations and study information were distributed through these platforms so as to allow participants to engage with the researcher at their convenience. Informed consent was obtained electronically prior to participation, ensuring all participants fully understood the nature of the study and voluntarily agreed to take part. Participation was both anonymous and strictly confidential, with participants receiving a survey link to complete the study remotely at a time and place of their choosing, such as their home or office, ensuring their privacy. Participants were also informed of their right to withdraw from the study at any time, further safeguarding their autonomy and privacy.

Procedure

The present study employed a scenario-based experimental method to test the hypotheses, simulating real-world decision-making in a controlled, hypothetical context. This approach was employed because of its high internal validity, allowing the manipulation of variables that would be otherwise challenging to test in a real-life setting due to ethical and logistical constraints.

Participants were randomly assigned to one of two conditions, each presenting a scenario involving an unfair promotion recommendation potentially influenced by gender bias, where the gender of the candidate was manipulated. In half of the cases, participants were told that a human decision-maker considered the man to be a more suitable candidate for the role, while in the other half, they were informed that an AI agent made this determination based on predictions of leadership potential. The scenarios were accompanied by a 16-item scale to explore participants' implicit beliefs about AI by asking them to evaluate the promotion decision's perceived qualities, such as rationality, objectivity, and fairness. Additionally, a post-scenario questionnaire was administered to further examine participants' explicit beliefs about AI's role in performance evaluation and leadership identification.

Moreover, upon accessing the link to the study, participants were welcomed, and after clicking "Let's proceed" below, they were presented with an informed consent form (see Appendix A), followed by a demographics questionnaire that was used to help understand the context of participants' views (see Appendix B). Following that, participants were provided with detailed instructions, and they were then sequentially presented with the two candidate profiles detailing their qualifications (see Appendix C) and asked to evaluate the promotion recommendation that was made by either a human or an AI agent, depending on their assigned condition (see Appendix D). Finally, after having evaluated the promotion recommendation, participants completed the post-scenario questionnaire assessing their explicit beliefs and attitudes about AI's role in promotion contexts (see Appendix E). The study then concluded with a thank-you message and confirmation of successful response submission.

Design

Implicit Attitudes Toward AI

Participants were presented with a biased promotion recommendation favoring a majority-status candidate over a minority-status candidate for a leadership position despite the latter having a stronger profile. The scenario presented two candidate profiles that differed in demographic characteristics, with the majority-group candidate (a man) being favored over the minority-status candidate (a woman) despite having a weaker profile, thereby revealing gender bias. The candidates' gender was evident by their names and pronouns. Participants in one condition were informed that a human decision-maker made the biased recommendation, while those in the other condition were told it was made by an AI agent. This between-subjects design aimed to assess how participants evaluated the decision in terms of rationality, scientific soundness, objectivity, impartiality, fairness, and trustworthiness.

To measure participants' implicit beliefs and attitudes toward AI-based decision-making by comparison with human-based decision-making, an investigator-developed 16-item scale was used. After reviewing the two candidate profiles, participants had to evaluate the recommendation on a 5-point Likert scale, ranging from 1 (= *Strongly disagree*) to 5 (= *Strongly agree*), indicating the extent to which they perceived it to have the aforementioned qualities by rating their agreement with statements such as "I find the AI agent's recommendation to be based on an objective consideration of all facts." The first item measured perceived fairness; items 2, 8, 9, 10, and 11 measured perceived objectivity; item 3 measured perceived impartiality; items 4 through 7 measured perceived scientific soundness; item 12 measured perceived rationality; and items 13 through 16 measured perceived trustworthiness.

Explicit Attitudes Toward AI

After they evaluated the promotion recommendation, participants were asked to complete a brief post-scenario questionnaire designed to gauge their explicit beliefs about AI's superiority

in the aforementioned qualities when evaluating performance and identifying leadership potential among employees. Participants had to rate their agreement with statements such as "Promotion decisions made by AI are more rational and fairer compared to human decision-making," again on a 5-point Likert scale ranging from 1 (= *Strongly disagree*) to 5 (= *Strongly agree*). This second investigator-developed scale consisted of four items: the first measured participants' beliefs about AI's superior rationality, objectivity, and fairness; the second measured their trust in AI's scientific soundness, impartiality, and trustworthiness; the third measured participants' views on potential AI bias; and the fourth measured their perceptions of AI's disparate impact on marginalized groups.

Results

All statistical analyses were conducted using IBM SPSS Statistics 29.0.1.

The internal consistency of the first scale, which measured participants' implicit attitudes toward AI-based decision-making (M = 44.28, SD = 12.56), was assessed using Cronbach's alpha. The result, α = .95, indicates excellent reliability, demonstrating that the scale items consistently measured the underlying construct. Additionally, the moderate variability in responses suggests that participants' scores were reasonably centered around the mean of 44.28, reflecting a relatively uniform set of attitudes. For the second scale, which measured explicit attitudes toward AI's fairness, rationality, and objectivity in decision-making (M = 12.52, SD = 2.71), Cronbach's alpha was found to be lower, at α = .55, indicating questionable internal consistency.

This low reliability may stem from the scale's small size (only four items) and its attempt to capture a complex, multi-dimensional construct. Notably, the scale comprised heterogeneous items, with two being reverse-scored to control for response bias. Although necessary for

thorough construct measurement, these items may have contributed to the lower alpha value due to the different facets of explicit attitudes they aimed to capture. The descriptive statistics (M = 12.52, SD = 2.71) indicate that participants' responses were moderately variable, with scores clustered relatively closely around the mean. Despite the low alpha, this suggests that participants generally shared similar explicit beliefs, though those beliefs may have been influenced by various facets of their understanding of AI decision-making.

To test H1, a Mann-Whitney U test was conducted to examine whether HRM professionals' ratings of the perceived qualities of promotion decisions differed based on the type of decision-maker (AI vs. human). The test revealed a statistically significant difference between the two groups, with the AI decision-maker receiving a significantly higher mean rank (M = 33.80) than the human decision-maker group (M = 17.20), (U = 105, $n1$ = 25, $n2$ = 25, z = -4.03, $p < .001$). The effect size, calculated as r = .58, indicates a large effect, suggesting a substantial difference in participants' perceptions of decision-making based on whether AI or a human was responsible. This result supports H1, as participants demonstrated a clear preference for AI-driven decision-making in terms of perceived qualities (see Table 1).

To test H2, a simple linear regression analysis was conducted to examine the predictive relationship between implicit and explicit attitudes toward AI use among participants. Specifically, this analysis was performed to evaluate whether implicit favoring of AI-based decision-making predicted explicit endorsement of AI's superiority over human decision-making in promotion decisions. A sample of 50 HRM professionals and people managers was analyzed, correlating implicit attitudes with corresponding explicit attitude scores. The regression coefficient for implicit attitudes was B = .06, (SE = .03), indicating that for every one-unit increase in implicit attitudes, there was a predicted increase of .06 units in explicit attitudes. This

positive relationship was statistically significant, $t(48) = 2.02$, $p = .05$, suggesting that implicit attitudes favoring AI were a significant predictor of explicit endorsement of AI in decision-making.

The mean age of the sample was 37 years ($SD = 10.11$), and the mean value for explicit attitudes was 11.52 ($SD = 2.8$). The regression model produced a statistically significant equation, ($F(1, 48) = 4.06$, $p = .05$), and explained approximately 8% of the variance in explicit attitudes, $R^2 = .08$. This result supports H2, demonstrating that implicit favoring of AI does predict explicit attitudes, though the effect size was small. The low-value R^2 indicates that implicit attitudes contribute meaningfully to explicit endorsement but do not constitute the primary drivers (see Table 2).

A multiple linear regression analysis was also carried out to explore whether participants' demographic characteristics, including gender, education level, age, and years of managerial experience, could predict their ability to perceive bias in promotion decisions. The results indicated that the demographic factors explained 31% of the variance in participants' bias perception in promotion decisions ($R^2 = .31$), indicating the proportion of variance explained. The overall model was statistically significant, $F(8, 41) = 2.31$, $p = .04$, suggesting a moderate correlation between the predictors and the outcome variable ($R = .56$), representing the multiple correlation coefficient between the independent variables and the dependent variable. Notably, gender, specifically identifying as a woman, was the only statistically significant variable ($B = -8.59$, $\beta = -.34$, $t = -2.18$, $p = .04$) (see Table 3).

Discussion

Organizations across a wide range of industries have been increasingly adopting AI tools to capture significant market share (McKinsey & Company, 2024). Sophisticated AI-driven

algorithms are revolutionizing decision-making processes by enhancing automation, with numerous studies showing algorithms consistently demonstrate superior performance compared to human decision-makers in various domains, which has led to their rapid implementation in business contexts (Grove et al., 2000; Stone et al., 2013). AI technologies have, therefore, become indispensable to modern organizational strategy, with HRM emerging as one of the key areas where AI is driving transformative changes across organizational functions.

AI-driven solutions are reshaping HRM processes by offering several advantages, such as reduced time and costs, increased productivity, and minimized risks (Suen et al., 2019; Woods et al., 2020). Beyond these economic advantages, organizations are also turning to AI, intending to reduce human biases and improve fairness, objectivity, and consistency in decision-making processes (Langer et al., 2019; Raghavan et al., 2020). However, a major challenge that has emerged is that algorithms can perpetuate human discrimination by producing biased outcomes based on the data they are trained on (Barocas & Selbst, 2016). These biases are further exacerbated by the lack of transparency and accountability in algorithmic decision-making (Pasquale, 2015). More often than not, training data reflects broader societal and historical inequities and prejudiced human judgments. In other words, models are frequently trained on data that reflect human decisions or the indirect consequences of systemic inequalities (Manyika et al., 2019).

Today, the vast majority of HR functions and business processes have been automated to varying degrees, with the aim of improving both effectiveness and efficiency across numerous activities, such as performance appraisals (Hendrickson, 2003). However, biased algorithmic decisions in HRM can lead to various problems, such as lack of transparency, critical misjudgments, and potential harm to an organization's reputation (Garg et al., 2021). Biases in

AI-powered performance management systems, specifically, have been shown to result in several detrimental outcomes, such as reduced trust in feedback quality, diminished reliability of the evaluation process (Minbaeva, 2021), and negative effects on fairness, trustworthiness, and organizational effectiveness (Zhang & Yencha, 2022). But despite these troubling findings, there remains a notable gap in understanding how AI-related biases and their perceptions impact the performance management process.

In spite of growing evidence that inherent human biases can often influence algorithmic decision-making, many people hold beliefs about algorithmic discrimination that may not align with objective reality (O'Neil, 2016). Psychological factors, such as stereotypes and intuitions about machines, can shape these beliefs, leading individuals to conclude that algorithms are less likely to discriminate than humans (Jago & Laurin, 2022). For instance, people are often misled into thinking that algorithms are objective and neutral arbiters in decision-making, free from biases and discriminatory attitudes. Therefore, the issues of bias and fairness in AI-based decision-making systems within HR functions constitute an increasingly relevant topic, especially considering the reputational and legal risks companies may face if their HR methods are found to be biased, unfair, or discriminatory (Brown, 2024). Nonetheless, research on inherent biases in algorithmic decision-making within HRM is still in its infancy, despite its growing significance amid the increasing digitization and automation in the field (Köchling & Wehner, 2020).

To address this significant gap in the extant literature, the current study set out to explore issues of bias and fairness in AI-based decision-making systems used in performance evaluations and their effectiveness in identifying leadership candidates. Specifically, the purpose of this study was to better understand how discrimination may arise from the implementation of AI-

based decision-making, as well as how these issues might be exacerbated by HRM professionals' ability, or lack thereof, to perceive AI-related biases and their attitudes toward such systems.

To achieve these objectives, a scenario-based experiment was conducted in which participants evaluated promotion recommendations for a leadership position made by either a human or an AI agent, employing a between-subjects design. The sample comprised 50 HRM professionals and people managers with diverse demographic characteristics recruited through word of mouth, personal emails, social media platforms, and professional networking sites, utilizing a combination of convenience sampling, purposive sampling, and snowball sampling methods. The scenario featured two candidate profiles that differed in demographic characteristics, specifically gender, with a man being favored over a woman despite having a weaker profile, thereby illustrating potential gender bias. Participants were asked to assess the rationality, scientific soundness, objectivity, impartiality, fairness, and trustworthiness of these recommendations. Additionally, the study included a short post-scenario questionnaire to explore participants' explicit beliefs about AI's advantages concerning these qualities compared to human decision-makers.

This dual approach aimed to provide comprehensive insights into how both implicit and explicit attitudes toward AI might influence promotion decisions and potentially contribute to workplace discrimination against women and minorities. It was hypothesized that (a) HRM professionals would perceive promotion decisions made by an AI agent as more rational, objective, impartial, fair, scientifically sound, and trustworthy than identical decisions made by a human agent, and (b) HRM professionals who hold these perceptions of AI would be more likely to endorse AI-based decision-making as superior in promotion contexts.

There are three key findings from the present research:

1. Consistent with the study's original hypothesis, the results showed that HRM professionals exhibited a clear preference for AI-driven decision-making. Participants rated the promotion recommendation made by the AI agent as significantly more rational, scientifically sound, objective, impartial, fair, and trustworthy than the identical decision made by a human agent in the other condition. Furthermore, the AI decision-maker group not only scored significantly higher than the human decision-maker group in these qualities, but the difference was also substantial.
2. In line with the study's second hypothesis, the results provided support by showing that implicit favoring of AI does, in fact, predict its explicit endorsement, though the effect size was small. The findings indicated that while implicit attitudes contributed meaningfully to the explicit endorsement of AI-based decision-making, they were not the primary drivers. One interpretation of this result is that other factors, such as personal experiences, societal views, or organizational context, likely play a more substantial role in shaping explicit beliefs about AI-based decision-making. Future research could thus benefit from exploring these additional variables to gain a deeper understanding of what factors might influence explicit attitudes toward the use of AI in performance evaluation and leadership potential identification.
3. The results also revealed that gender, specifically identifying as a woman, was the most significant predictor of the ability to perceive gender bias and unfairness in the promotion decision, with women being more likely to perceive bias compared to men. Identifying as a man did not significantly predict the ability to perceive bias in this model. Other demographic variables, such as participants' level of education, age, and amount of managerial experience, were also not found to contribute significantly. This suggests that,

given the substantial variance left unexplained, additional factors like organizational context, personal experiences with bias, or attitudes toward decision-making processes should be explored in future research. It also highlights that gender, at least in this sample, may be more influential in shaping perceptions of bias in promotion decisions than other demographic factors. This finding could be explained by the idea that women, due to their lived experiences with bias and discrimination, both professionally and personally, may perhaps be more attuned to recognizing gender bias occurrences in these contexts.

The present study raises awareness of potential biases and discrimination arising from algorithmic decision-making, as well as in relation to perceptions and attitudes toward it in HRM. The findings contribute to the existing literature by offering researchers and practitioners valuable insights into the risks associated with algorithmic decision-making in the HRM context, particularly with regard to ethical concerns such as bias, fairness, and discrimination in performance management and succession planning. Additionally, the results emphasize the growing importance of business ethics research related to algorithmic literacy (Cotter & Reisdorf, 2020) while also highlighting the need to address the risks of individuals being misled into uncritically accepting algorithmic evaluations.

Moreover, the first key finding of the present research provides strong evidence that mechanical objectivity and automation bias might have been at play, with participants perceiving the AI-generated decision as significantly less biased, fairer, and more objective than the exact same decision made by a human manager. In addition to that, the second key finding indicated that while implicit attitudes contributed meaningfully to the explicit endorsement of AI-based decision-making, they were not found to be the primary drivers. A potential explanation for this

could be the occurrence of *algorithmic aversion*. Algorithmic aversion, the opposite of automation bias, occurs when the machine heuristic leads to negative rather than positive reactions to AI (Dietvorst et al., 2015). There is evidence to suggest that algorithmic aversion can occur, with users favoring human decision-making over AI, even when AI's performance is actually better (Alvarado-Valencia & Barrero, 2014; Bucher, 2017; Dietvorst et al., 2015).

A common reason cited for the occurrence of algorithmic aversion is the violation of AI's "perfection scheme," which refers to the expectation that AI should always function perfectly (Dietvorst et al., 2015), which is linked to the concept of mechanical objectivity. This also often includes concerns about ethical issues related to relying on machines for important decisions (Dawes, 1979), as well as AI's perceived inability to integrate contextual factors (Grove & Meehl, 1996). Therefore, when AI fails, the initial high expectations placed on it are violated, leading to a significant decrease in people's preference for AI and resulting in the avoidance of its use. In contrast, when it comes to decision-making by human agents, people tend to have more realistic expectations, acknowledging that human beings often make errors in judgment. In this study, it is, therefore, possible that participants exposed to the AI decision-maker condition were essentially primed to view AI distrustfully due to the unfair promotion recommendation it offered, hence inducing algorithmic aversion, which then led to decreased explicit endorsement of AI-based decision-making among participants.

Implications for Practice

The findings of the present research suggest that integrating AI-driven solutions in HR departments must first be aligned with an organization's capacity to predict, detect, and mitigate potential biases in these systems in order to ensure fairness (Akter et al., 2021). This can be achieved through various methods, such as auditing the behavior of underlying algorithms using

empirically sound methodologies as well as diverse perspectives (Tuffaha, 2023). Additionally, it is essential to cultivate a heterogeneous and well-educated workforce capable of collaboratively scrutinizing, detecting, and addressing issues of bias and fairness, thereby minimizing the risk of potentially harmful effects (Rozado, 2020). It is, therefore, crucial for HR professionals to be able to effectively tackle AI-related biases to foster a fair and inclusive workplace while enhancing HR efficiency using AI-powered systems (Brown, 2024). Moreover, without thorough testing and diverse teams, unconscious biases can easily seep into ML models, resulting in AI systems that automate and perpetuate these biases (Marr, 2022). To this end, relevant literature emphasizes the need for data generated by AI-powered solutions to support decision-makers by providing key insights into strategic performance and developing metrics around critical success factors (Raffoni et al., 2018).

Bias not only harms those who are directly discriminated against but also has broader societal impacts by limiting people's participation in the economy and society. Therefore, collective action against bias is crucial. Bias undermines AI's potential for business and societal benefits by fostering mistrust and skewing results (Manyika et al., 2019). Eliminating bias in AI could also contribute to the enhancement of decision-making and provide significant benefits to traditionally disadvantaged groups, a concept referred to as the "disparate benefits from improved prediction" (Kleinberg et al., 2019). Therefore, business leaders ought to ensure that AI not only improves decision-making but also adheres to standards and research aimed at reducing bias (Manyika et al., 2019). Furthermore, extensive research on AI bias highlights two critical imperatives for action (Silberg & Manyika, 2019). First, AI must be used wisely to enhance traditional decision-making processes, as humans may be unaware of or deceptive about the factors influencing their decisions, such as hiring or rejecting job applicants (Silberg &

Manyika, 2019). Second, as previously discussed, advancing the mitigation of AI bias requires addressing complex challenges in defining and measuring fairness comprehensively (Silberg & Manyika, 2019).

AI undoubtedly has the potential to enhance management efficiency and reduce certain biases, but it requires very careful oversight. HRM professionals play a critical role in ensuring that AI-related biases do not result in discriminatory practices based on gender or other protected characteristics. To address these challenges, HR should collaborate with AI vendors to ensure that algorithms are trained on balanced, representative data, regularly update datasets to reflect diverse and current candidate pools, and meticulously clean data to eliminate historical biases (Brown, 2024). While AI holds great promise for HR by streamlining processes and enhancing decision-making, it is crucial to implement proactive measures to address potential biases and discrimination. This includes staying informed about legal developments, adhering to evolving regulations, providing thorough training, conducting regular audits, vetting algorithms, and ensuring detailed contracting (Brown, 2024). Ultimately, HRM professionals must balance AI's innovative benefits with fairness and compliance to ensure that AI positively impacts their organization's management practices and promotes a more inclusive and equitable workplace (Brown, 2024).

This work highlights ongoing research that addresses challenges related to AI bias and suggests practical solutions for moving forward. We still lack a full understanding of how the complex interactions between algorithms and societal structures operate. As a result, scholars have called for "algorithmic accountability" to deepen our understanding of the influences, biases, and power dynamics that algorithms exert in society (Diakopoulos, 2014). To this end, the current study sheds light on the complexities of these issues and emphasizes the need for

human vigilance in critically examining instances of unfair bias that often becomes baked in and perpetuated by AI systems.

As a final point, this paper also emphasizes two key opportunities. First, the opportunity to utilize AI to identify and mitigate human biases and their effects. Second, the opportunity to advance AI systems themselves so as to prevent them from reproducing and reinforcing human and societal biases, as well as creating new biases of their own. Realizing these opportunities, of course, requires interdisciplinary collaboration to advance ethical standards and operational practices. Additionally, business leaders can support progress by making more data available to practitioners and researchers working on these issues while staying mindful of potential risks and privacy concerns (Silberg & Manyika, 2019).

Limitations and Future Directions

The findings of this study should be considered in light of some limitations. First, a significant limitation pertains to the normality of the data. The Kolmogorov–Smirnov test that was conducted to assess normality revealed that the data was not normally distributed. This non-normality could be attributed to the small sample size of this pilot study (25 participants assigned in each condition). Nevertheless, this study provided valuable insights into the potential outcomes of a larger-scale research project by testing the efficacy of research methods, instruments, and procedures that could be improved upon in future studies. Additionally, another limitation pertains to the study's sample being unbalanced in terms of gender, with participants identifying as women making up the majority. As a result, the findings may primarily reflect the perspectives or experiences of this group, potentially skewing conclusions and recommendations. This imbalance could potentially limit the generalizability of the obtained results to the entire

population. Future research should, therefore, aim to include a more balanced sample in terms of gender.

Moreover, another limitation was that the Cronbach's alpha for the second investigator-developed scale, which measured explicit attitudes among participants, was lower than the typically accepted threshold ($\alpha \geq .70$), raising concerns about the scale's internal consistency. However, this could be attributed to the scale being composed of heterogeneous items. While these items were necessary for thorough construct measurement, they may have contributed to the lower alpha value due to the different facets of the explicit attitudes they aimed to capture. Explicit attitudes toward AI decision-making likely encompass several dimensions—such as fairness, trust, objectivity, and concerns about bias—that may not be perfectly correlated. Therefore, the complexity of the construct, along with the small number of items, could explain the lower alpha value. This limitation could also be addressed in future research.

Moreover, the use of convenience, purposive, and snowball sampling in this study constitutes non-random methods that could potentially introduce bias. Since participants were recruited through personal networks, online platforms, and word-of-mouth, the sample may not fully represent the broader population of HRM professionals, which could potentially limit the generalizability of these findings. Additionally, gender was found to be a significant predictor of bias perceptions, while other demographic factors, such as age, education, and managerial experience, were not. This raises the question of whether other unexamined variables, such as race, cultural background, and industry type, might influence attitudes toward AI decision-making and bias perception. Future research could therefore benefit from including a broader range of demographic and contextual variables. Moreover, the study measured participants' attitudes immediately after exposure to the scenario but could not of course assess whether these

attitudes remained stable over time. Future studies could, therefore, incorporate longitudinal designs to examine if initial perceptions of AI-driven decision-making evolve with greater exposure or real-world experience.

Conclusion

All in all, this study highlights critical challenges related to bias and fairness in AI-driven decision-making as perceived by HRM professionals. As AI technologies become integral to organizational strategies and are increasingly adopted in HR processes—such as performance appraisals and promotion decisions—it is paramount to be aware of how AI systems can perpetuate human biases. The findings of this work contribute to the growing body of research on AI biases in HRM by offering new insights into how professionals perceive issues of bias and fairness in algorithmic decision-making. Furthermore, this work emphasizes the importance of ongoing research to address the ethical implications of AI in HRM and highlights the need for interdisciplinary collaboration to effectively tackle these challenges. To mitigate AI-related biases, organizations must take proactive steps through comprehensive auditing, diverse teams, and ongoing monitoring, to ensure that AI systems enhance fairness without replicating harmful biases.

Ultimately, the effective and ethical integration of AI in HRM requires a balance between leveraging AI's capabilities and addressing its potential pitfalls to support more inclusive and equitable workplace practices. Advancing the field necessitates conducting empirical research and systematically reviewing existing knowledge on biases and discrimination in AI-based decision-making while identifying new research avenues. Minimizing AI biases is crucial for building trust in these systems, as confidence in their reliability is essential for unlocking their full potential, driving business and economic growth through productivity gains, and addressing

pressing societal issues. Interdisciplinary collaboration, involving social scientists, ethicists, and other experts, is therefore needed to move the field forward, given the complexities of each application area.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. M. (2018). A reductions approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning, 80,* (pp. 60–69). https://doi.org/10.48550/arXiv.1803.02453

Aktepe, A., & Ersoz, S. (2012). A quantitative performance evaluation model based on a job satisfaction-performance matrix and application in a manufacturing company. *International Journal of Industrial Engineering: Theory, Applications, and Practice, 19*(6), 637. https://doi.org/10.23055/ijietap.2012.19.6.637

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y., D'Ambra, J., & Shen, K. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management, 60,* 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior, 36,* 102–113. https://doi.org/10.1016/j.chb.2014.03.047

Amabile, T. M., Schatzel, E. A., Moneta, G. B., & Kramer, S. J. (2004). Leader behaviors and the work environment for creativity: Perceived leader support. *The Leadership Quarterly, 15*(1), 5–32. https://doi.org/10.1016/j.leaqua.2003.12.003

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals. And it's biased against*

blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Appio, F., La Torre, D., Lazzeri, F., Masri, H., & Schiavone, F. (2024). *Impact of artificial intelligence in business and society: Opportunities and challenges*. Routledge. https://doi.org/10.4324/9781003304616

Arrow, K. J. (1973). The theory of discrimination. *Discrimination in Labor Markets, 3*(10), 3–33. https://www.scirp.org/reference/referencespapers?referenceid=580126

Atluri, V., Dahlström, P., Gaffey, B., García de la Torre, V., Kaka, N., Lajous, T., Singla, A., Sukharevsky, A., Travasoni, A., & Vieira, B. (2024, February 22). *Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom*. McKinsey & Company. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/beyond-the-hype-capturing-the-potential-of-ai-and-gen-ai-in-tmt

Banks, J. (2020). Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI, 7*, Article 62. https://doi.org/10.3389/frobt.2020.00062

Barbara Wimmer, B. (2018, December 6). *Computer says no: Algorithm gives women fewer chances at the AMS*. Futurezone.at. https://futurezone.at/netzpolitik/computer-sagt-nein-algorithmus-gibt-frauen-weniger-chancen-beim-ams/400345297

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732. http://www.jstor.org/stable/24758720

Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, *64*(1), 87–123. https://doi.org/10.1177/0001839217751692

Bellur, S., & Sundar, S. S. (2014). How can we tell when a heuristic has been used? Design and analysis strategies for capturing the operation of heuristics. *Communication Methods and Measures*, *8*(2), 116–137. https://doi.org/10.1080/19312458.2014.903390

Bertrand M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. https://www.nber.org/papers/w9873

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit Discrimination. *American Economic Review*, *95*(2), 94–98. https://doi.org/10.1257/000282805774670365

Bohnet, I., van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225–1234. https://doi.org/10.1287/mnsc.2015.2186

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *30th Conference on Neural Information Processing Systems*, 4356–4364. https://doi.org/10.48550/arXiv.1607.06520

Brown, J. J. (2024, August 26). *The role of HR in managing AI and mitigating bias*. HR Daily Advisor. https://hrdailyadvisor.blr.com/2024/08/26/the-role-of-hr-in-managing-ai-and-mitigating-bias/

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W W Norton & Co. https://wwnorton.com/books/the-second-machine-age/

Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*, 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Budhwar, P., Malik, A., De Silva, M. T. T., & Thevisuthan, P. (2022). Artificial intelligence – challenges and opportunities for international HRM: A review and research agenda. *The International Journal of Human Resource Management*, *33*(6), 1065–1097. https://doi.org/10.1080/09585192.2022.2035161

Buhmann, A., Paßmann, J., & Fieseler, C. (2020). Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse. *Journal of Business Ethics*, *163*(2), 265–280. https://doi.org/10.1007/s10551-019-04226-4

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, (pp. 77–91). https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Burton-Harris, V., & Mayor, P. (2020, June 24). *Wrongfully arrested because face recognition can't tell black people apart*. American Civil Liberties Union. https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292. https://doi.org/10.1007/s10618-010-0190-x

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *2009 IEEE International Conference on Data Mining Workshops*, 13–18. https://doi.org/10.1109/ICDMW.2009.83

Calders, T., Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In Custers, B., Calders, T., Schermer, B., Zarsky, T. (Eds.), *Discrimination and privacy in the information society* (pp. 43–57). Springer. https://doi.org/10.1007/978-3-642-30487-3_3

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://www.science.org/doi/10.1126/science.aal4230

Calmon, F. D., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Neural Information Processing Systems*. https://dl.acm.org/doi/pdf/10.5555/3294996.3295155

Carey, D. & Smith, M. (2016, April 22). *How companies are using simulations, competitions, and analytics to hire*. Harvard Business Review. https://hbr.org/2016/04/how-companies-are-using-simulations-competitions-and-analytics-to-hire

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, *134*(3), 1163–1224. https://doi.org/10.1093/qje/qjz008

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Chander, A. (2017). The racist algorithm? *Michigan Law Review*, *115*(6), 1023–1045. http://www.jstor.org/stable/44984908

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.48550/arXiv.1703.00056

Christin, A. (2016). From daguerreotypes to algorithms: Machines, expertise, and three forms of objectivity. *ACM Computers & Society*, *46*(1), 27–32. https://doi.org/10.1145/2908216.2908220

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023, June 14). *The economic potential of generative AI: The next productivity frontier*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-

insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

Citron, D. K., & Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, *89*(1), 1–33. https://ssrn.com/abstract=2376209

Civin, D. (2018, May 21). *Explainable AI could reduce the impact of biased algorithms*. VentureBeat. https://venturebeat.com/ai/explainable-ai-could-reduce-the-impact-of-biased-algorithms/

Cloudy, J., Banks, J., & Bowman, N. D. (2021). The str(AI)ght scoop: Artificial intelligence cues reduce perceptions of hostile media bias. *Digital Journalism*, *11*(9), 1577–1596. https://doi.org/10.1080/21670811.2021.1969974

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, *24*(1), 14730–14846. https://doi.org/10.48550/arXiv.1808.00023

Cotter, K., & Reisdorf, B. C. (2020). Algorithmic knowledge gaps: A new horizon of (digital) inequality. *International Journal of Communication*, *14*, 745–765. https://ijoc.org/index.php/ijoc/article/view/12450/2952

Cropanzano, R., Bowen, D. E., & Gilliland, S. W. (2007). The management of organizational justice. *The Academy of Management Perspectives*, *21*(4), 34–48. https://doi.org/10.5465/AMP.2007.27895338

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, *36*(3), 316–328. https://doi.org/10.1006/jesp.1999.1418

Dastin, J. (2018, October 11). *Insight - Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations*, *40*, 81–128. https://doi.org/10.2307/2928741

Daugherty, P. R., & Wilson, H. J. (2018, March 20). *Human + machine: Reimagining work in the age of AI*. Harvard Business Review. https://store.hbr.org/product/human-machine-reimagining-work-in-the-age-of-ai/10163

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Demetis, D. S., & Lee, A. S. (2018). When humans using the IT artifact becomes IT using the human artifact. *Journal of the Association for Information Systems*, *19*(10), 929–952. https://doi.org/10.17705/1jais.00514

Diakopoulos, N. (2014). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319. https://doi.org/10.1111/1467-9280.00262

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds & Machines*, *28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Ford, M. (2015). *Rise of the robots: Technology and the threat of a jobless future*. Basic Books. http://digamo.free.fr/marford15.pdf

Frey, C. B., & Osborne, M. A. (2016). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2021). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*, *71*(5), 1590–1610. https://doi.org/10.1108/IJPPM-08-2020-0427

Garr, S. S., & Jackson, C. (2019). *Diversity & inclusion technology: The rise of a transformative market*. Mercer. https://info.mercer.com/rs/521-DEV-513/images/Mercer_DI_Report_Digital.pdf

Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepi, A. (2021). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, 1–26. https://doi.org/10.1080/0960085X.2021.1927213

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review*, *18*(4), 694–734. https://doi.org/10.2307/258595

Goldman, B. M., Gutek, B. A., Stein, J. H., & Lewis, K. (2006). Employment discrimination in organizations: Antecedents and consequences. *Journal of Management*, *32*(6), 786–830. https://doi.org/10.1177/0149206306293544

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). *Implicit Association Test (IAT)* [Database record]. APA PsycTests. https://doi.org/10.1037/t03782-000

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The

clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Gualtieri, M. (2021, April 7). *The evolution of ML platforms to AI platforms: A spectrum of capabilities* [On-demand webinar]. Forrester Research, Inc. https://www.forrester.com/webinar/The%2BEvolution%2BOf%2BML%2BPlatforms%2BTo%2BAI%2BPlatforms%2BA%2BSpectrum%2BOf%2BCapabilities/WEB33085

Gunton, R., M., Stafleu, M. D. & Reiss, M. J. (2021) A general theory of objectivity: Contributions from the reformational philosophy tradition. *Foundations of Science*, *27*(3), 1–15. https://doi.org/10.1007/s10699-021-09809-x

Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication*, *1*, 37–54. https://doi.org/10.30658/hmc.1.3

Hendrickson, A. R. (2003) Human resource information systems: Backbone technology of contemporary human resources. *Journal of Labor Research*, *24*(3), 382–394. http://dx.doi.org/10.1007/s12122-003-1002-5

Hoffman, B. (2024, March 10). *Automation bias: What it is and how to overcome it*. Forbes. https://www.forbes.com/sites/brycehoffman/2024/03/10/automation-bias-what-it-is-and-how-to-overcome-it/

House, R. J., & Shamir, B. (1993). Toward the integration of transformational, charismatic, and visionary theories. In M. M. Chemers & R. Ayman (Eds.), *Leadership theory and research: Perspectives and directions* (pp. 81–107). Academic Press.

Hsu, J. (2020). Can AI hiring systems be made antiracist? Makers and users of AI-assisted recruiting software reexamine the tools' development and how they're used - [News]. *IEEE Spectrum*, *57*(9), 9–11. https://doi.org/10.1109/MSPEC.2020.9173891

Jago, A. S., & Laurin, K. (2022). Assumptions about algorithms' capacity for discrimination. *Personality and Social Psychology Bulletin*, *48*(4), 582–595. https://doi.org/10.1177/01461672211016187

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with Applications in R*. Springer Science & Business Media. https://www.stat.berkeley.edu/users/rabbee/s154/ISLR_First_Printing.pdf

Jones-Jang, S. M. & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, *28*(1), 1–8. https://doi.org/10.1093/jcmc/zmac029

Jones, K. P., Peddie, C. I., Gilrane, V. L., King, E. B., & Gray, A. L. (2016). Not so subtle: A meta-analytic investigation of the correlates of subtle and overt discrimination. *Journal of Management*, *42*(6), 1588–1613. https://doi.org/10.1177/0149206313506466

Kaibel, C., Koch-Bayram, I., Biemann, T., & Mühlenbock, M. (2019). Applicant perceptions of hiring algorithms - Uniqueness and discrimination experiences as moderators. *Academy of Management Proceedings*, *2019*(1), 18172. https://doi.org/10.5465/ambpp.2019.210

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *2010 IEEE International Conference on Data Mining*, 869–874. https://doi.org/10.1109/ICDM.2010.50

Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2018). Exploiting reject option in classification for social discrimination control. *Information Sciences*, *425*(3), 18–33. https://doi.org/10.1016/j.ins.2017.09.064

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2012. Lecture Notes in Computer Science*, *7524*, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

Kaplan A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Khan, Z. A., Nawaz, A., & Khan, I. (2016). Leadership theories and styles: A literature review. *Journal of Resources Development and Management*, *16*, 1–7. https://core.ac.uk/download/pdf/234696192.pdf

Kim, J.-Y., & Heo, W. (2022). Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians. *Information Technology & People*, *35*(3), 861–878. https://doi.org/10.1108/ITP-04-2019-0173

Kim, P. T. (2016). Data-driven discrimination at work. *William & Mary Law Review*, *58*(3), 857. https://scholarship.law.wm.edu/wmlr/vol58/iss3/4

Kite, M. E., & Whitley, B. E., Jr. (2016). *Psychology of prejudice and discrimination* (3rd ed.). Psychology Press. https://books.google.co.zw/books?id=yHlTDAAAQBAJ&printsec=frontcover#v=onepage&q&f=false

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings*, *108*, (pp. 22–27). https://doi.org/10.1257/pandp.20181018

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174. https://doi.org/10.1093/jla/laz001

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS), 67,* (pp. 43:1–43:23). https://doi.org/10.48550/arXiv.1609.05807

Köchling, A., & Wehner, M. C. (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. *BuR – Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kuang, C. (2017, November 21). *Can A.I. be taught to explain itself?* The New York Times. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), 217–234. https://doi.org/10.1111/ijsa.12246

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). *How we analyzed the COMPAS recidivism algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 1–16. https://doi.org/10.1177/2053951718756684

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics*, *160*(2), 377–392. https://doi.org/10.1007/s10551-019-04204-w

Lindebaum, D., Vesa, M., & Den Hond, F. (2020). Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review*, *45*(1), 247–263. https://doi.org/10.5465/amr.2018.0181

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *BMJ. British Medical Journal*, *296*(6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. *Knowledge Discovery and Data Mining*, 502–510. https://doi.org/10.1145/2020408.2020488

Mantzaris, K., & Myloni, B. (2023). Human vs technology: A cross-cultural comparison of HR professionals' perceptions. *International Journal of Manpower*, *44*(1), 58–76. https://doi.org/10.1108/IJM-05-2020-0197

Manyika, J., Silberg, J., & Presten, B. (2019, October 25). *What do we do about the biases in AI?* Harvard Business Review. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Marr, B. (2022, September 30). *The problem with biased AIs (and how to make AI better)*. Forbes. https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=685ec6214770

Martin, K. E. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, 160(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

Massrur, R., Nejad, A. F., & Sami, A. (2014). The surveying of the effect of the incentive pays to the degree of the attraction of resources in bank branches through the data mining technique. *2014 Iranian Conference on Intelligent Systems (ICIS)*, 1–6. https://doi.org/10.1109/IranianCIS.2014.6802590

McColl, R., & Michelotti, M. (2019). Sorry, could you repeat the question? Exploring video-interview recruitment practice in HRM. *Human Resource Management Journal*, *29*(4), 637–656. https://doi.org/10.1111/1748-8583.12249

McGregor, D. (1960) *The human side of enterprise.* McGraw-Hill Book Co.

McKinsey & Company. (2024, April 2). *What is generative AI?* *https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai#/*

Minbaeva, D. (2021). Disrupted HR? *Human Resource Management Review*, *31*(4), 100820. https://doi.org/10.1016/j.hrmr.2020.100820

Möhlmann, M., & Zalmanson, L. (2017). *Hands on the wheel: Navigating algorithmic management and Uber drivers' autonomy. In Proceedings of the International Conference on Information Systems (ICIS)*. (pp. 10–13). https://aisel.aisnet.org/icis2017/DigitalPlatforms/Presentations/3/

Moon, W. K., Chung, M., & Jones-Jang, S. M. (2022). How can we fight partisan biases in the COVID-19 pandemic? AI source labels on fact-checking messages reduce motivated reasoning. *Mass Communication and Society*, *26*(4), 646–670. https://doi.org/10.1080/15205436.2022.2097926

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(41), (pp. 16474–16479). https://doi.org/10.1073/pnas.1211286109

Nasir, S. Z. (2017). Emerging challenges of HRM in 21st century: A theoretical analysis. *The International Journal of Academic Research in Business and Social Sciences*, *7*(3), 216–223.

https://hrmars.com/papers_submitted/2727/Emerging_Challenges_of_HRM_in_21st_Century.pdf

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., . . . Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery/Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1356

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group. https://doi.org/10.5860/crl.78.3.403

Ozkazanc-Pan, B. (2019). Diversity and future of work: Inequality abound or opportunities for all? *Management Decision*, *59*(11), 2645–2659. https://doi.org/10.1108/MD-02-2019-0244

Packer, B., Halpern, Y., Guajardo-Céspedes, M., & Mitchell, M. (2018, April 13). *Text embedding models contain bias. Here's why that matters*. Google for Developers. https://developers.googleblog.com/en/text-embedding-models-contain-bias-heres-why-that-matters/

Paschen, U., Pitt, C., & Kietzmann, J. (2020). Artificial intelligence: Building blocks and an innovation typology. *Business Horizons*, *63*(2), 147–155. https://doi.org/10.1016/j.bushor.2019.10.004

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://www.hup.harvard.edu/books/9780674970847

Persson, A. (2016). Implicit bias in predictive data profiling within recruitments. In A. Lehmann, D. Whitehouse, S. Fischer-Hübner, L. Fritsch, & C. Raab (Eds.), *Privacy and identity management. Facing up to next steps. Privacy and identity 2016* (Vol. 498). Springer. https://doi.org/10.1007/978-3-319-55783-0_15

Pethig, F., & Kroenung, J. (2022). Biased humans, (un)biased algorithms? *Journal of Business Ethics*. Advance online publication. https://doi.org/10.1007/s10551-022-05071-8

Petruzzellis, S., Licchelli, O., Palmisano, I., Semeraro, G., Bavaro, V., & Palmisano, C. (2006). Personalized incentive plans through employee profiling. *Proceedings of the Eighth International Conference on Enterprise Information Systems - AIDSS*, (pp. 107–114). https://doi.org/10.5220/0002493401070114

Qamar, Y., Agrawal, R. K., Samad, T. A., & Jabbour, C. J. (2021). When technology meets people: The interplay of artificial intelligence and human resource management. *Journal of Enterprise Information Management*, *34*(5), 1339–1370. https://doi.org/10.1108/JEIM-11-2020-0436

Raffoni, A., Visani, F., Bartolini, M., & Silvi, R. (2017). Business performance analytics: Exploring the potential for performance management systems. *Production Planning & Control*, *29*(1), 51–67. https://doi.org/10.1080/09537287.2017.1381887

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). *Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. (pp. 469–481). https://doi.org/10.1145/3351095.3372828

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Reiter, A. (2019, May 11). *The office and my data*. Zeit Online. https://www.zeit.de/2019/20/digitale-verwaltung-behoerden-aemter-effizienzsteigerung-probleme

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 12). *Local interpretable model-agnostic explanations (LIME): An introduction*. O'Reilly Media. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Roehling, M. V., Boswell, W. R., Caligiuri, P., Feldman, D., Graham, M. E., Guthrie, J. P., Morishima, M., & Tansky, J. W. (2005). The future of HR management: Research needs and directions. *Human Resource Management*, *44*(2), 207–216. https://doi.org/10.1002/hrm.20066

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access, 8*, 42200–42216. https://ieeexplore.ieee.org/document/9007737

Rozado, D. (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLoS ONE*, *15*(4), e0231189. https://doi.org/10.1371/journal.pone.0231189

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. (3rd ed.). Prentice-Hall. https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

Sakka, F., El Maknouzi, M. E., & Sadok, H. (2022). Human resource management in the era of artificial intelligence: Future HR work practices, anticipated skill set, financial and legal implications. *Academy of Strategic Management Journal*, *21*(1), 1–14. https://www.abacademies.org/articles/human-resource-management-in-the-era-of-artificial-intelligence-future-hr-work-practices-anticipated-skill-set-financial-and-legal-13536.html

Savage, D. D., & Bales, R. (2017). Video games in job interviews: Using algorithms to minimize discrimination and unconscious bias. *ABA Journal of Labor & Employment Law*, *32*(2), 211–228. http://www.jstor.org/stable/44648549

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Shellenbarger, S. (2019, February 13). *A crucial step for averting AI disasters*. The Wall Street Journal. https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865

Silberg, J., & Manyika, J. (2019, June 6). *Tackling bias in artificial intelligence (and in humans)*. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans?cid=eml-web

Silverman, R. E., & Waller, N. (2015, March 13). *The algorithm that tells the boss who might quit*. The Wall Street Journal. https://www.wsj.com/articles/the-algorithm-that-tells-the-boss-who-might-quit-1426287935

Simbeck, K. (2019). HR analytics and ethics. *IBM Journal of Research and Development, 63*(4/5), 1–9. https://doi.org/10.1147/JRD.2019.2915067

Singla, A., Sukharevsky, A., Yee, L., & Chui, M. (2024, May 30). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Sprenger, J. M., & Reiss, J. (2020). Scientific Objectivity. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2020 ed.) https://plato.stanford.edu/entries/scientific-objectivity/

Stone, D. L., Lukaszewski, K. M., Stone-Romero, E. F., & Johnson, T. L. (2013). Factors affecting the effectiveness and acceptance of electronic selection systems. *Human Resource Management Review*, *23*(1), 50–70. https://doi.org/10.1016/j.hrmr.2012.06.006

Suen, H., Chen, M. Y., & Lu, S. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, 93–101. https://doi.org/10.1016/j.chb.2019.04.012

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019)*. *538*, (pp. 1–9). https://doi.org/10.1145/3290605.3300768

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *ArXiv*, *abs/1901.10002*. https://doi.org/10.48550/arXiv.1901.10002

Sweeney, L. (2013). Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *ACM Digital Library*, *11*(3), 10–29. https://doi.org/10.1145/2460276.2460278

The Royal Society. (2017, April). *Machine learning: The power and promise of computers that learn by example*. https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, *42*(2), 1600–1631. https://doi.org/10.1002/smj.3322

Tuffaha M. (2023). The impact of artificial intelligence bias on human resource management functions: Systematic literature review and future research directions. *European Journal*

of Business and Innovation Research, *11*(4), 35–58. https://doi.org/10.37745/ejbir.2013/vol11n43558

Tuffaha, M., Perello-Marin, M., & Suarez-Ruz, E. (2022). Key elements in transferring knowledge of the AI implementation process for HRM in COVID-19 times: AI consultants' perspective. *International Journal of Business Science and Applied Management*, *17*(1), 81–97. https://doi.org/10.69864/ijbsam.17-1.159

Varathan, P. (2017, August 27). *Most stay-at-home dads now actually want to stay at home*. Quartz. https://qz.com/1061668/most-stay-at-home-dads-now-actually-want-to-stay-at-home

Walker, J. (2012, September 20). *Meet the new boss: Big data*. The Wall Street Journal. https://www.wsj.com/articles/SB10000872396390443890304578006252019616768

Weisshaar, K. (2018, February 22). *Stay-at-home moms are half as likely to get a job interview as moms who got laid off*. Harvard Business Review. https://hbr.org/2018/02/stay-at-home-moms-are-half-as-likely-to-get-a-job-interview-as-moms-who-got-laid-off

Westerman, D., Edwards, A. P., Edwards, C., Luo, Z., & Spence, P. R. (2020). I-It, I-Thou, I-Robot: The perceived humanness of AI in human-machine communication. *Communication Studies*, *71*(3), 393–408. https://doi.org/10.1080/10510974.2020.1749683

Wojcieszak, M., Thakur, A., Ferreira Gonçalves, J. F., Casas, A., Menchen-Trevino, E., & Boon, & M. (2021). Can AI enhance people's support for online moderation and their openness

to dissimilar political views? *Journal of Computer-Mediated Communication*, *26*(4), 223–243. https://doi.org/10.1093/jcmc/zmab006

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*(1), 64–77. https://doi.org/10.1080/1359432X.2019.1681401

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web*, (pp. 1171–1180). https://doi.org/10.1145/3038912.3052660

Zemel, R. S., Wu, L. Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *International Conference on Machine Learning*, *28*(3), 325–333. https://www.cs.toronto.edu/~toni/Papers/icml-final.pdf

Zhang, L., & Yencha, C. (2022). Examining perceptions towards hiring algorithms. *Technology in Society*, *68*, 101848. https://doi.org/10.1016/j.techsoc.2021.101848

Table 1

Results of the Mann-Whitney U Test Comparing Perceived Qualities of AI and Human Decision-Making

Group	*N*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*Z*	*p*
AI	25	33.80	845.00	105.00	-4.03	< .001
Human	25	17.20	430.00			

Note. The results suggest that the AI decision-maker group has significantly higher scores than the human decision-maker group.

Table 2

Results of the Simple Linear Regression Predicting Participants' Explicit Endorsement of AI Based on Their Implicit Attitudes Toward AI

Predictor	*B*	*SE*	*β*	*t*	*p*
Intercept	11.52	2.8	–	–	–
Implicit Attitudes	.06	.03	–	2.02	.049

Note. The results suggest that the model explained 7.8% of the variance in explicit endorsement of AI-based decision-making (R^2 = .078), $F(1, 48)$, p = .049, indicating that the model is significant but with a small effect size.

Table 3

Results of Multiple Linear Regression Analysis Predicting Participants' Bias Perception Based on their Demographic Characteristics

Variable	*B*	*SE*	*β*	*t*	*p*
Constant	56.71	9.71		5.84	<.001
Age	-.14	.28	-.12	-0.51	.61
Experience	-.79	.68	-.24	-1.16	.25
Gender (Woman)	-8.59	3.95	-.34	-2.18	.04
High school	-5.31	12.07	-.06	-.44	.66
Further_Ed	4.11	9.25	.07	.44	.66
BS	5.81	3.89	.22	1.50	.14
Professional_D	7.73	9.75	.12	.79	.43
Doctoral_D	-3.63	7.06	-.07	-.51	.61
R^2	.31				
Adj. R	.18				
F	2.31				

Note. The results suggest a moderate correlation between the predictors and the outcome, with identifying as a woman being the only significant variable ($B = -8.59$, $\beta = -.34$, $t = -2.18$, $p = .04$).

Appendix A

Informed Consent Form

Please take the time to carefully read this informed consent form. If you have any questions or need clarification, don't hesitate to reach out to the researcher conducting this study.

Purpose of the Study
My name is Ioannis Barrett, and I'm inviting you to participate in a study I'm conducting as part of my master's thesis in Organizational Psychology. The purpose of this study is to explore the effectiveness of various evaluation methods in identifying leadership potential among employees.

Procedure
If you agree to participate in this study, you'll be first asked to review performance feedback materials for two employees who have applied for promotion to a leadership position within their organization. These materials, completed by the employees' coworkers, pertain to one employee being considered for the Level III Procurement Administrator role. You'll then have to assess the degree to which you agree with the recommendation for promoting one of these two employees to this role. Following this, you'll be asked to complete a short questionnaire about your views on using different systems for identifying suitable candidates for leadership roles. The entire study will take approximately 30 minutes.

Potential Benefits/ Risks
There are no known risks associated with participating in this study. By participating, you'll gain valuable experience in psychological research and contribute to enhancing our understanding of the role of different systems in performance management and evaluation.

Anonymity/ Confidentiality
This study is anonymous. You'll only be asked to provide some demographic information without revealing any personal details. Only the researcher and the study supervisor will have access to the data, which will be stored securely.

Voluntary Participation/ Right to Withdraw
Your participation is voluntary. You have the right to withdraw at any time without explanation or consequence.

Contacts and Questions
If you have any questions or concerns about the study, either during or after your participation, please get in touch with me at I.Barrett@acg.edu. Alternatively, you may contact the project supervisor at Okyriakidou@acg.edu.

*** Consent**
I, the undersigned, freely agree to participate in the study described above. I confirm that I have read and understood the information and am willing to participate.

- o Yes, I agree to participate
- o No, I do not agree to participate

Appendix B

Participant Demographics

Demographic Questions

Please provide some general demographic information about yourself by selecting the option that best applies to you. If you're stuck between two options, please select the one you feel best describes you.

1. What is your gender?

- Man
- Woman
- Non-binary
- Prefer to self-describe: _____
- Prefer not to disclose

2. What is your age?

- 18–24 years
- 25–34 years
- 35–44 years
- 45–54 years
- 55–64 years
- 65 years or older
- Prefer not to disclose

3. What is the highest level of education you have completed?

- High school diploma or equivalent
- Further education (A-levels, BTEC, etc.)
- Associate's degree
- Bachelor's degree
- Master's degree
- Professional degree (M.D., J.D., etc.)
- Doctoral degree (Ph.D., Ed.D., etc.)
- Prefer not to disclose

4. What is your level of managerial responsibility?

- Individual contributor (no managerial responsibility)
- First-line manager (supervises team members or individual contributors)

- Middle manager (manages first-line managers)
- Senior manager/executive (manages middle managers, heads of departments)
- Executive leadership (C-suite, VP, etc.)
- Prefer not to disclose

5. How many years of managerial experience do you have?

- Less than 1 year
- 1–3 years
- 4–6 years
- 7–10 years
- More than 10 years
- Prefer not to disclose

6. The industry you work in:

- Technology (e.g., IT, Software, Hardware)
- Healthcare (e.g., Hospitals, Pharmaceuticals, Medical Devices)
- Education (e.g., K-12, Higher Education)
- Finance (e.g., Banking, Insurance, Investments)
- Manufacturing (e.g., Automotive, Consumer Goods, Industrial Equipment)
- Retail (e.g., E-commerce, Brick-and-Mortar Stores)
- Public Sector (e.g., Government, Non-profit, NGOs)
- Professional Services (e.g., Consulting, Legal, Accounting)
- Transportation & Logistics (e.g., Shipping, Warehousing)
- Energy & Utilities (e.g., Oil & Gas, Renewable Energy)
- Other: _____
- Prefer not to disclose

7. What is the type and size of your organization?

- Startup (1–50 employees)
- Small business (51–200 employees)
- Medium-sized business (201–500 employees)
- Large enterprise (501–5,000 employees)
- Multinational corporation (5,001+ employees)
- Non-profit organization
- Government agency
- Prefer not to disclose

8. What is your job title? _____

Appendix C

Instructions

1. Overview: Please take the time to carefully review the performance feedback materials for two employees, *John* and *Elizabeth*, who are both being considered for promotion to the Level III Procurement Administrator position, a leadership role, at a big manufacturing company. These feedback materials were completed by their coworkers.

1.2 Materials Provided: Below, you'll find the job description for this management position, along with Employee Information Forms containing background information on John's and Elizabeth's work history with the company. You'll also be presented with feedback rating forms completed by their coworkers for both candidates.

*** Next Steps: After reviewing this information, you'll be asked to answer a series of questions about these two employees' performance on the next page.**

2. Job Description

2.1 Position Overview:
Job Title: Level III Procurement Administrator
Location: Athens
Position Type: Full-time

2.2 About Us: At Glomax, we're a forward-thinking organization committed to innovation and excellence in manufacturing. Our diverse team is driven by a shared passion for positive change and exceptional results. We foster an inclusive culture where every individual is valued and empowered to contribute their unique talents. We celebrate diversity and ensure equal opportunities for all employees.

2.3 Job Summary: As a Level III Procurement Administrator, you'll play a crucial role in our procurement team, managing the acquisition of goods and services essential to our operations. You'll oversee the procurement process from supplier sourcing to contract negotiation while ensuring compliance with company policies and regulations.

2.4 Key Responsibilities:
• Lead project teams to deliver on schedule and within budget.
• Develop detailed project plans, including timelines, milestones, and resource allocation.
• Communicate project status, risks, and issues to stakeholders and executive leadership.
• Foster a collaborative, high-performance team environment.
• Collaborate with internal stakeholders to understand procurement needs.
• Identify and evaluate potential suppliers.

• Negotiate contracts, terms, and pricing.
• Manage supplier relationships, including performance evaluations and issue resolution.
• Prepare reports and analyze procurement data to support decision-making and identify areas for improvement.

2.5 Qualifications:
• Bachelor's degree in business administration, Supply Chain Management, or a related field.
• 7 years of experience in procurement and supply chain management, especially in manufacturing.
• Proven track record of managing multiple projects successfully.
• Strong negotiation skills and the ability to build and maintain supplier relationships.
• Excellent communication and interpersonal skills.
• Proficiency in procurement software and data analysis.
• Detail-oriented with strong organizational, analytical, and problem-solving skills.
• Knowledge of procurement regulations and best practices.

*** Next Steps: Keeping the above job description in mind, please carefully review the profiles of the two candidates considered for this leadership position below.**

3. Employee Information Forms

3.1 Candidate Profile 1: John Smith

3.1.1 John Smith's Information Form:
• Work Department: Purchasing
• Job Title: Level II Administrator
• Tenure with the Company: 7 years
• Tenure in Current Position: 5 years
• Work Group: 4-person team
• Age: 29

3.1.2 John Smith's Feedback Form (Completed by Jack Evans, Coworker)
Here's how Jack rated John's skills:
• Planning and Organizing: 80th percentile
• Strategic Vision: >99th percentile
• Communication Skills: 80th percentile
• Decision-Making: >90th percentile
• Problem-Solving: 80th percentile
• Integrity: 99th percentile
• Following Through: 80th percentile
• Accepting Responsibility: 90th percentile

Additional Comments: *"John is a great fit for this role. We attended university together, and he has always been a fantastic collaborator. Smart, driven, and knowledgeable—I highly recommend him."*

3.2 Candidate Profile 2: Elizabeth Williams

3.2.1 Elizabeth Williams's Information Form:

• Work Department: Purchasing
• Job Title: Level II Administrator
• Tenure with the Company: 7 years
• Tenure in Current Position: 5 years
• Work Group: 4-person team
• Age: 29

3.2.2 Elizabeth Williams's Feedback Form (Completed by Emily Johnson, Coworker)

Here's how Emily rated Elizabeth's skills:
• Planning and Organizing: 90th percentile
• Strategic Vision: 99th percentile
• Communication Skills: 90th percentile
• Decision-Making: 99th percentile
• Problem-Solving: 80th percentile
• Integrity: >99th percentile
• Following Through: 90th percentile
• Accepting Responsibility: 90th percentile

Additional Comments: *"Elizabeth was a wonderful coworker three years ago—extremely conscientious, kind, and easy to work with. She would be a great cultural fit here. I highly recommend her for this position."*

* **I've read and understood the information above.**

- Let's proceed.

Appendix D

Scenario-Based Experiment

1st Condition

*** After reviewing the profiles of John Smith and Elizabeth Williams, the manager has determined that John Smith (Candidate Profile 1) is the most suitable candidate and has decided to recommend him for this leadership position. Senior leadership will make the final decision based on this recommendation.**

Instructions: Please evaluate this recommendation by reading the following statements and selecting a number from 1 (= *Strongly disagree*) to 5 (= *Strongly agree*) to indicate how much you agree with each one. There are no right or wrong answers, and all responses will be kept strictly confidential.

I find the manager's recommendation to be:

- Fair.
- Objective.
- Impartial.
- Based on scientific knowledge and methods.
- Supported by good scientific evidence.
- Scientifically provable.
- Based on facts.
- Reasonable and logical.
- Based on an objective consideration of all facts.
- Rational and objective.
- Based on logical analysis.
- Based on facts, not opinions.
- Sincere.
- Made by someone reliable.
- Made by someone trustworthy.
- Made by someone considerate.

*** Next Steps: Now that you've completed this part of the study, you'll be asked to share your views on using different systems for identifying suitable candidates for leadership roles on the next page, the final component of this study.**

2nd Condition

*** After reviewing the profiles of John Smith and Elizabeth Williams, *an AI agent that predicts leadership potential has determined that John Smith (Candidate Profile 1) is the most suitable candidate* and has decided to recommend him for this leadership position. Senior leadership will make the final decision based on this recommendation.**

Instructions: Please evaluate this recommendation by reading the following statements and selecting a number from 1 (= *Strongly disagree*) to 5 (= *Strongly agree*) to indicate how much you agree with each one. There are no right or wrong answers, and all responses will be kept strictly confidential.

I find the AI agent's recommendation to be:

- Fair.
- Objective.
- Impartial.
- Based on scientific knowledge and methods.
- Supported by good scientific evidence.
- Scientifically provable.
- Based on facts.
- Reasonable and logical.
- Based on an objective consideration of all facts.
- Rational and objective.
- Based on logical analysis.
- Based on facts, not opinions.
- Sincere.
- Made by someone reliable.
- Made by someone trustworthy.
- Made by someone considerate.

*** Next Steps: Now that you've completed this part of the study, you'll be asked to share your views on using different systems for identifying suitable candidates for leadership roles on the next page, the final component of this study.**

Appendix E

Post-Scenario Questionnaire

Instructions: Please read each statement carefully and select the number that best reflects your level of agreement. Don't spend too much time on any one statement and remember there are no right or wrong answers. All responses will be kept strictly confidential. Response options: 1= *Strongly disagree*; 2= *Disagree*; 3= *Neither agree nor disagree*; 4= *Agree*; 5= *Strongly agree*.

Survey Questions

1. Promotion decisions informed by AI are more rational, objective, and fairer compared to those based on human decision-making.
    1. Strongly disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly agree

2. I'm likely to endorse the use of AI-based systems in promotion decisions for leadership roles over traditional human-based methods, as they're more scientifically sound, impartial, and trustworthy.
    1. Strongly disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly agree

3. AI-based promotion decisions can produce biased outcomes similar to or more biased than those made by humans.
    1. Strongly disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly agree

4. I believe that AI-based systems used in automating promotion decisions for leadership positions disproportionately impact marginalized groups (e.g., based on gender).
    1. Strongly disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly agree

Curriculum Vitae

Ioannis Barrett
Graduate Student in Organizational Psychology

CONTACT DETAILS

Address: 16 Pirrou St., Athens, 116 33 Greece
Phone: +30 (694) 005-9203
E-mail: I.Barrett@acg.edu

PERSONAL INFORMATION

Date of Birth: Mar. 6, 1996

Citizenship: Greek

EDUCATION

Deree – The American College of Greece – Athens, Greece
Apr. 2023 – Present *MS in Organizational Psychology*

Deree – The American College of Greece – Athens, Greece
Sept. 2014 – Dec. 2021 *BA in Psychology, Cum Laude*
BA in Psychology, Second Class Honours – The Open University, UK
(Dual Degree Awarded)

- Presented my undergraduate thesis at the 17th European Congress of Psychology and participated in related conferences and activities

General Lyceum (GEL) of Limni – Evia, Greece
Sept. 2011 – Jun. 2014 *High School Diploma*

SKILLS

- Competent in academic writing and research
- Highly organized with strong time management abilities
- Team-oriented and capable of working independently
- Proficient in MS Office Suite, IBM SPSS, and social media platforms

ACTIVITIES & CERTIFICATIONS

Sept. 2023 – Feb. 2024 **NGO Boroume ("We Can")**, *Community-Based Work Volunteer*

- Volunteered in initiatives to reduce food waste and malnutrition, enhancing social responsibility and community engagement

July 2023 **University of South Florida, Corporate Training & Professional Education**, *USF - ACG Negotiation and Conflict Resolution Certificate*

- Completed a comprehensive workshop on negotiation strategies and conflict resolution techniques for organizational settings

Oct. 2021 – Dec. 2021 **"Safe Touches" Project, Eliza Charity Organization**, *Data Collection Volunteer*

- Assisted in administering psychological assessments to children, focusing on child abuse prevention and awareness