Employee Preferences for Human-driven versus Artificial Intelligence-driven Performance Evaluation Systems

DO NOT SHARE WITHOUT AUTHOR PERMISSON

Jasmijn Bol*
Tulane University

Conor Brown
Grand Valley State University

Lisa LaViers**
Tulane University

November 2021

*The authors would like to thank the Institute of Management Accountants for funding this research. We would also like to thank Maximilian Margolin, Steven Salterio, Daniel Schaupp, Suyun Wu, Michael Yip, and participants of the 2021 AAA Management Accounting Section conference and workshop participants at Tulane University, Monash University, and Middlesex University for their helpful comments on this paper.

**Corresponding Author: Please contact at llaviers@tulane.edu

ABSTRACT:
New developments in artificial intelligence (AI) have made AI-driven performance evaluation systems a realistic alternative to human-driven systems for an increasing number of jobs. However, prior research suggests employees are concerned about the use of AI for performance evaluation because they perceive that AI decontextualizes performance data. Decontextualization means that the system does not sufficiently consider the performance context and consequently conducts less fair evaluations. Contrary to prior literature, we predict that employees' concerns about decontextualization and subsequently their preferences for human-driven and AI-driven performance evaluation systems differ based on firm characteristics and individual differences. Using two experiments, we examine several factors that influence their preferences. We find that employees have relatively stronger preferences for AI when the operating environment of the organization is stable, when they feel that they have been subjected to discrimination by managers in the past, or when they have lower social intelligence. We also hypothesize and find that these effects are moderated by the amount of social contact between employees and managers by manipulating whether employees and managers have either a shared or remote workspace. These findings help guide firms that are considering investment in AI-driven performance evaluation systems.

KEYWORDS: Performance Evaluations, Artificial Intelligence, Objective Performance Measurement, Subjective Performance Evaluation, Discrimination, Social Intelligence, Remote work.

I. INTRODUCTION

Although objective performance evaluation systems are praised for their strong motivational impact and simplicity, traditionally, very few jobs have had objective measures which could completely capture every dimension of performance (Prendergast 1999). Because of the limited amount of suitable objective performance data, most jobs' performance evaluation systems also include elements of subjective evaluation based on human judgements (Baker et al. 1994). The addition of subjectivity allows managers to consider the elements of performance which are not easily measured and to incorporate the impact of contextual factors (Prendergast 1999; Bol 2008). While including human judgement can make systems more holistic, subjectivity is a double-edged sword that also introduces human biases (Prendergast 1999; Moers 2005). Thanks to recent technological advances in artificial intelligence (AI), the practical limitations on objective performance measurement are changing. New AI-driven technology now allows for a far greater number of performance attributes and contextual factors to be evaluated objectively, because it enables different types of data collection and more sophisticated analysis. These large advancements mean that for the first time in history, a significant number of firms can use objective performance measurement through AI to capture a more complete picture of employee performance, but it is yet unknown in which situations it would be more beneficial than using more subjective human-driven systems.

In order to add to the collective understanding of when AI-driven performance evaluation may improve firm outcomes, we examine employees' relative preferences for AI-driven versus human-driven systems in different environments and for different types of employees. Understanding employee preferences and fairness concerns regarding performance evaluation systems and how these preferences vary is critically important because the motivational effectiveness of a performance evaluation system is driven by employees' acceptance and perceived fairness of that system (Miceli et al. 1991; Taylor et al. 1995; Janssen 2001; Jawahar 2007). When employees feel that a performance evaluation system does not fairly capture and reward their effort, they will not be willing to increase, and might even decrease the effort they are providing to the organization (Bol 2011).

The technological advancements of the last five years have allowed AI-driven systems to emerge as a realistic alternative to using human-driven systems with subjectivity (Mackenzie, Wehner, and Kennedy 2020). These new systems' advanced data processing capabilities not only have the ability to objectively capture a wider range of performance dimensions, but also to control for a large variety of contextual factors that influence employee performance. For example, natural language processing allows AI to interpret the meaning, sentiment, and tone of textual information, such as customer comments in customer satisfaction surveys, allowing these surveys to be objectively evaluated in the same way for every employee (Ghiassi, Skinner, and Zimbra 2013; Hirschberg and Manning 2015; Zhang, Wang, and Liu 2018). AI can also determine which of a large set of complex economic factors like the price of oil, a competitors' changing percentage of market share, or inflation should be considered when adjusting for uncontrollable contextual factors (Chung et al. 2019). Prior to AI, this type of information could only be interpreted by humans; purely objective performance evaluation was largely limited to simple counts of output and basic computations in spreadsheets (see Lazear (2000) for an example of this type of system).

These improvements in AI are happening at the same time that the COVID-19 pandemic forced many employees to work remotely (Lund et. al 2021). This switch to remote work is important as it fundamentally changed how many employees are being evaluated by their managers (Knight 2020). More employees than before are being evaluated using predominantly digital data, as managers' ability to collect observational data through social interactions with employees is often limited to online video or chat conversations. This forced switch to more digital interactions increased organizations' interest in AI as they are already investing in more digitalization. For example, the CEO of Enaible, a start-up that sells an AI-based employee productivity tracking software, says that the firm has quadrupled sales of the firm's turn-key AI evaluation system since the start of the pandemic (Heaven 2020).[1] Thus, our investigation into employee preferences for AI-driven systems and how these preferences might be different for remote workers is very timely.

While AI-driven systems improve, organizations are also trying to improve human-driven systems through new innovations (i.e., multi-rater systems, calibration). However, the 2019 WorldatWork survey shows that dissatisfaction with human systems, driven by cost and ineffectiveness, are still high, even for those organizations that invested in new innovations (WorldatWork, 2019). A management consulting company estimated that a company of ten thousand employees spends $35 million each year conducting employee reviews and that a single manager spends an average of two hundred hours each year conducting them (Cunningham 2015). Despite the level of investment firms make into human-driven systems, managers report high levels of dissatisfaction with the process (Cappelli and Tavis 2016). Performance evaluation is not only time consuming; managers also dislike the task as subjective assessments of employees often leads to costly confrontations (Bol 2011). It is therefore not surprising that organizations are interested in learning about the cost savings and improved outcomes that AI-driven systems could potentially bring to the firm.

Although employees are generally also dissatisfied with the performance evaluation system, it is not clear that they would be as enthusiastic about AI-driven performance evaluation

[1] This type of turn-key system is available for sale today and collects data from common business software like Word and Slack and allows a wide range of firms to use AI-driven evaluation systems.

systems as top management might be (Kellogg, Valentine, and Christin 2020). Prior research on algorithm aversion finds that people are resistant to following the recommendations and judgments that AI makes *for* them as a decision aid (Dietvorst, Simmons, and Massey 2015; Burton, Stein, and Jenson 2019). Research examining how employees would feel about AI making judgments *about* them is scarcer, but also shows employee reluctance towards AI. In one of the first papers about AI-driven performance evaluation, Newman et al. (2020) argue that employees prefer human-driven systems to AI-driven systems because employees believe AI-driven systems decontextualize performance information more than humans. Decontextualization means the evaluator (i.e., human or AI) considers only the performance measures, not the context in which the employee operates. Interestingly, Newman et al. (2020) find that these perceptions hold even when the two systems make the exact same performance assessment.

In this study, we refine these findings to demonstrate that employee preferences against AI-driven performance evaluation systems are not constant, but instead depend on characteristics of the organization's operating environment, individual differences between employees, and opportunity for social contact (i.e, a shared or remote workspace). First, we predict that in stable operating environments employees will be less concerned about decontextualization than in unstable operating environments. In stable operating environments, the context in which employees perform their jobs is standard and the contextual factors that influence performance are known. However, when the operating environment is unstable, employees will believe that a fair evaluation will require more contextualization, i.e. consideration for both known factors and factors that were not previously known. We predict that employees will perceive human-driven systems' subjective adjustments to be better at incorporating the context of an unstable operating environment than AI-driven systems even when those systems are programed to consider context.

Consequently, employees will have a relatively higher preference for AI-driven systems versus human-driven systems in stable environments than in unstable environments.

While we expect the organizational operating environment to affect all types of employees similarly, we posit that employees' individual differences and the opportunities for social contact within the employees' workspace will result in differing preferences for human-driven and AI-driven performance evaluation systems. We predict that employees who believe they have experienced workplace discrimination will have a relatively higher preference for AI-driven systems than employees who do not believe they have experienced workplace discrimination. We argue that this preference is driven by the employee's perception that human managers over-contextualize, meaning that they include performance-irrelevant contextual factors like gender, race and sexual orientation in their subjective performance assessments. Since employees who have experienced workplace discrimination have experienced this downside of managers' subjectivity firsthand, we posit that they will prefer a system that they perceive to be more objective.

Based on the intergroup contact theory of prejudice, we also predict that the relative preference for AI-driven performance evaluation systems of employees who have experienced workplace discrimination will be larger depending on the opportunities for social contact in their workplace. Research shows that when social contact between group members is relatively scarce intergroup prejudice and biases are more prevalent (Pettigrew et al. 2011, Pettigrew 1998). As a result, we predict that for those who have experienced workplace discrimination there is even more concern for managers' subjectivity in performance evaluation when there is limited opportunity for social contact, like when they work remotely, compared to when there is ample opportunity for social contact, like when they share an office space with their manager and peers.

Although we predict some employees are wary of over-contextualization by managers, we posit that others will feel that this type of evaluation is to their advantage. Employees with high social intelligence will likely believe that their social skills will cause them to benefit from their managers' subjective assessments resulting in a relatively higher preference for human managers. Again, we predict that this effect will not be uniform, but dependent on the opportunities for social contact. Specifically, when there are many opportunities to interact with the manager, social intelligence will have a larger effect on employees' relative preference for an AI-driven evaluation system than when social contact is limited.

We use two scenario-based experiments in an online labor market. Each experiment has a manipulation and measured variables. In both of our experiments, participants assume the role of employees at a hypothetical firm. After learning about the firm and nature of the performance evaluation, participants are asked to indicate their preference for a human-driven or AI-driven performance evaluation system. In experiment 1, we vary the stability of the business environment and measure perceived past workplace discrimination. In experiment 2, we manipulate the opportunity for social contact by varying the employees' workspace; employees either work remotely and have lower opportunities for social contact or in a shared office with their managers and have higher opportunities for social contact. We also measure employees' social intelligence scores using the Tromsø Social Intelligence scale (Silvera, Martinussen, and Dahl 2001) and perceptions of past workplace discrimination.

We find that employees' preferences for human-driven versus AI-driven systems are indeed not uniform. Employees show a significantly higher relative preference for AI-driven systems in a stable environment than in an unstable one. Through mediation analysis, we show that the effect of operating environments on preferences is driven by a difference in employees' views of how the systems consider the operating environments. That is, we find that employees are more concerned about an AI-driven system's ability to fairly consider context in an unstable environment than in a stable environment. In contrast, their concerns about a human manager's ability to fairly consider the context of the environment are unchanged between stable and unstable environments.

Using data from both experiments 1 and 2, we show that employees who feel they have faced workplace discrimination in the past have a relatively stronger preference for AI-driven systems than employees that have not faced workplace discrimination. We also predict and find that the preference for AI-driven performance evaluation systems among participants who believe they have experienced workplace discrimination is stronger when participants have less social contact in the workspace. Lastly, using data from experiment 2, we find that participants with higher social intelligence have a stronger relative preference for human-driven systems than participants with lower social intelligence and that this effect differs based on the opportunities for social contact. When working remotely, social intelligence plays no significant role in preference for AI-driven versus human-driven systems. When working in a shared office, however, those with higher social intelligence prefer human-driven systems. We extend on these results in supplemental analysis. Using a moderated-mediation model, we show that the effect of social intelligence on preference for an AI-driven or human-driven evaluation system is driven by a difference in perceived fairness of the evaluator. Together these results show that the effect of social intelligence on preferred type of evaluator is mediated by the perceived fairness of the evaluator, and this indirect effect is moderated by employees' opportunity for social contact.

This paper contributes to the accounting literature and to accounting practice in several ways. First, its findings contribute to the theoretical knowledgebase about subjective and objective performance evaluations. Prior to the advent of AI, for most jobs objective performance measurement could not reasonably capture employee performance and, consequently, subjective judgment by the manager needed to be included to avoid perverse incentives. With advances in AI, objective measurement can be more complete for many more jobs, and consequently, many organizations are rethinking the costs and benefits of more subjective human-driven systems versus more objective AI-driven evaluations. We contribute to the literature by highlighting that employee preferences need to be considered in this reassessment because those preferences directly impact motivational effectiveness.

Incorporating employee preferences is not a straightforward task. We show that employees' preferences for human-driven versus AI-driven performance evaluation systems are not uniform across all types of employees and in all performance contexts. That is, we show that how employees perceive the AI-driven performance evaluation system is not just driven by a blanket effect of algorithm aversion, but also by their beliefs about what is the fairest way to be evaluated. In some cases, employees want a more subjective human-driven system that they perceive will contextualize their performance, while in other cases, employees prefer an objective AI-driven system that they perceive considers only their performance and ignores the other factors. Our findings also connect with other research on AI in accounting which shows that auditors have different trust levels in using AI when the inputs are more subjective or more objective (Commerford et al. 2021). Together our work indicates that the qualities of the data that the AI uses may have as much of an impact on trust in, and preference for AI as the actual design of the AI itself.

Our study also contributes to the growing literature on differences in how to manage employees who work in a shared or remote workspace. Our findings suggest that the relative preference between the AI-driven versus human-driven systems depends on the opportunity for social contact between employees and their managers. This insight is important because the percentage of professionals who have left the office to work remotely has increased significantly over the last decade, and due to the COVID-19 pandemic, this number increased even more rapidly in March 2020. Although many of these employees will return to the office, a significant percentage will continue to work remotely and consequently have less opportunity for social interaction. In this new more remote world, firms might feel AI-driven systems are the natural next step because employee performance needs to be captured digitally regardless of which person or system will evaluate it. While some employees will likely welcome this approach, like those who have been discriminated against in the past, others will likely be warier, for example employees with high social intelligence. This highlights again that managers should not look at just the cost-benefit trade-off of *implementing* AI-driven systems but also consider the preferences of their workforce as this influences the motivational effect of the performance evaluation system.

This research also contributes to the work of computer scientists who are studying AI development and algorithm aversion (Burton, Stein, and Jensen 2020). While these researchers are working to advance the technology, they lack a deep understanding of the corporate performance evaluation processes. By showing how AI functions in business environments, we better illustrate how end users will react to this technology and how it can be best implemented. We hope that by working together, we can gain a better understanding of where advancement of this technology might be most beneficial to organizations.

II. LITERATURE REVIEW

Organizations use performance evaluation systems to motivate their employees. Research has shown that rewards linked to performance assessments can result in increased productivity (Engellandt and Riphahn 2011). Prior studies have, however, also shown that in order for performance evaluation to have a motivational effect, employees need to believe that increased effort will result in higher performance ratings (Downes and Choi 2014; Trevor, Reilly, and Gerhart 2012). Employees must believe that their current period effort will be fairly rewarded during a future evaluation for them to be willing to put forward high effort levels (Bol 2011). If employees do not feel that effort will be fairly rewarded, they will reduce their effort levels or leave the firm (Simons and Roberson 2003). Because of this need for confidence in and acceptance of the performance evaluation system, employees' preferences and fairness perceptions are vital. Organizations can implement elaborate systems to capture performance, but if the employees do not accept the system as fair, the system will not motivate higher effort (Trevor et al. 2012).

While a plethora of studies have been conducted in management accounting and other fields on performance evaluation (Prendergast and Topel 1993; Jacob and Lefgren 2008; Demeré, Sedatole, and Woods 2019), almost all of these studies have focused on human-driven evaluation because, until recently, it was the only system possible. As a result, the question of whether employees prefer subjective or objective evaluation has not been the focus of the literature because the use of supervisor discretion (i.e., subjectivity) was not a choice but a necessity for most organizations as reasonably complete objective performance evaluation data was not available. As AI-driven evaluation systems emerge as a viable alternative, supervisor discretion becomes a choice, and consequently firms need to re-think the costs and benefits of different approaches. That is, research that examines employee preferences and perceptions of AI-driven versus human-driven systems is necessary.

In this study we start by discussing the advantages and disadvantages of each system and then develop hypotheses regarding employees' preferences. Note, though it is possible for AI and humans to work together on decision-making, in this study we specifically examine the AI-driven performance evaluation systems where AI and humans act independently from one another, rather than a system in which a human manager uses an AI recommendation as a decision aid.

Human-Driven Performance Evaluation

Traditionally, human managers have evaluated employee performance (Prendergast 1999; Bol 2008). In general, managers need to apply subjective judgement to complete performance evaluations for most job functions because performance cannot be captured by a predetermined formula that combines purely objective, quantitative performance measures. For most jobs, the manager is asked to at least determine the weight placed on the different performance measures and/or adjust for uncontrollable or unanticipated factors when deemed appropriated (Prendergast and Topel 1993; Ittner, Larker and Meyer 2003). The process of making weighting decisions or adjustment in order to take circumstances, like firm competition and economic factors into account is referred to as contextualization (Newman et al. 2020). Often, however, the manager is also asked, in addition to measuring objective performance dimensions, to make assessments on dimensions that cannot be measured easily, like leadership skills and the quality of project execution (Ittner, Larker and Meyer 2003). Allowing managers to have the discretion to apply their judgment can lead to more complete assessments of the employees' performance, skills, and long-term contributions, leading to a performance evaluation that employees perceive to be fairer (Gibbs et al. 2005; Fisher et al. 2005; Bol 2011).

While managers' subjectivity can be advantageous because of its ability to contextualize and make assessments more complete, it also has a dark side. Managers can add the *wrong* context to the process. According to an international survey conducted by Glassdoor (2019), around 50% of respondents report having personally seen or experienced ageism, sexism, racism, or homophobia in their workplace. These experiences occur because human managers are subconsciously or consciously including inappropriate information about employees in the evaluation process. Because of these experiences, employees are concerned about the fairness of performance assessments that include judgement by the manager (Chan and Dimauro 2020; Moise and Cruise 2020).

Beyond identity-based discrimination, research has also shown that managers have many other biases like ones that result from personal relationships and office politics (Prendergast and Topel 1996; Higgins, Judge, and Ferris 2003). Managers often display favoritism to certain employees that results in overly favorable ratings for employees who have a personal relationship with the manager which employees can perceive to be unfair (Ittner, Larcker, and Meyer 2003; Bandiera, Barankay, and Rasul 2009). Thus, employees often perceive that their managers are influenced by factors other than performance and that this leads to biased performance ratings, a process we refer to as over-contextualization. We posit that perceived over-contextualization negatively affects the perceived fairness of the performance evaluation system.

Beyond the problem of over-contextualization, the other major disadvantage to human-driven evaluation is costliness. The evaluation process is typically slow and labor intensive (Rogel 2020). Estimates of the average time spent on performance evaluations are as high as hundreds of hours per year (Cappelli and Tavis 2016). Before dropping annual performance reviews, Deloitte Inc. estimated that the company spent nearly two million hours each year on performance evaluations (Cappelli and Tavis 2016). Many organizations limit employee performance evaluation to one evaluation per year because firms simply cannot afford to do more, even though many employees desire increased feedback (Tzuo 2017). Firms that are looking for ways to cut costs are therefore attracted to the idea that an investment in an AI-driven system could free up managers' time and lead to increased firm productivity (Aspan 2020).

AI-driven Performance Evaluation

Technical progress in the field of AI has made it possible for companies to switch to an AI-driven performance evaluation system. Some firms are developing their own AI-driven systems for this purpose. For example, IBM uses its self-developed AI-driven system (called Watson) to predict future employee performance (Greene 2018). IBM claims this has resulted in large cost savings with a reduction of human resources staff by 30% (Rosenbaum 2019). Other firms like Enaible and ButterFly.Ai are selling AI-driven performance evaluation systems. While these AI-driven systems may make up only a small portion of all performance evaluation currently, industry experts predict that the trend will grow (Holsinger et. al 2019).

Indeed, AI-driven systems are attractive to managers because of the perceptions that AI provides a simple, cost-effective, and efficient way to improve evaluations (Fecheyr-Lippens, Schaninger, & Tanner 2015; Cheng & Hackett 2021). AI allows firms to capture the benefits of fast and repeatable objective evaluation for more jobs. Although some of the objective information collected might have been available before, the processing speed to both collect and analyze the information was not yet there. Moreover, AI systems can quickly analyze large amounts of data and develop weights that consistently adjust across every employee for known performance influencing factors like economic conditions and competitors' actions. Because of its processing speed, it can do complex math faster than humans can and, unlike human managers, it does not suffer from cognitive overload (Bol, Margolin and Schaupp 2021).

While the top management of a firm may be interested in AI because of its potential to cut costs and streamline the evaluation process, employees' reaction to AI-driven performance evaluation is less clear. There is some research on algorithm appreciation that suggests that people may prefer the recommendation of an algorithm to one from a human (Castelo, Bos, and Lehmann 2019; Berger et al. 2021). However, there is a larger body of research on algorithm aversion that finds that most people prefer to rely on guidance from a human advisor more than an algorithm (Dietvorst, Simmons, and Massey 2015; Prahl and Van Swol 2017; Burton, Stein, and Jenson 2020). Both of these literature streams mainly look at the acceptance of a recommendation by an algorithm. The research on how employees perceive their performance being evaluated completely by an AI is more limited. One of the small number of studies in this area, Newman et. al (2020), shows that employees perceive performance evaluations made by AI to be less fair than those made by human evaluators. They argue that this is caused in part by employees' belief that AI does not take the context in which employees operate into account. Consistent with their predictions, the authors find that employees prefer human-driven systems, even when the outcomes are identical.

We contribute to this line of research by showing that employees' perceptions on the extent to which AI-driven performance evaluation system decontextualizes and how problematic this is for performance evaluation are not uniform, and consequently their relative preference for AI-driven versus human-driven performance evaluation systems are not constant. In this paper we have identified several different areas of interest to management accounts that we predict influence these relative preferences: The stability of the organizational operating environment, individual differences in employees and the opportunities for social contact in their workspace.

Note, in our research setting the human manager and AI are using the same set of relevant performance measures to capture performance and have access to the same contextual information. Moreover, we assume that the AI is well functioning and that the manager will take their job seriously. The hypothesized difference in preferences are therefore predicted to come from employees' perceptions of how well the information is processed and analyzed, not the availability of different types of information.

III. HYPOTHESIS DEVELOPMENT

Stability of the Organizational Operating Environment

When an organization is operating in a stable environment, the context in which employees' performance occurs is relatively unchanging. With many periods of comparable information, an organization can develop reasonable benchmarks and prediction models of not only employees' performance but also of other factors that are known to influence performance like economic factors and competitor actions. We predict that employees will perceive that the comparable historical data makes contextualization of their performance easier and subjective adjustments for context less necessary. Under these circumstances, employees are less concerned about potential decontextualization of AI-driven systems. Moreover, Hu (2021) finds that in stable environments, employees build trust in AI-driven systems because of the consistency of their judgements. In contrast, employees may doubt that managers will match this level of consistency in judgement in stable environments due to their cognitive limitations and tendency to bias (Prendergast and Topel 1993; Ittner, Larker, and Meyer 2003).

In an unstable operating environment, historical data may not be perceived by employees to be representative of the current context in which employees' performance occurs. Employees will likely feel that there are new factors that impact their performance which need to be contextualized for their assessment to be an accurate reflection of their effort. For example, in the context of the COVID-19 pandemic, employees may prefer their performance to be assessed in the context of stay-at-home orders, not just the economic and competitive factors that were considered before. As a result, AI's perceived decontextualization may make AI-driven performance evaluation less attractive to employees in unstable environments than in stable environments.

This prediction, however, is not without tension. It is not clear whether a human or an AI would *actually* perform better at performance evaluation in unstable environments. While AI uses models explicitly built on historical data, human managers also have mental models of what "high performers" are like based on the same data (Bol and Leiby 2018). Thus, it is also not easy for human managers to fairly consider new contextual factors that influence performance. Moreover, humans may be cognitively overloaded by needing to make complex assessments using unfamiliar data sources or larger adjustments from their mental models (Simon 1990). When humans are cognitively overloaded, they are more likely to engage in biases and inaccuracies. AI will not be cognitively overloaded and may in fact be better at making the types of large adjustments needed to examine the data in the unstable period. Thus, it is not clear which system would actually assess performance more fairly.

Despite the fact that both human-driven and AI-driven evaluators will have a harder time evaluating performance in unstable operating environments, we still predict that a lack of stability will increase the relative preference for human-driven versus AI-driven evaluation systems. We predict that this change in preferences is driven by employees' increased concerns about AI's decontextualization. In contrast, when the operating environment is stable, we predict that employees will have a relatively higher preference for AI-driven systems because the historical data will allow the AI to perform in a very consistent and fair fashion. Thus, we hypothesize that the stability of the operating environment influences employees' preferences for human-driven versus AI-driven performance evaluation systems.

H1: Employees in stable business environments show relatively higher preferences for AI-driven systems versus human-driven systems than employees in unstable business environments.

Employee Discrimination and Opportunities for Social Contact

Employees who work in the same type of job in the same type of operating environment will not necessarily have similar preferences for human-driven versus AI-driven evaluation systems. We predict that individual employees are also influenced by their past experiences with performance evaluation systems (or other organizational controls). Specifically, we predict that if employees feel that they have experienced workplace discrimination in the past, then they will have a greater fear that a human manager will over-contextualize and generate a biased evaluation. These employees will also have fewer concerns about AI-driven systems' decontextualization as they welcome being evaluated on just their performance and not have contextual factors like their gender or race subjectively considered. Thus, contrary to Newman et al. (2020), we argue that employees who believe they have suffered from discrimination in the past will have an appreciation for the decontextualization of the AI-driven system which will, all else equal, increase their preference for AI-driven systems relative to human-driven systems. This prediction is also consistent with the finding that members of underrepresented groups are more likely to select into firms which are already using AI-driven systems (Brown, Burke, and Sauciuc 2021). As a result, we hypothesize that employees who feel they have been discriminated against in the past by a human manager have a relatively greater preference for AI-driven performance evaluation versus human-driven performance evaluation compared to employees who do not believe they have been discriminated against.

H2a: Employees who believe they have been discriminated against in the past will show relatively higher preferences for AI-driven systems versus human-driven systems than employees who do not.

We posit that the relative preference that employees who have been discriminated against in the past have for AI-driven systems versus human-driven systems also depends on the opportunities for social contact available to the employees during work. Intergroup contact theory of prejudice posits that bias towards people in a different social group can be reduced by social interaction (Pettigrew 1998). In their meta-analysis of 515 research studies, Pettigrew et al. (2011), find that intergroup social contact is effective at increasing trust and forgiveness and reducing prejudice between groups. Consistent with this theory, we hypothesize that employees that have experienced past discrimination will have a stronger preference for AI-driven systems versus human-driven systems when there are only limited opportunities for social contact during work, like when they work remotely, compared to when there are ample opportunities for social contact during work, like when they work in a shared workspace, because the lack of intergroup social contact makes the over-contextualization of human managers even more pronounced.

H2b: The effect of past discrimination on preferred evaluator type will be larger when employees have limited versus ample opportunities for social contact with their manager in the workspace.

See Figure 1, Panel A for a graphical depiction of H2a and H2b.

Social Intelligence and Opportunities for Social Contact

While employees who feel they have been discriminated against may be wary of contextualization by humans, other employees may feel that contextualization is to their advantage in the performance evaluation process. Some employees are particularly good at navigating social situations because they have high social intelligence. Research has shown that positive relationships and being an effective "political player" results in positively biased performance ratings (Prendergast and Topel 1996; Ittner, Larcker, and Meyer 2003; Bandiera et al. 2009). As a result, these employees will likely want their manager to contextualize and find decontextualization of AI-driven systems to be undesirable. On the other end of the spectrum, employees with low social intelligence may be more likely to embrace AI as they find socialization tiresome and wish to avoid it. These employees will not want to be compared to other employees by the human managers with whom that manager may have stronger social relationships. They would prefer an AI that will not use friendships or political connections in its evaluation process. We predict a main effect of social intelligence on preference for human-driven versus AI-driven systems.

H3a: Employees who have lower social intelligence will show relatively higher preferences for AI-driven systems versus human-driven systems than employees who have higher social intelligence.

Besides a main effect, we also predict that this effect will be stronger in situations where there will be more opportunities for social contact. When managers interact with their employees and they are able to informally monitor them and not just examine their work output, there will be more context that the manager can consider when evaluating performance. There will also be more opportunities for employees with social intelligence to influence these judgments, and hence, more opportunities for over-contextualization. We therefore predict that the level of social intelligence will have a greater effect on preferences for AI-driven versus human-driven systems when there are more opportunities for social contact.

H3b: The effect of social intelligence on preferred evaluator type will be larger when employees have ample versus limited opportunities for social contact with their manager in the workspace.

See Figure 2, Panel A for a graphical depiction of H3a and H3b. In the next section, we detail the methods by which these hypotheses were tested and describe the experiments conducted.

IV. METHODS

We test our hypotheses using two experiments. Both were conducted on Amazon's Mechanical Turk (AMT) platform and were approved by the IRB of the major US-based research institutions involved in the study. Prior research has shown that this platform supplies participants who behave similarly to traditional student-driven samples, while also being more demographically representative of the American labor force (Paolacci, Chandler, Ipeirotis 2010; Farrell, Grenier, and Leiby 2017; Buchheit et al. 2018). A representative sample of the broader labor force is important for our research question, in which we explore the preferences and perceptions of rank-and-file employees at firms. In addition, AMT participants typically have experience working both online and in person meaning that they have received digital performance reviews for their AMT work done in a remote setting with no social interaction and traditional human-driven evaluations from bosses in shared office space with social interactions. This set of dual work experiences allows them to be uniquely suited to imagining a world where both types of reviews are possible and to be able to express a preference between them. Participants earned a base pay of $2.00 and then an additional $0.10 for each of two attention check questions and one manipulation check question they answered correctly. Participants were told the task would take up to 20 minutes of their time. All participants were adults who were at least 18 years old.

Several steps were taken to increase the quality of the data collected from AMT. The experiments were conducted using Cloud Research, an independent research platform which allows for more rigorous subject filtering. The participant pool was limited to workers who had completed at least 1,000 tasks with a 90% approval rating and whose IP addresses were located in the United States. Participants were also excluded if they had duplicate IP addresses or had IP addresses identified by Cloud Research as “suspicious.” Participants first completed a consent form and then a captcha, a picture of text that participants needed to translate into machine readable data. Participants who could not successfully complete the captcha were not allowed to continue. The use of a captcha is in keeping with best practices surrounding AMT usage and helps filter for automated, non-human, participants.

Both experiments use a similar design and company setting. They are 1x2 experiments with additional measured variables. In each, participants are asked to assume the role of a salesperson at ABC robotics. They learn that they spend half their time selling robots and the other half on site with clients conducting trainings. They are evaluated on both tasks using objective, and subjective performance metrics to determine whether or not they will receive a monthly performance-based bonus. See Figure 3 for the detailed information participants saw regarding their evaluation. They are told that this bonus is meaningful to them, and that in the months that they receive it, they use it to buy themselves a special treat.

After their job has been explained to them, they are told that ABC robots is forming a new division to which they will be reassigned. At the moment, ABC robots is using human-driven evaluation for some departments and using AI-driven evaluation for others and participants are told that these systems have been equally successful. The company needs to determine which system to use in the new division and are therefore conducting a vote to get employees' opinions. Participants are asked to vote on which they would prefer. Their votes are measured on a 1-5 scale where 1 is "strongly prefer humans," 3 is no preference, and 5 is "strongly prefer AI." After participants vote, they complete a post-experiment questionnaire (PEQ). The PEQ contains demographic data and the measured variables used in analysis and tests of hypotheses.

While the base of the experiments is the same between the two experiments, as discussed above, they each have a different manipulation. The PEQ also differs slightly between the two experiments. The next two sections detail these differences and Figure 4 provides a timeline which highlights similarities and differences.

Experiment 1 Design

In experiment 1, we manipulate the *Operating Environment*. The company and the job are as described in the previous section, but the operating environment of the firm is manipulated between subjects at two levels: *Stable* and *Unstable*. In the *Stable* condition, participants learn that the firm's current operating environment is not different from prior years, and the company has extensive historical employee performance data and experience conducting performance evaluations under the current economic conditions. In the *Unstable* condition, participants learned that the COVID-19 pandemic and related financial crisis have significantly disrupted the company's operations and the firm does not have experience evaluating employee performance under the current operating conditions. See Figure 5 for the exact manipulation. All participants are told that they work from home when they are selling the robots and in their client's offices when they are training. As a manipulation check, participants were asked which economic condition they are in. Participants who cannot answer this question correctly are removed from the sample for data analysis.

In a PEQ, participants reported whether they felt they had been subject to workplace discrimination during their career (outside the experiment). Participants were asked to rate their agreement with the statement, "I have been subject to discrimination at work." Responses were recorded on a seven-point scale from "strongly disagree" (coded as 1) to "strongly agree" (coded as 7). These responses constitute the *Past Discrimination* variable in the analysis. Demographic information and other perceptions relating to AI-driven and human-driven performance evaluations were also collected in the PEQ.

Experiment 2 Design

As stated before, experiment 2 uses the same company setting, job, and dependent variable (*Preferred Evaluator*) as experiment 1. All participants are told that they are in a stable operating environment. In experiment 2, we manipulate opportunities for social contact between the employee and the manager in the workspace and name the variable *Workspace*. *Workspace* is manipulated between subjects at two levels: *Shared* and *Remote*. In the *Shared* condition, participants were informed that they work in a shared central office whenever they are not on-site with customers. In the *Remote* condition, participants were told that they work from home whenever they are not on-site with customers. In order to increase the saliency of this manipulation, participants are also shown a photo of a desk on a white background and are told to imagine it is their desk. They are also given the same details about features of their offices. For example, they are told that they have snacks and coffee available to them. See Figure 6 for the exact manipulation. After participants see the *Workspace* manipulation, they are asked to report where they work for their job at ABC robots. Participants are not allowed to advance in the experiment until they correctly respond to this manipulation check.

After the manipulation check, participants indicate their preference for an AI-driven or human-driven evaluation system and then complete a PEQ. The PEQ is similar to the one of experiment 1, but the Tromsø Social Intelligence Scale is added. This scale was developed and validated in Silvera, Martinussen and Dahl (2001). Participants responded to 21 items using a seven-point scale ranging from "strongly disagree" (coded as 1) to "strongly agree" (coded as 7). The questions include statements like "I understand other peoples' feelings" and "I find people unpredictable." See the appendix for the complete scale.

V. RESULTS

A total of 150 and 155 participants were recruited from AMT for experiments 1 and 2, respectively. Participants had a mean age of 38 years (36 years in experiment 2) with a reported mean work experience of 16 years (12 years in experiment 2). For sample descriptive statistics, see Table 1 for experiment 1 and Table 2 for experiment 2. All participants completed manipulation checks. In experiment 1, 124 participants (83%) correctly responded to the manipulation check. These participants are retained for analysis in experiment 1. In experiment 2, all 155 participants passed the manipulation check because participants could not move forward in the task without correctly answering the question and were referred back to the pertinent information for additional review if they submitted an incorrect response.

Tests of H1

In the full sample of experiment 1, the mean value of *Preferred Evaluator* is 2.73 out of 5 where 3 is the midpoint of the scale, indicating that participants slightly prefer human evaluators over AI on average. However, consistent with H1 which predicts that employees will show a stronger relative preference for AI-driven systems when the operating environment is stable than when it is unstable, mean *Preferred Evaluator* in the *Stable* condition is 2.93 and 2.43 in the *Unstable* condition. See Table 3, Panel A for simple means by condition. To more formally test for the effect of *Environment Stability* on *Preferred Evaluator*, we perform an analysis of covariance including *Environment Stability* and *Past Discrimination* (Table 3, Panel B). We find a significant effect of *Environment Stability* on *Preferred Evaluator* ($F_{(1,121)} = 11.38$; $p < 0.01$), consistent with H1.[2]

In order to further investigate the mechanism by which *Environment Stability* affects *Preferred Evaluator*, we perform a mediation analysis using the simultaneous regression method outlined in Hayes (2018). In development of H1, we argue that employees believe that an AI-driven system is less able than a human-driven system to adjust performance for the relevant context when the operating environment is unstable, resulting in performance evaluations that are perceived to be less fair. We therefore test whether the perceived ability of AI and human managers to consider the context of the operating environment mediates the effect of *Environment Stability* on *Preferred Evaluator*. To capture concerns about ability to consider context, we use participants' responses to two post-experimental questionnaire items: (1) "When casting my vote for who would evaluate me (AI or a human manager), I was concerned that the AI wouldn't be able to fairly

[2] In an untabulated result, we repeat the analysis removing *Past Discrimination* from the model and continue to find a significant effect of *Environment Stability* on *Preferred Evaluator* ($p = 0.01$).

consider the circumstances I was in." (2) "When casting my vote for who would evaluate me (AI or a human manager), I was concerned that a human manager wouldn't be able to fairly consider the circumstances I was in." These responses are recorded on a seven-point scale from "strongly disagree" (coded as 1), to "strongly agree" (coded as 7). Responses to the first of these PEQ items are captured in the *AI Context* variable, and responses to the second item are captured in the *Human Context* variable. Responses to both items are included in the mediation analysis, and *Past Discrimination* is included as a covariate.

Full results of the mediation analysis are presented in Table 4 and depicted graphically in Figure 7. We find evidence that *Environment Stability* affected participants' concerns about AI's ability to consider the relevant context (p < 0.01, see Table 4, Panel A), but we do not find evidence of a significant effect of *Environment Stability* on concerns about a human manager's ability to fairly consider the context of employee performance (p = 0.12, see Table 4, Panel B). We also find that each of these measures significantly predicted *Preferred Evaluator* (p < 0.01 for both *AI Context* and *Human Context*, see Table 4, Panel C). We find evidence that the total indirect effect of *Environment Stability* on *Preferred Evaluator* through *AI Context* is significant (95% confidence interval: [0.04, 0.44], see Table 4 Panel D). We do not find evidence of a similar indirect effect through *Human Context* (95% confidence interval: [-0.04, 0.40]). These results indicate that participants preferred an AI-driven evaluation system more in a stable environment than an unstable environment because they were less concerned about AI's ability to contextualize performance in a stable environment. Consistent with the underlying reasoning for our hypothesis, environment stability did not affect participants' concern about a human manager's ability to contextualize performance information. Overall, we find strong support for H1.

Tests of H2a and H2b

H2a predicts that employees who believe they have been discriminated against in the past will show relatively higher preferences for AI-driven systems versus human-driven systems than employees who do not. Recall that we measure perceived past discrimination by capturing participants' agreement with the following statement, "I have been subject to discrimination at work." (see Table 1, Panel A). Participants in experiment 1 rated their agreement as a mean of 3.44 out of 7 for experiencing discrimination in the past. Thirty-two percent of participants reported that they at least "somewhat agree" (a response of five or more) with the statement. Interestingly, feelings of being discriminated against are not confined to participants who also report being a member of an underrepresented demographic group in their chosen field. People of all different demographics are included in this measure.[3] We check that randomizing our participants into conditions was successful and find that *Past Discrimination* is not significantly different between *Stable* and *Unstable* conditions. See Table 5 Panel A for more details. We formally test H2a with the model presented in Table 3. We find that *Past Discrimination* significantly predicts *Preferred Evaluator* (p = 0.01), consistent with H2a.

Additional tests of H2a and H2b are conducted using data collected in experiment 2. We test the effect of *Past Discrimination, Workspace*, and their interaction on *Preferred Evaluator* using an ANCOVA model.[4] We again find that *Past Discrimination* significantly predicts *Preferred Evaluator* (p < 0.01; Table 6, Panel A).[5] Our study thus finds robust support for H2a

[3] Among participants that also report being part of an underrepresented demographic in their chosen field, 43% responded that they at least "somewhat agree", while among participants that do not belong to an underrepresented group, 26% report that they at least "somewhat agree." These statistics indicate that perceptions of workplace discrimination are not limited to members of underrepresented groups.

[4] In untabulated results, we run a similar analysis while including the interaction between *Social Intelligence* and *Workspace* as an additional independent variable. Inferences made from both models are consistent. In fact, results of this alternate model are generally stronger.

[5] We once again check that *Past Discrimination* is not significantly different between the two *Workspace* conditions (Table 5, Panel B).

using two different experiments with two different experimental samples. H2b predicts that the effect of past discrimination on preferred evaluator type will be larger when employees have limited opportunities for social contact compared to when they have ample opportunities for social contact. We find an interactive effect between *Workspace* and *Past Discrimination* such that past discrimination has a stronger effect on preferred evaluator when working remotely than when in a shared office (p = 0.03). Further analysis (Table 6, Panel A) shows that the effect of *Past Discrimination* on *Preferred Evaluator* is significant (p < 0.01) when *Workspace* is remote, and not significant in a shared office (p = 0.23).

To determine if the interaction is consistent with the predicted pattern, we separate participants into four groups using a median split of *Past Discrimination* and *Workspace.* We then examine the mean *Preferred Evaluator* in each of these groups (see Table 6, Panel B). This pattern is consistent with our prediction. We find that participants who have experienced more workplace discrimination have a stronger preference for AI-driven evaluation systems than participants who have experienced less workplace discrimination in both conditions of *Workspace.* Further, as predicted in H2b, we find that participants that have experienced more past discrimination have a stronger preference for an AI-driven performance evaluation when working remotely than when working in a shared office.

Tests of H3a and H3b

H3a predicts that employees who have lower social intelligence will show relatively higher preferences for AI-driven systems versus human-driven systems than employees who have higher social intelligence. H3b predicts that the effect of social intelligence on preferred evaluator type will be larger when employees have more opportunities for social contact in a shared office. The minimum *Social Intelligence* in our sample is 51, the maximum is 145, the mean is 94.28 and the standard deviation is 17.96 (see Table 2). In order to ensure that there is balance between the cells, we test that *Social Intelligence* is not significantly different between the two conditions (see Table 5, Panel B). The results show no significant difference between conditions. We conduct a median split on *Social Intelligence* and find participants with low social intelligence have a stronger relative preference for an AI-driven system (mean *Preferred Evaluator* of 3.37) than participants with high social intelligence (mean *Preferred Evaluator* of 2.44; results presented in Table 7, Panel A).

We use an ANCOVA model to test the effect of *Social Intelligence*, *Workspace,* and their interactive effect on *Preferred Evaluator* while controlling for *Past Discrimination* (see Table 7, Panel B). We find a significant main effect of *Social Intelligence* ($F_{(1,150)} = 4.07$, $p = 0.04$), and a marginally significant interactive effect of *Social Intelligence* and *Workspace* ($F_{(1,150)} = 3.53$, $p = 0.06$). Consistent with H3b, we find that the effect of *Social Intelligence* on *Preferred Evaluator* is stronger in the *Shared* condition than in the *Remote* condition. These results support H3a and H3b.

Additional Analysis

We posit in the theory section that employee preferences are based on their perceptions of when a performance evaluation would be fairer. However, it is possible that instead of fairness, participants value some other feature of the two systems. For example, they may believe that the human will be a more lenient judge of performance or that the AI will be easier to deceive, resulting in greater likelihood of a reward. In order to confirm that participants are selecting the evaluator based on fairness concerns, we also directly asked participants which evaluator they believed would be the fairest. *Fairest Evaluator* is measured using a scale from 1 to 5 where 1 is a human and 5 is an AI.

We first examine the univariate relationship between *Fairest Evaluator* and *Preferred Evaluator*. The two are significantly positively correlated with one another in both experimental populations ($p < 0.01$) (Table 1 Panel B, and Table 2, Panel B). This raw correlation provides initial evidence supporting our theory. We then re-examine our hypotheses by performing similar tests but replacing *Preferred Evaluator* with *Fairest Evaluator* as the dependent variable. Consistent with H1, we find that *Environmental Stability* significantly predicts *Fairest Evaluator* ($p = 0.02$) (Table 8, Panel B) where AI is relatively more preferred in a stable environment. We also find a significant effect of *Past Discrimination* on *Fairest Evaluator* with the data from experiment 1 ($p < 0.01$, Table 8, Panel B), consistent with H2a. People who report having been discriminated against are more likely to believe that AI-driven systems will be fairer than those who did not report that. In Table 9, we test whether *Past Discrimination* and *Workspace* have an interactive effect on *Fairest Evaluator.* Although we find support for H2b in our main analysis, we do not find a significant result ($p = 0.97$) in this supplemental analysis. It is unclear whether there is a theoretical reason for this null result or whether it is driven by limitations of our study design. We leave the examination of this to future research. As shown in Table 10, Panel B, we do not find a significant main effect of *Social Intelligence* on *Fairest Evaluator.* We do, however, find a significant interactive effect ($F_{(1, 150)} = 3.62$, $p = 0.06$) of *Social Intelligence* and *Workspace* on *Fairest Evaluator*. We also demonstrate through simple effects tests that the effect of *Social Intelligence* on *Fairest Evaluator* is significant ($p = 0.02$) in the *Shared* condition, and not significant in the *Remote* condition ($p = 0.79$). These results are consistent with H3b.

We continue to explore the role of fairness by testing whether *Fairest Evaluator* mediates the effect of *Social Intelligence* on *Preferred Evaluator* and whether this indirect effect is moderated by *Workspace* (see Figure 8 for visual depiction). To test this moderated mediation model, we use the simultaneous OLS regression method and Model 7 of the PROCESS macro as outlined in Hayes (2018). This method allows us to simultaneously test whether *Fairest Evaluator* mediates the effect of *Social Intelligence* on *Preferred Evaluator* and whether such an effect is moderated by *Workspace.* In doing so, we connect our prior analyses on how the independent variables of interest affect participants' perceptions of the fairest evaluator and, in turn, how perceptions of fairness affect their preferred evaluator type. Regressions are conducted using 5,000 bootstrap samples and a 95% confidence interval. We find a significant interactive effect of *Social Intelligence* and *Workspace* on *Fairest Evaluator* (t = 1.90, p = 0.03, one-tailed test) (see Table 11). Additionally, we find that *Fairest Evaluator* significantly predicts *Preferred Evaluator* (t = 6.18, p < 0.01). We also find evidence of a statistically significant indirect effect of *Social Intelligence* on *Preferred Evaluator* through *Fairest Evaluator* in the *Shared* condition (95% confidence interval: [-0.0160, -0.0007]) but not in the Remote condition (95% confidence interval: [-0.0062, 0.0091]). The index of moderated mediation tests for a significant difference in the strength of the indirect effect between levels of *Workspace* and is nearly significant at a 95% confidence level (confidence interval: [-0.0003, 0.0208]). These results indicate that in an environment where there are able opportunities for social contact, people with higher social intelligence are less likely to prefer an AI-driven evaluation system than people with lower social intelligence because they believe AI is less capable of making a fair evaluation than a human manager. In sum, our results not only support the hypotheses, but also the underlying reasoning.

VI. CONCLUSION

While previous research that examines employee preferences for AI-driven systems versus human-driven systems finds that employees always prefer human evaluators, this study documents that preferences between AI and human evaluators are not uniform. We show that the stability of the organizational operating environment, employee history with discrimination, employee social intelligence, and the available opportunities for social contact during work change employee preferences in predictable ways. In more stable operating environments, employees show a relatively higher preference for AI-driven performance evaluation systems. Employees who have been discriminated against also show a relatively higher preference for AI than those who have not, and this effect is stronger in remote settings than in shared offices. We also find social intelligence predicts preferences for AI, but only when employees are working in a shared office space where there are able opportunities for social contact. Furthermore, we show that these changes in preference occur because depending on the situation employees have different beliefs about which evaluator will be the fairest. Firms can use these findings to guide investments into AI-driven evaluation systems by better understanding in which settings employees are more likely to accept the systems as fair. For example, our study suggests that companies in mature industries with more predictable operating environments or with historical struggles with workplace discrimination may be better candidates for AI-driven performance evaluation.

While this paper is an important first step to examine cross-sectional differences in employees' preferences towards AI, it only examines a few situations. We hope that future research will continue to examine other settings and add to these results. Furthermore, we examine only one of Newman et al.'s two mechanisms, decontextualization. We see fertile ground for future research that examines the second mechanism, quantification. Also, our results rest on employees' *perceived* decontextualization. Since each AI-driven evaluation system functions differently, it would also be valuable to examine how perceptions of decontextualization and quantification differ depending on the specific (technological) features of the AI. The impact of how these perceptions change when more explanation is provided on the working of the technology will also be an important addition to this line of research.

Our study also has limitations due to it being a scenario-based experiment. While studying employee perceptions and preferences is key in understanding AI, extending this work to real effort tasks in a laboratory setting or a field setting will help build on what this study has shown. We can show how employees believe they will react to the systems, but documenting how employees actually do react is a valuable future direction for managerial accounting work in general and our study in particular. Another limitation of our study is the fact that our experiment takes place using participants in an online labor market. Because participants are all employed doing online remote work, they may have a much higher baseline level of trust in new technologies than other employees. Extending this work to different subject groups will be important to understanding AI in different types of workspaces.

Despite its limitations, our study is an important first step in management accounting research investigating the effects of AI on performance evaluation. This new technology is being used by a small but increasing number of firms. It is important that management accountants join in the discussion of how and when to implement it in firms. This discipline's deep understanding of the evaluation process can help developers create more effective technology. Moreover, it can help managers understand which employee groups may be a better fit for adoption of an AI-driven system and communicate to employees how these systems will differ from traditional human-driven systems.

BIBLIOGRAPHY

Aspan, M. (2020, January 24). This tech giant says A.I. has already helped it save $1 billion. Fortune. https://fortune.com/2020/01/24/ai-ibm-human-resources/

Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. The quarterly journal of economics, 109(4), 1125-1156.

Bandiera, O., Barankay, I., & Rasul, I. (2009). Social connections and incentives in the workplace: Evidence from personnel data. Econometrica, 77(4), 1047-1094.

Berger, B., Adam, M., Rühr A., & Benlian, A. (2021). Watch me improve – Algorithm aversion and demonstrating the ability to learn. Business & Information Systems Engineering. 63.1: 55-68.

Bol, J. C. (2008). Subjectivity in Compensation Contracting. Journal of Accounting Literature, 27, 1-24.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. The Accounting Review, 86(5), 1549-1575.

Bol, J. C., & Leiby, J. (2018). Subjectivity in Professionals' Incentive Systems: Differences between Promotion-and Performance-Based Assessments. Contemporary Accounting Research, 35(1), 31-57.

Bol, J. C., Margolin M., Schaupp D., (2021). Multi-Rater Performance Evaluation and Calibration: Managing Multiple Opinions. Working paper.

Bol, J. C., & Smith, S. D. (2011). Spillover effects in subjective performance evaluation: Bias and the asymmetric influence of controllability. The Accounting Review, 86(4), 1213-1230.

Brown, J., Burke, J., & Sauciuc, A. (2021). Workforce Diversity and Artificial Intelligence: Implications for AI Integration into Performance Evaluation Systems. Working Paper.

Buchheit, S., Doxey, M. M., Pollard, T., & Stinson, S. R. (2018). A technical guide to using Amazon's Mechanical Turk in behavioral accounting research. Behavioral Research in Accounting, 30(1), 111-122.

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), 220-239.

Cappelli, P., & Tavis, A. (2016). The performance management revolution. Harvard Business Review. https://hbr.org/2016/10/the-performance-management-revolution

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809-825.

Chan, H & Dimauro, J. (2020). Black Lives Matter movement sparks outcry for corporations to show diversity gains. Retrieved from https://blogs.thomsonreuters.com/answerson/black-lives-matter-corporate-diversity-gains/

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. Human Resource Management Review, 31(1), 100698.

Chung, D. J., Huber, I., Murthy, V., Sunku, V., & Weber, M. (2019). Setting better sales goals with analytics. Harvard Business Review. Retrieved from https://hbr.org/2019/07/setting-better-sales-goals-with-analytics.

Commerford, B. P., Dennis, S. A., Joe, J. R., & Ulla, J. (2021). Man versus machine: Complex estimates and auditor reliance on artificial intelligence. Forthcoming in the Journal of Accounting Research

Cunningham, L. (2015) In big move, Accenture will get rid of annual performance reviews and rankings. Retrieved from https://www.washingtonpost.com/news/on-leadership/wp/2015/07/21/in-big-move-accenture-will-get-rid-of-annual-performance-reviews-and-rankings/.

Demeré, B. W., Sedatole, K. L., & Woods, A. (2019). The role of calibration committees in subjective performance evaluation systems. Management Science, 65(4), 1562-1585.

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.

Downes, P. E., & Choi, D. (2014). Employee reactions to pay dispersion: A typology of existing research. Human Resource Management Review, 24(1), 53-66.

Enaible. (2021). New technology for a new era. https://enaible.io/how-it-works/

Engellandt, A., & Riphahn, R. T. (2011). Evidence on incentive effects of subjective performance evaluations. ILR Review, 64(2), 241-257.

Farrell, A. M., Grenier, J. H., & Leiby, J. (2017). Scoundrels or stars? Theory and evidence on the quality of workers in online labor markets. The Accounting Review, 92(1), 93-114.

Fecheyr-Lippens, B., Schaninger, B., & Tanner, K. (2015). Power to the new people analytics. McKinsey Quarterly. https://www.mckinsey.com/business-functions/organization/our-insights/power-to-the-new-people-analytics#

Fisher, A. (2019). An Algorithm May Decide Your Next Pay Raise. Retrieved from https://fortune.com/2019/07/14/artificial-intelligence-workplace-ibm-annual-review/

Fisher, J. G., Maines, L. A., Peffer, S. A., & Sprinkle, G. B. (2005). An experimental investigation of employer discretion in employee performance evaluation and compensation. The Accounting Review, 80(2), 563-583.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications, 40(16), 6266-6282.

Gibbs, M. J., Merchant, K. A., Stede, W. A. V. D., & Vargus, M. E. (2005). The benefits of evaluating performance subjectively. Performance Improvement, 44(5), 26-32.

Glassdoor. (2019). Diversity & Inclusion Study 2019. Retrieved from https://about-content.glassdoor.com//app/uploads/sites/2/2019/10/Glassdoor-Diversity-Survey-Supplement-1.pdf.

Greene, T. (2018, July 10). IBM is using its AI to predict how employees will perform. Retrieved from https://thenextweb.com/artificial-intelligence/2018/07/10/ibm-is-using-its-ai-to-predict-how-employees-will-perform/.

Hayes, A. F. (2018). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford publications.

Heaven, W. (2020). This startup is using AI to give workers a "productivity score". Retrieved from https://www.technologyreview.com/2020/06/04/1002671/startup-ai-workers-productivity-score-bias-machine-learning-business-covid/

Higgins, C. A., Judge, T. A., & Ferris, G. R. (2003). Influence tactics and work outcomes: A meta-analysis. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 24(1), 89-106.

Hu, W. (2021). The Lost Productivity: An Experimental Investigation of Human Versus Algorithm-Based Discretion in Incomplete Compensation Contracts. Working paper.

Hughes, C., Robert, L., Frady, K., & Arroyos, A. (2019). Artificial intelligence, employee engagement, fairness, and job outcomes. In Managing Technology and Middle-and Low-skilled Employees. Emerald Publishing Limited.

Ittner, C. D., Larcker, D. F., & Meyer, M. W. (2003). Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. The accounting review, 78(3), 725-758.

Heaven, W. (2020). This startup is using AI to give workers a "productivity score". MIT Technology Review. https://www.technologyreview.com/2020/06/04/1002671/startup-ai-workers-productivity-score-bias-machine-learning-business-covid/

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

Holsinger, L., et al. (2019). Performance Transformation in the Future of Work. Mercer. https://www.mercer.com/our-thinking/career/performance-transformation-in-the-future-of-work.html

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. Journal of labor Economics, 26(1), 101-136.

Janssen, O. (2001). Fairness perceptions as a moderator in the curvilinear relationships between job demands, and job performance and job satisfaction. Academy of management journal, 44(5), 1039-1050.

Jawahar, I. M. (2007). The influence of perceptions of fairness on performance appraisal reactions. Journal of Labor Research, 28(4), 735-754.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366-410.

Knight, R. (2020). How to Do Performance Reviews — Remotely. Harvard Business Review. https://hbr.org/2020/06/how-to-do-performance-reviews-remotely

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, *90*(5), 1346-1361.

Lund, S. Madgavkar, A., Manyika, J., Smit, S., Ellingrud, K., and Robinson, O. (2021). The future of work after COVID-19. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-after-covid-19

Mackenzie, L. N., Wehner, J., & Kennedy, S. (2020). How do you evaluate performance during a pandemic? Harvard Business Review. https://hbr.org/2020/12/how-do-you-evaluate-performance-during-a-pandemic

Miceli, M. P., Jung, I., Near, J. P., & Greenberger, D. B. (1991). Predictors and outcomes of reactions to pay-for-performance plans. Journal of Applied Psychology, 76(4), 508.

Moers, F. (2005). Discretion and bias in performance evaluation: the impact of diversity and subjectivity. Accounting, Organizations and Society, 30(1), 67-80.

Moise, I. & Cruise, S. (2020). Goldman Sachs executive's email making plea for racial equality goes viral at firm. Retrieved from https://www.reuters.com/article/us-usa-goldman-sachs-race/goldman-sachs-executives-email-making-plea-for-racial-equality-goes-viral-at-firm-idUSKBN23C086

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions. Organizational Behavior and Human Decision Processes, 160, 149-167.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. Judgment and Decision making, 5(5), 411-419.

Pettigrew, T. F. (1998). Intergroup contact theory. Annual review of psychology, 49(1), 65-85.

Pettigrew, T. F. (2021). Advancing intergroup contact theory: Comments on the issue's articles. Journal of Social Issues, *77*(1), 258-273.

Poon, J. M. (2004). Effects of performance appraisal politics on job satisfaction and turnover intention. Personnel review.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted?. Journal of Forecasting, *36*(6), 691-702.

Prendergast, C. (1999). The provision of incentives in firms. Journal of economic literature, 37(1), 7-63.

Prendergast, C., & Topel, R. (1993). Discretion and bias in performance evaluation. European Economic Review, 37(2-3), 355-365.

Prendergast, C., & Topel, R. H. (1996). Favoritism in organizations. Journal of Political Economy, 104(5), 958-978.

Rogel, C. (2020). How much do performance reviews actually cost and are they really worth it? Decisionwise. https://decision-wise.com/how-much-do-performance-reviews-actually-cost-and-are-they-really-worth-it/

Rosenbaum, E. (2019, April 3). IBM artificial intelligence can predict with 95% accuracy which workers are about to quit their jobs. Retrieved from https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html

Silvera, D., Martinussen, M., & Dahl, T. I. (2001). The Tromsø Social Intelligence Scale, a self-report measure of social intelligence. Scandinavian journal of psychology, 42(4), 313-319.

Simon, H. A. (1990). Bounded rationality. In Utility and probability (pp. 15-18). Palgrave Macmillan, London.

Simons, T., & Roberson, Q. (2003). Why managers should care about fairness: the effects of aggregate justice perceptions on organizational outcomes. Journal of applied psychology, 88(3), 432.

Taylor, M. S., Tracy, K. B., Renard, M. K., Harrison, J. K., & Carroll, S. J. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. Administrative science quarterly, 495-523.

Trevor, C. O., Reilly, G., & Gerhart, B. (2012). Reconsidering pay dispersion's effect on the performance of interdependent work: Reconciling sorting and pay inequality. Academy of Management Journal, 55(3), 585-610.

Tzuo, T. (2017) Why millennials actually want more feedback at work. Fortune. https://fortune.com/2017/02/14/leadership-career-advice-millennials-instant-feedback-retention-work-life-balance/

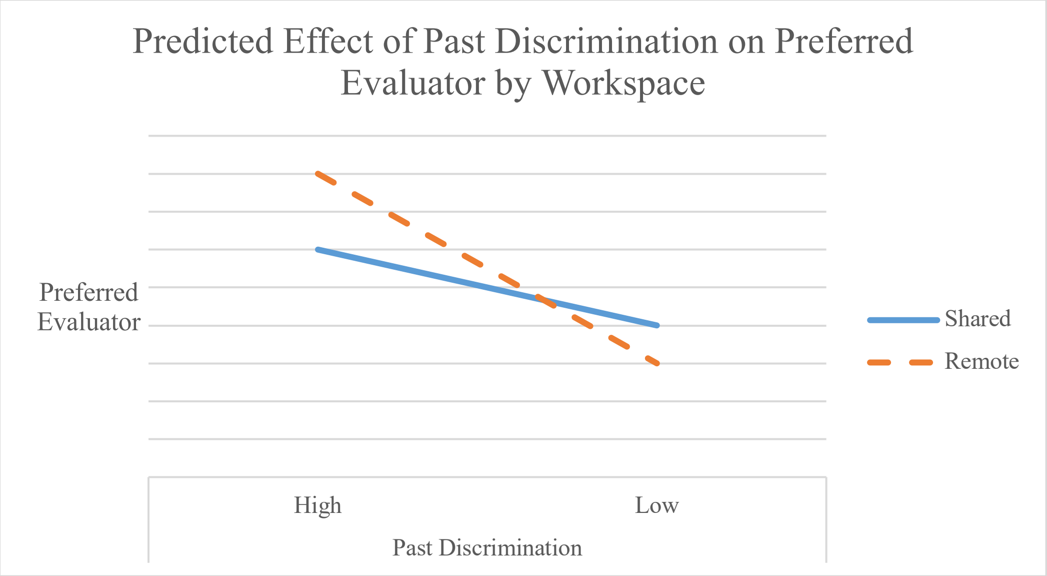
WorldatWork. (2019). Contemporary Performance Evaluation Practices Survey.

Young, S. M., Du, F., Dworkis, K. K., & Olsen, K. J. (2016). It's all about all of us: The rise of narcissism and its implications for management control system research. Journal of Management Accounting Research, 28(1), 39-55..

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

Figure 1: Predicted and Actual Results for H2a and H2b

Panel A: Predicted Results



Panel B: Actual Results

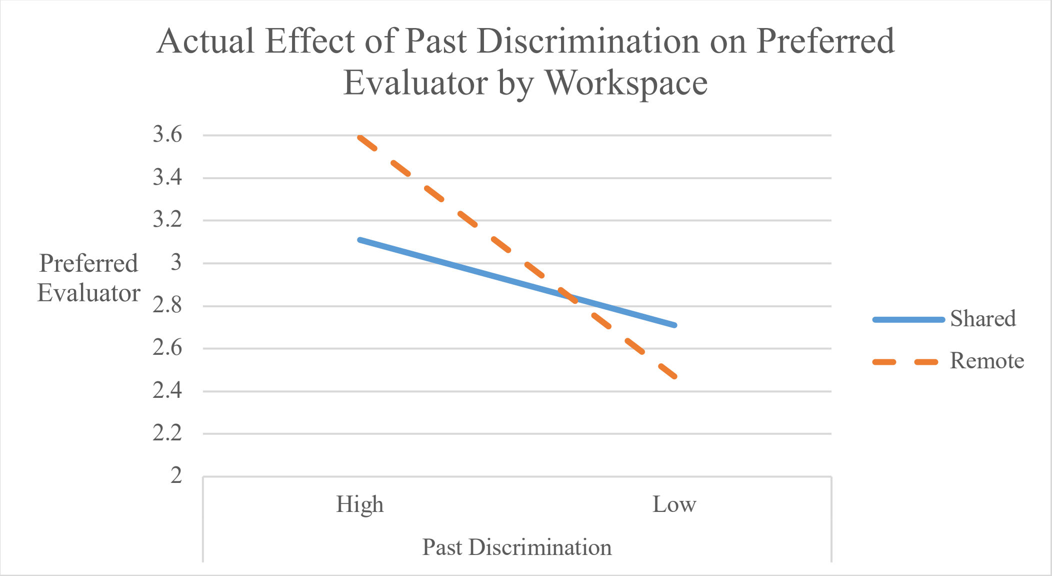
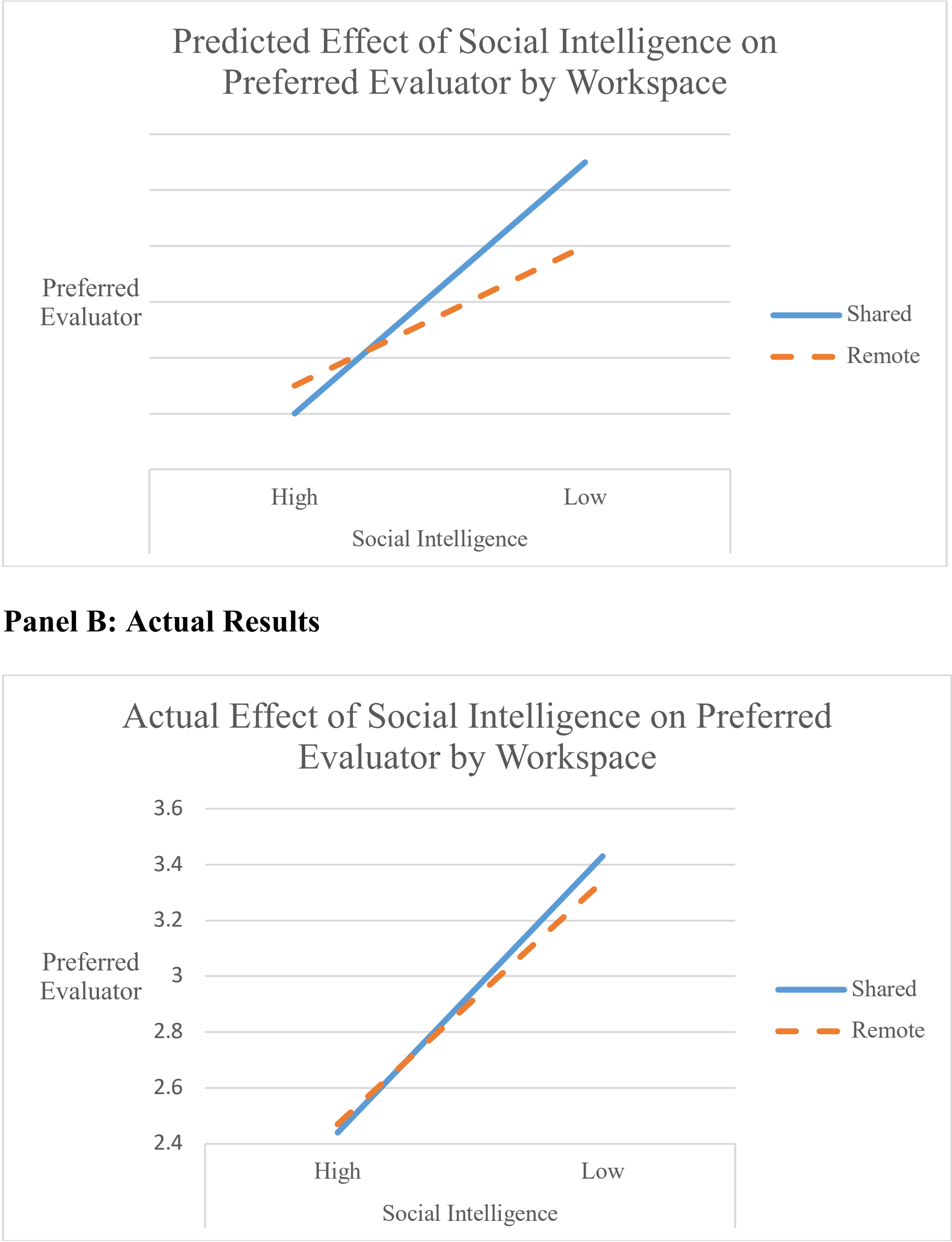


Figure 2: Predicted and Actual Results for H3a and H3b



Panel A: Predicted Results

Panel B: Actual Results

Figure 3: Performance Data Format Information

Participants in both experiments saw the following information:

To assess your sales performance, the company records your total sales, the prices you negotiated, and the premium add-ons that you sold. In addition, to help put your sales numbers into perspective and adjust for factors outside your control, the company collects several other pieces of information. For example, the company collects national and local economic indicators that are relevant to your sales performance during the month such as global oil prices, local unemployment levels, federal interest rates, and extreme weather. The company also collects information on competitors like their market share and pricing strategy. At the end of the month, your sales performance is evaluated by analyzing your sales number with consideration for these context variables described above.

To assess your performance as a trainer, customers fill out a survey about their satisfaction with the training. Some of the survey questions are numerical and answered on a standardized scale, while others are open-ended and allow the customer to provide a free response. Open-ended questions are just as important as the numerical questions because they provide more specific information about your strengths and weaknesses as a trainer. The organization also collects information that could affect customer satisfaction with the training such as trends in robot usage and the prior experience the customers you trained had with robotics. As with sales data, the training performance is evaluated by analyzing the survey responses in light of the customer types that you served.

Your performance is weighted so that 50% of your final evaluation is based on sales performance, and 50% is based on your effectiveness as a trainer.

Figure 4: Timeline of Experiments

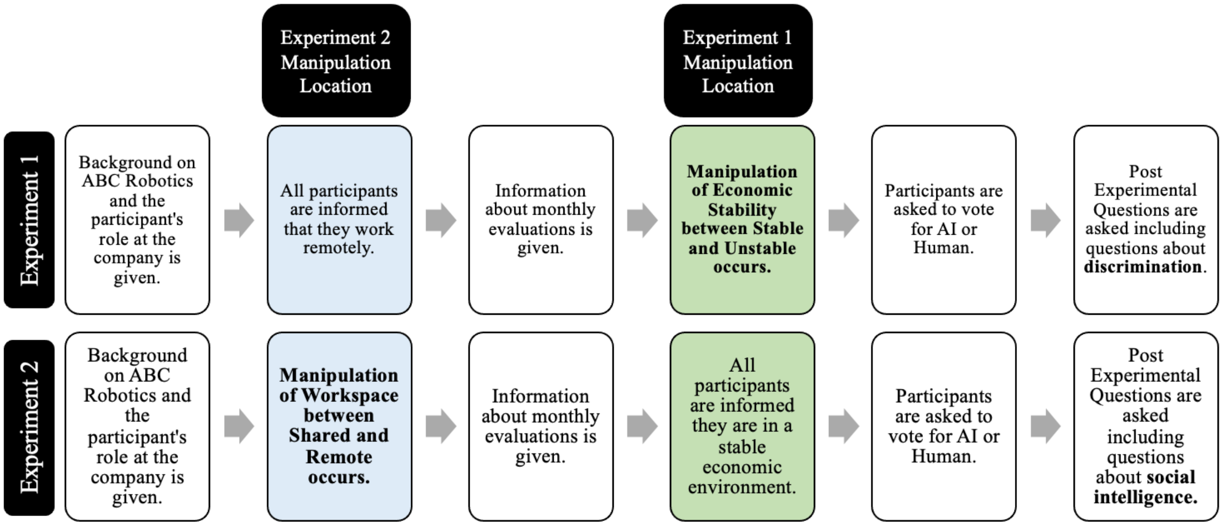

Experiment 2 Manipulation Location
Experiment 1 Manipulation Location
Experiment 1
Background on ABC Robotics and the participant's role at the company is given.
All participants are informed that they work remotely.
Information about monthly evaluations is given.
Manipulation of Economic Stability between Stable and Unstable occurs.
Participants are asked to vote for AI or Human.
Post Experimental Questions are asked including questions about discrimination.
Experiment 2
Background on ABC Robotics and the participant's role at the company is given.
Manipulation of Workspace between Shared and Remote occurs.
Information about monthly evaluations is given.
All participants are informed they are in a stable economic environment.
Participants are asked to vote for AI or Human.
Post Experimental Questions are asked including questions about social intelligence.

Figure 5: Environmental Stability Manipulation

Stable:

The quality of the data at ABC Robotics is high. There is historical data on all variables for several years and the data set is complete. The current business environment mostly stable. Both the AI algorithm and the human manager have experience assessing performance under these circumstances.

Unstable:

The quality of the data at ABC Robotics is high. There is historical data on all variables for several years and the data set is complete. The current business environment, however, is highly unstable: the world has been hit by the COVID-19 pandemic and related financial crisis. Neither the AI algorithm nor the human manager has any experience with assessing performance under these circumstances.

Figure 6: Opportunity for Social Contact in the Workspace Manipulation

Shared Office:

You work as a salesperson selling the robots and conducting on-site trainings with employees at the companies that buy the robots. At ABC Robotics, everyone, including senior leadership, works in a shared office space whenever they are not with their clients. You spend about half of your time with clients and the other half of your time working from the office finding sales leads, doing paperwork, etc.

ABC Robots has a comfortable and well decorated office space. You have a large desk with a top quality office chair, multiple computer monitors, and lots of healthy snacks for you to eat throughout the day. You even have a fancy coffee machine for your morning pick me up.

Below you will see a picture of the desk you work at.

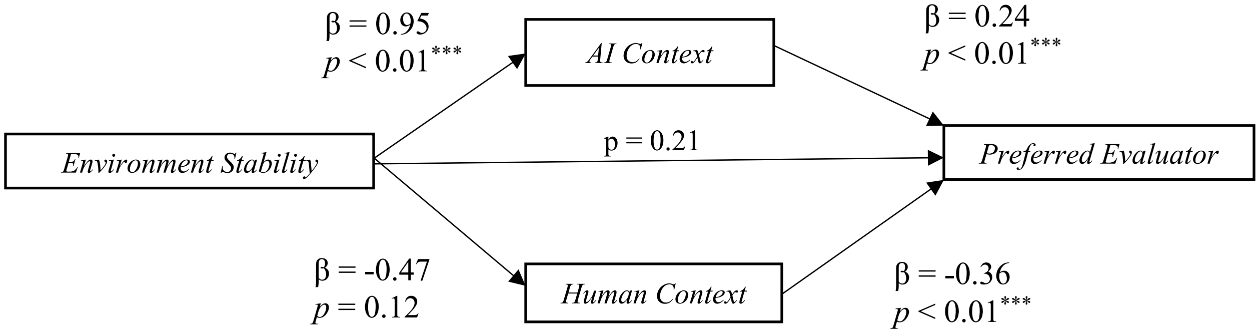
Remote:

You work as a salesperson selling the robots and conducting on-site trainings with employees at the companies that buy the robots. At ABC Robotics, everyone, including senior leadership, works from home whenever they are not with their clients. You spend about half of your time with clients and the other half of your time working from home finding sales leads, doing paperwork, etc.

You have a comfortable and well decorated home office space. You have a large desk with a top quality office chair, multiple computer monitors, and lots of healthy snacks for you to eat throughout the day. You even have a fancy coffee machine for your morning pick me up.

Below you will see a picture of the desk you work at.

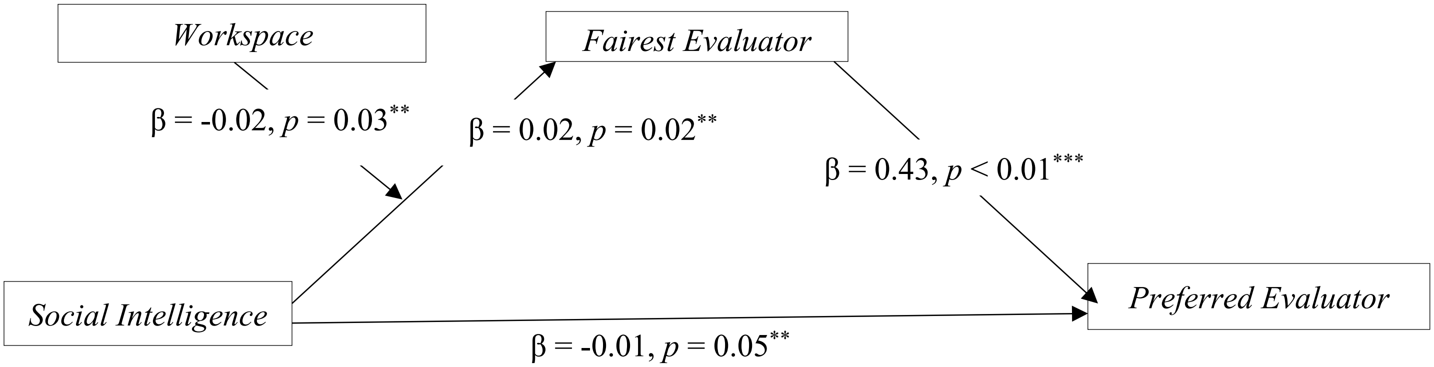
Figure 7: Mediation of the Effect of *Environment Stability* on *Preferred Evaluator*



Indirect effects of *Environment Stability* on *Preferred Evaluator*

	Lower CI	Upper CI
through *AI Context*	0.04	0.44
through *Human Context*	-0.04	0.40

Figure 8: The Effect of *Social Intelligence* on *Preferred Evaluator* through *Fairest Evaluator*, Moderated by *Workspace*



Indirect effects of *Social Intelligence* on *Preferred Evaluator* through *Fairest Evaluator* for each level of *Workspace*

	Lower CI	Upper CI
Shared	-0.0160	-0.0007
Remote	-0.0062	0.0091

Table 1: Experiment 1 Descriptive Statistics and Pearson Correlation Table

Panel A: Descriptive Statistics

Parameter	*Min*	*Max*	*Median*	*Mean*	*SD*
Preferred Evaluator	1	5	3	2.73	1.30
Fairest Evaluator	1	5	3	2.76	1.15
Past Discrimination	1	7	3	3.44	1.89
Age (years)	18	73	35	38	11.46
Work Experience (years)	1	56	15	16	11.11
Education	2	5	4	3.6	0.85

Panel B: Pearson Correlation Matrix

Parameter	*1*	*2*	*3*	*4*	*5*	*6*
1. Preferred Evaluator	1					
2. Fairest Evaluator	0.63^{***}	1				
3. Past Discrimination	0.23^{*}	0.27^{***}	1			
4. Age (years)	--0.01	-0.14	0.04	1		
5. Work Experience	-0.06	-0.14	-0.07	0.86^{***}	1	
6. Education	0.19	0.14	0.00	0.03	-0.06	1

Variable Definitions:

1. *Preferred Evaluator*: participants' responses to the statement, "Which would you rather have evaluating you at ABC Robotics: a human manager or an artificial intelligence algorithm?" Responses were recorded on a five-point scale from "I strongly prefer a human manager to evaluate me" (coded as 1) to "I strongly prefer an artificial intelligence algorithm to evaluate me" (coded as 5).
2. *Fairest Evaluator:* participants' responses to the statement, "Which would you expect to be better at making a fair final evaluation, AI or a human manager?" Responses were recorded on a five-point scale from "AI is much better" (coded as 1) to "A human manager is much better" (coded as 5).
3. *Past Discrimination:* participants' responses to the statement, "I have been subject to discrimination at work." Responses were recorded on a seven-point scale from "Strongly disagree" (coded as 1) to "Strongly agree" (coded as 7).
4. *Age:* participants' self-reported age in years.
5. *Work Experience:* participants' self-reported work experience in years.
6. *Education*: participants' self-reported education level recorded on a five-point scale from "less than a high school degree" (coded as 1) to "higher than a college degree" (coded as 5).

Throughout the paper: *, **, *** denote significance at the, 0.1, 0.05 and <0.01 level.

Table 2: Experiment 2 Descriptive Statistics and Correlation Tables

Panel A: Descriptive Statistics

Parameter	*Min*	*Max*	*Median*	*Mean*	*SD*
Preferred Evaluator	1	5	3	2.91	1.33
Social Intelligence	51	145	89	94.28	17.96
Fairest Evaluator	1	5	3	3.23	1.30
Age (years)	20	69	32	36	10.32
Work Experience (years)	0	47	10	12	9.36
Education	2	5	4	3.73	0.78

Panel B: Correlations

Parameter	*1*	*2*	*3*	*4*	*5*	*6*
1. Preferred Evaluator	1					
2. *Social Intelligence*	-0.28***	1				
3. *Fairest Evaluator*	0.46***	-0.11	1			
4. *Age*	0.04	0.15	-0.10	1		
5. *Work Experience*	-0.12	0.31***	-0.05	0.78***	1	
6. *Education*	0.26***	-0.14	0.22***	-0.04	-0.10	1

Variable Definitions:
2. Social Intelligence: Sum of the responses to 21 items of the Social Intelligence Scale created and validated in Silvera et al. (2001). See appendix for complete scale.
See Table 1, Panel A for all other variable definitions.

Table 3: Main Test of H1

Panel A: *Preferred Evaluator* by *Environment Stability*

	n	*Preferred Evaluator*	SD
Unstable	53	2.43	1.37
Stable	71	2.96	1.21

Panel B: Main Test of H1: ANCOVA

Model: *Preferred Evaluator = Environment Stability + Past Discrimination +* ϵ

	df	*MS*	*p*
Environment Stability	1	11.38	$<0.01^{***}$
Past Discrimination	1	13.97	$<0.01^{***}$
Error	121		

Panel A reports cell sizes, means, and standard deviations of *Preferred Evaluator* for each condition of *Environment Stability,* collected from experiment 1. See Table 1 for variable definitions

Panel B reports the results of an ANCOVA model including *Environment Stability* and *Past Discrimination* as independent variables and *Preferred Evaluator* as the dependent variable. See Table 1 for variable definitions.

Table 4: Mediation Analysis of the Effect of *Environment Stability* on *Preferred Evaluator* While Controlling for *Past Discrimination*

Panel A: Outcome Variable: *AI Context*

	β	SE	*t*	*p*
Environment Stability	0.95	0.35	2.74	$<0.01^{***}$
Past Discrimination	0.05	0.91	0.54	0.59

Panel B: Outcome Variable: *Human Context*

	β	SE	*t*	*p*
Environment Stability	-0.47	0.30	-1.58	0.12
Past Discrimination	-0.39	0.08	-4.90	$<0.01^{***}$

Panel C: Outcome Variable: *Preferred Evaluator*

	β	SE	*t*	*p*
Environment Stability	0.24	0.19	1.26	0.21
AI Context	0.22	0.05	4.65	$<0.01^{***}$
Human Context	-0.36	0.06	-6.48	$<0.01^{***}$
Past Discrimination	0.03	0.05	0.59	0.55

Panel D: Indirect Effects of *Environment Stability* on *Preferred Evaluator*

	β	SE	*Lower CI*	*Upper CI*
Total	0.38	0.16	0.09	0.70
AI Context	0.21	0.10	0.04	0.44
Human Context	0.17	0.11	-0.04	0.40

We conduct a mediation analysis of the effect of *Environment Stability* on *Preferred Evaluator* through two parallel mediators: *AI Context* and *Human Context*. We include *Past Discrimination* as a covariate. We conduct this analysis using the simultaneous OLS regression method and Model 4 outlined in Hayes (2018). See Figure 7 for a visual depiction of the model and results. Tests are conducted using a 95% confidence interval and 5,000 bootstrap samples. *AI Context* is participants' response to the following PEQ item on a seven-point scale from strongly disagree (coded as 1) to strongly agree (coded as 7): "When casting my vote for who would evaluate me (AI or a human manager), I was concerned that a human manager wouldn't be able to fairly consider the circumstances I was in." *Human Context* is participants' response to the following PEQ item on the same scale: "When casting my vote for who would evaluate me (AI or a human manager), I was concerned that a human manager wouldn't be able to fairly consider the circumstances I was in." See Table 1 for other variable definitions.

Table 5: Randomization Checks

Panel A: Experiment 1 *Past Discrimination* between Levels of *Environment Stability*

	within *Stable*	within *Unstable*	Difference	*t*	*p*
Past Discrimination	3.74	3.21	0.53	1.54	0.13

Panel B: Experiment 2 *Past Discrimination* and *Social Intelligence* between Levels of *Workspace*

	within *High*	within *Low*	Difference	*t*	*p*
Past Discrimination	3.91	3.73	0.17	0.55	0.58
Social Intelligence	95.03	93.57	1.46	0.50	0.62

We test for successful randomization of participants by comparing mean levels of measured variables across conditions of our manipulated variables. We find no significant differences in any of the measured variables between conditions, suggesting successful randomization.

Table 6: Main Tests of H2a and H2b

Panel A: The Effect of *Past Discrimination* and *Workspace* on *Preferred Evaluator*

ANCOVA: *Preferred Evaluator = Workspace + Past Discrimination + Past Discrimination*Workspace +* ϵ

	df	*MS*	*F*	*p*
Workspace	1	4.89	3.09	0.08^{*}
Past Discrimination	1	24.95	15.77	$<0.01^{***}$
Past Discrimination Workspace*	1	7.25	4.58	0.03^{**}
Error	151			

Effects of *Past Discrimination* on *Preferred Evaluator* at each level of *Workspace*

	df	*MS*	*F*	*p*
Shared	1	2.50	1.46	0.23
Remote	1	31.53	21.67	$<0.01^{***}$

Panel B: *Preferred Evaluator* by *Workspace* and Median Split of *Past Discrimination*

	Workspace		
Past Discrimination	*Shared*	*Remote*	Difference
High	3.11	3.59	0.48
Low	2.71	2.47	-0.24
Difference	0.40	1.12	

In Panel A, we conduct an ANCOVA using *Preferred Evaluator* as the dependent variable and *Workspace*, *Past Discrimination*, and the interactive effect as independent variables. See Table 1 for variable definitions.

Panel B reports the mean *Preferred Evaluator* of the four groups created by performing a median split of the sample by both *Workspace* and *Past Discrimination*. See Table 1 for variable definitions.

Table 7: Main Tests of H3a and H3b

Panel A: The Effect of *Social Intelligence* and *Workspace* on *Preferred Evaluator*

ANCOVA: *Preferred Evaluator = Social Intelligence* + *Workspace* +*Social Intelligence* Workspace* + *Past Discrimination* + ϵ

	Df	MS	F	p
Social Intelligence	1	6.37	4.07	0.04**
Workspace	1	5.11	3.27	0.07*
Social Intelligence Workspace*	1	5.51	3.53	0.06*
Past Discrimination	1	11.93	7.63	<0.01***
Error	150			

Effects of *Social Intelligence* on *Preferred Evaluator* at each level of *Workspace*

	Df	MS	F	p
Shared	1	21.95	15.31	<0.01***
Remote	1	0.97	0.66	0.42

Panel B: *Preferred Evaluator* by *Workspace* and Median Split of *Social Intelligence*

	Workspace		
Social Intelligence	*Shared*	*Remote*	Difference
High	2.44	2.47	-0.03
Low	3.43	3.34	0.09
Difference	-0.99	-0.87	

In Panel A, we conduct an ANCOVA using *Preferred Evaluator* as the dependent variable and *Workspace*, *Social Intelligence*, and the interactive effect as independent variables. We also include *Past Discrimination* as a covariate. See Table 1 for variable definitions.

Panel B reports the mean *Preferred Evaluator* of the four groups created by performing a median split of the sample by both *Workspace* and *Social Intelligence*. See Table 1 for variable definitions.

Table 8: Effect of *Environment Stability* and *Past Discrimination* on *Fairest Evaluator*

Panel A: *Fairest Evaluator* by *Environment Stability*

	n	*Fairest Evaluator*	SD
Unstable	53	2.54	1.22
Stable	71	2.92	1.08

Panel B: ANCOVA

Model: *Fairest Evaluator = Environment Stability + Past Discrimination +* ϵ

	df	*MS*	*F*	*p*
Environment Stability	1	6.38	5.34	0.02^{**}
Past Discrimination	1	14.05	11.76	$<0.01^{***}$
Error	121			

See Table 1 for variable definitions.

Table 9: The Effect of *Past Discrimination* and *Workspace* on *Fairest Evaluator*

Model: *Fairest Evaluator = Workspace + Past Discrimination + Workspace*Past Discrimination +* ϵ

	df	*MS*	*F*	*p*
Workspace	1	0.14	0.08	0.77
Past Discrimination	1	1.59	0.93	0.34
Workspace Past Discrimination*	1	<0.01	<0.01	0.97
Error	151			

See Table 1 for variable definitions.

Table 10: Effect of *Social Intelligence* and *Workspace* on *Fairest Evaluator*

Panel A: *Fairest Evaluator* by *Social Intelligence* and *Workspace*

	n	*Fairest Evaluator*	SD
High Social Intelligence	77	2.57	1.30
Low Social Intelligence	78	2.97	1.28
Shared	76	2.86	1.28
Remote	79	2.70	1.32

Panel B: The Effect of *Social Intelligence* and *Workspace* on *Fairest Evaluator*

ANCOVA: *Fairest Evaluator = Social Intelligence + Workspace+*
*Social Intelligence*Workspace + Past Discrimination +* ϵ

	df	MS	F	p
Social Intelligence	1	2.40	1.44	0.23
Workspace	1	6.80	4.07	0.04^{**}
Social Intelligence Workspace*	1	6.04	3.62	0.06^{*}
Past Discrimination	1	0.18	0.11	0.74
Error	150			

Effects of *Social Intelligence* on *Fairest Evaluator* at each level of *Workspace*

	df	MS	F	p
Shared	1	9.01	5.83	0.02^{**}
Remote	1	0.13	0.07	0.79

See Table 1 for variable definitions.

Table 11: Moderated Mediation Regression Analysis of the Effect of *Social Intelligence* on *Preferred Evaluator* through *Fairest Evaluator,* Moderated by *Workspace*

Outcome Variable: *Fairest Evaluator*				
	B	SE	*t*	*p*
Social Intelligence	-0.02	0.01	-2.13	$0.02^{**\dagger}$
Workspace	-2.25	1.12	-2.02	$0.02^{**\dagger}$
*Social Intelligence * Workspace*	0.02	0.01	1.90	$0.03^{**\dagger}$
Past Discrimination	0.02	0.06	0.32	0.74
Outcome Variable: *Preferred Evaluator*				
	B	SE	*t*	*p*
Social Intelligence	-0.01	0.01	-1.65	$0.05^{**\dagger}$
Fairest Evaluator	0.43	0.07	6.18	$<0.01^{***\dagger}$
Past Discrimination	0.15	0.05	2.88	$<0.01^{***}$
Indirect Effect of *Social Intelligence* on *Preferred Evaluator* through *Fairest Evaluator* for each level of *Workspace*				
	B	SE	Lower CI	Upper CI
Shared	-0.01	0.004	-0.0160	-0.0007
Remote	<0.01	0.004	-0.0062	0.0091
Index of moderated mediation	Index	SE	Lower	Upper
Workspace	0.0096	0.006	-0.0003	0.0208

We test a moderated mediation model using the simultaneous OLS regression method and Model 7 outlined in Hayes (2018). This approach allows an examination of whether a) there is an indirect effect of the independent variable on the dependent variable through a mediator, and b) whether such an indirect effect is conditional on a moderating variable. In this model, we test whether *Social Intelligence* affects *Preferred Evaluator* through *Fairest Evaluator* and whether this indirect effect is conditional on *Workspace.* Tests are conducted using 5,000 bootstrap samples and a 95% confidence interval.

† One-tailed test consistent with directional prediction.

Appendix: Tromsø Social Intelligence Scale (Silvera, Martinussen, and Dahl 2001)

All responses were recorded on a seven-point Likert scale from “strongly agree” to “strongly disagree.” Items 2, 4, 5, 8, 11, 12, 13, 15, 16, 20, 21 are reverse coded.

1. I can predict other peoples’ behavior.
2. I often feel that it is difficult to understand others’ choices.
3. I know how my actions will make others feel.
4. I often feel uncertain around new people who I don’t know.
5. People often surprise me with the things they do.
6. I understand other peoples’ feelings.
7. I fit in easily in social situations.
8. Other people become angry with me without me being able to explain why.
9. I understand others’ wishes.
10. I am good at entering new situations and meeting people for the first time.
11. It seems as though people are often angry or irritated with me when I say what I think.
12. I have a hard time getting along with other people.
13. I find people unpredictable.
14. I can often understand what others are trying to accomplish without the need for them to say anything.
15. It takes a long time for me to get to know others well.
16. I have often hurt others without realizing it.
17. I can predict how others will react to my behavior.
18. I am good at getting on good terms with new people.
19. I can often understand what others really mean through their expression, body language, etc.
20. I frequently have problems finding good conversation topics.
21. I am often surprised by others’ reactions to what I do.