Held to a Higher Standard? Fairness Perceptions of Artificial Intelligence in HR Decision Making

Sandra L. Fisher
Münster School of Business
Münster University of Applied Sciences

Garett N. Howardson*
U.S. Army Research Institute for the Behavioral and Social Sciences
Team Leader, Predictive Analytics and Modeling Research Unit (PAMU)

*The views expressed in this chapter are those of the author and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.

Suggested Citation:

Fisher, S. A., & Howardson, G. N. (2022). Fairness of artificial intelligence in human resources –held to a higher standard? In S. Strohmeier (Ed.), *Handbook of research on artificial intelligence in human resource management* (pp. 303-322). Edward Elgar Publishing.

Abstract

Implementation of decision-making tools based on artificial intelligence (AI) has introduced a new set of concerns related to fairness in organizations. It is clear that human decision makers demonstrate biases and make errors in decisions related to hiring, compensation, discipline, and other human resource (HR) topics. Regardless, applicants and employees are expressing high levels of concern about the fair and appropriate use of AI in HR decision making. This chapter reviews research and theoretical perspectives about fairness in HR, including organizational justice and social exchange. We then compare these perspectives with research from the AI and information systems literatures to examine different viewpoints on fairness and why the fairness of decisions made by humans and AI systems might be judged differently. We highlight a potential paradox in thinking about automation and intentionality of AI decision tools, and discuss the potential impact on the use of these systems within HR.

Introduction

The fairness of decisions made about human resource management (HRM) issues in organizations has been a concern of both research and practice for decades (Colquitt et al., 2001; Gilliland, 1993; McCarthy et al., 2017). While people generally have an intuitive sense of what is fair and what is unfair, these terms are defined in different ways in different academic traditions. We start from the point of considering fairness as an evaluation of the appropriateness of different allocation decisions (e.g., Leventhal, 1980) or how rewards and resources are distributed. Research has consistently demonstrated that such allocation decisions in the human resources realm (e.g., hiring, compensation, performance evaluation, discipline) that are perceived as unfair have negative effects for both individuals and organizations. Individual employees experience lower levels of satisfaction with their jobs, lower commitment to the organization, and reduced trust when they perceive unfair decisions have been made about themselves or others (Colquitt et al., 2001; McCarthy et al., 2017). Organizations then may see decreases in their reputation as an employer, productivity losses among employees, and higher employee turnover.

Many organizations see the use of decision tools that use artificial intelligence (AI) and the corresponding algorithms as a way to improve their HR decision-making processes and outcomes (Gonzalez et al., 2019). Following Strohmeier (2020), we define algorithmic decision making in the context of HR as "employing computer-based HR decision algorithms for drawing or supporting HR decisions" (p. 54). Knowing that humans have biases and make errors in HR decisions ranging from selection to discipline to negotiations, AI and the related algorithms are viewed as a way to make consistent decisions based on concrete data. Such decisions can be made more quickly, freeing HR staff and managers from time-consuming tasks such as reviewing job application materials (Gonzalez et al., 2019). These benefits notwithstanding, new technologies often have unintended consequences on organizational processes (Orlikowski, 1992), as may be the case with AI and algorithmic decision making (Kellogg et al., 2020).

Evidence is beginning to accumulate that AI-based decision making can also be perceived as unfair. Biases that provide a distinct advantage or disadvantage to members of certain groups have been identified in some selection algorithms, such as the one reported by Amazon that evaluated female applicants as less qualified. These biases are also present in decision making algorithms affecting daily life outside of the employment context (e.g., the Apple credit card). Although there are clearly concerns with AI-based decision making, it appears that people may actually be holding these tools to a higher standard than human decision making (Dietvorst et al., 2015; Zerilli et al, 2019). Consider the example of self-driving cars. Human operated motor vehicles are associated with over 1 million deaths per year globally (World Health Organization, 2020). In 2017, there were over 35,000 deaths from motor vehicle accidents in the United States and over 25,000 deaths in the European Union (Eurostat, 2017; National Highway Traffic Safety Administration, n.d.). However, a single 2018 death associated with a test of a self-driving car in Arizona was considered as a major setback for implementation of the technology. This seems to be an example of how "the errors that we tolerate in humans become less tolerable when machines make them" (Dietvorst et al., 2015, p. 115), a phenomenon that we argue is also observable in HR decision making.

In this chapter, we explore why people might react differently to HR decisions made by humans and those made by AI supported tools. We review theories and models used to evaluate and predict the fairness of HR decisions and compare those to how fairness has been addressed in the AI literature. We identify decision maker intentionality as a critical feature of fairness from the social exchange perspective and examine how intentionality, autonomy, and social context play important roles in judgement of fairness.

Fairness in Organizational Decision Making

There have been decades of research on fairness in organizational decision making in the fields of industrial-organizational psychology, human resource management, and organizational behavior. Much of this research examines fairness perceptions by individuals (such as job applicants or employees) about decisions that various organizational representatives have made about them. This research has been described extensively elsewhere (e.g., Colquitt et al. 2001; Colquitt et al., 2013; McCarthy et al., 2017; Rupp et al., 2014) and in the context of AI decision-making (Robert et al., 2020); therefore, we will not review it in detail but rather identify some key themes and theoretical perspectives present in this research stream.

General Fairness Theories

Organizational Justice. One approach to the study of fairness in organizations is through the lens of organizational justice. This line of research has focused on different types of justice, or the rules people use to decide if a decision or event is fair (Colquitt et al., 2001; Colquitt et al., 2013; Gilliland, 1993). The literature has largely converged on four different types of justice: distributive, procedural, interpersonal, and informational. Distributive justice focuses on the outcomes of decisions about valued resources in organizations such as selection, pay, and promotion. Procedural justice focuses more on the process or procedures used to reach the decision. Informational justice narrows in on the information that people receive about decisions that are made, and the extent to which this information can explain "why procedures were used in a certain way or why outcomes were distributed in a certain fashion" (Colquitt et al., 2001, p. 427). Interpersonal justice addresses the extent to which the actors involved in making a decision treated the focal person with respect. These four types of justice are empirically distinct and show different patterns of relationships with important outcome variables in the literature (Colquitt et al., 2001; Colquitt et al., 2013). This literature has examined HR decision making in organizations broadly rather than focusing on one type of decision (in contrast, below we discuss literature that focuses on selection decisions). For example, the Colquitt et al. (2001) meta-analysis included papers that dealt with individual justice perceptions in HR topics including employee discipline, performance appraisal, layoffs, promotions, pay, negotiations, drug screening, and electronic control systems.

This model of organizational justice is based on a set of criteria set forth by Leventhal (1980). For decisions to be considered fair, they should be based on decision rules that have the following characteristics: (a) applied consistently over time and across people, (b) free from bias, meaning that the decision is made without self-interest or preconceptions about people, (c) based on accurate information and informed opinion, (d) include a mechanism for correcting errors or oversights in the decision process, (e) opinions of various groups affected by the decision are represented and considered, and (f) conform to standards of ethics or morality "accepted by that individual" (Leventhal, 1980, p. 33). While these are intended to be general rules that apply across people, the individual judgments regarding any one decision could vary based on individual characteristics (e.g., personal standards of ethics). The HR research based on these rules has focused on systematic efforts to improve fairness perceptions across people, such as making decisions more consistent and gathering accurate information for making decisions.

The AI literature (e.g., Robert et al. 2020) has been addressing some of these concepts, with some studies explicitly using the terminology of organizational justice theory and others more implicitly. The focus has been on distributive fairness outcomes, with some attention to procedural justice and less for interpersonal and informational. For example, in their

experimental study on perceptions of fairness in workplace decision making, Ötting and Maier's (2018) vignette-based study found that the decision maker (human team leader, humanoid robot, or intelligent computer system) did not matter in terms of perceptions of procedural justice. The study, however, did not investigate interpersonal or informational justice.

The justice dimensions do still seem to apply to a decision-making context in which AI plays a significant role. We can question the relevance of interpersonal justice with an AI-based decision agent, as this type of justice generally refers to the interactions between two people. Early research on this topic as reviewed by Glikson and Woolley (2020) suggests that AI agents such as chatbots can meet some of the criteria of interpersonal justice, such as treating people with "politeness, dignity, and respect" when "executing procedures or determining outcomes" (Colquitt et al., 2001, p. 427), particularly when they take on human attributes (Kim & Duhachek, 2020).

Social Exchange Perspective. Another view of organizational justice is based on the social exchange perspective (Colquitt et al., 2013; Rupp et al., 2014). Instead of focusing on the policies, procedures, or rules of how a decision is made, this perspective highlights the actor(s) involved in making the decision. A key question becomes, who is responsible for an act of injustice? The resulting attitudes and behavioral reactions are then targeted toward the party at fault, typically the supervisor or the organization as a whole (Rupp et al., 2014). If the decision maker acts in a just manner, then social exchange processes suggest that the individual will respond with positive reactions such as organizational citizenship behaviors, commitment, and trust. The focus of the responsible party and the individual reactions should match, such that supervisor-focused justice results in supervisor-oriented outcomes, examples of which are leader-member exchange (LMX), and organization-focused justice results in organization-oriented outcomes such as commitment (Colquitt et al, 2013). For social exchange processes to truly develop, a relatively long-term time frame is required. As such, the social exchange perspective may be more relevant to justice perceptions of employees that can unfold over multiple decisions in performance management, feedback, and labor relations processes rather than the short-term viewpoint of the typical applicant in a selection scenario (Colquitt et al., 2013; Rupp et al., 2014).

Fairness theory (Folger & Cropanzano, 2001) similarly looks at decision fairness in the context of attribution to different actors, with interest in the process of making accountability judgments or assigning blame. This approach to fairness in HR decisions examines if there is someone who should be held accountable for the unfair decision. Fairness theory requires that there is a negative event that could be considered unfair, that the actions leading to the negative event were discretionary (i.e., the actions were intentional), and it violates some kind of moral code. The person who was negatively affected by the event would then use counterfactual thinking to decide when to assign blame. For example, they might imagine what alternative actions the decision maker could have taken. If there were no other options that seem feasible, the actions may be considered not under control of the actor, and therefore this person cannot be blamed.

Conclusions on General Fairness Theories. The organizational fairness approaches discussed above are complementary. The types of justice approach (Gilliland, 1993; Colquitt et al., 2001) focuses on the "what/how" component of justice perceptions and offers specific rules through which organizations may improve fairness perceptions about their HR processes. The social exchange and fairness theory approaches (Folger & Cropanzano, 2001; Rupp et al., 2014) focus more on the "who/when" component of fairness judgments, thinking about who is responsible for an unfair decision or when they may be absolved from blame, such as when they could not have acted differently. These social exchange and fairness theory perspectives strive to account for situations where no particular individual warrants blame, such as when circumstances are beyond an individual's control. In such cases, holding an

individual accountable for events that were beyond their control would, regardless of the specific rule in question, be inappropriate or unfair. Fairness, in other words, may be more about the general perception that the application of a specific rule is appropriate given a specific set of circumstances (Colquitt & Rodell, 2015; Colquitt & Zipay, 2015; Carmody, 2015). In some circumstances, it may even be appropriate to hold no one accountable at all in that, "If no one is to blame, there is no social injustice" (Folger & Cropanzano, 2001, p. 1).

The importance of blame, accountability, and appropriateness in social exchange and fairness theory perspectives raises several important questions about AI and fairness. Can and should we blame AI-decision tools? Can and should they be held accountable? Under what conditions, if any at all, might we hold AI-decision tools accountable? While we cannot claim to offer definitive answers here, we do explore such questions a bit further in the final sections of this chapter. Before doing so, however, it is helpful to contextualize the general fairness perspectives above within some specific HR applications where questions about AI and fairness are beginning to emerge.

Application of General Fairness Theories to HR Contexts

Applicant Reactions to Selection. A more concrete stream of fairness research focuses specifically on applicant reactions to the hiring process and decision (Gilliland, 1993; Hausknecht et al., 2004; Konradt et al., 2017; McCarthy et al., 2017). This research frequently applies the four types of organizational justice described above (Gilliland, 1993; Colquitt et al, 2001), examining the antecedents and consequences of applicant reactions. Reviews of the literature, both meta-analytic and more qualitative reviews, have found general support for Gilliland's (1993) framework of the importance of distributive and procedural justice in applicant reactions to selection processes (e.g., Hausknecht et al., 2004; McCarthy et al., 2017). Studies have consistently found that job relevant selection procedures, provision of explanations, and availability of feedback were particularly important to fairness perceptions. Konradt et al. (2017) demonstrated that the importance of different fairness predictors can differ at different points in the selection process, such that interpersonal treatment was more important for procedural fairness perceptions in their study early in the process while availability of explanations was more important later in the process.

McCarthy et al. (2017) noted that the continued application of emerging technologies into the selection process has efficiency and cost advantages for organizations but mixed results from applicants in terms of their reactions. Some technology-based features such as easy to use online applications are viewed positively, but those that degrade the quality of the face-to-face experience or are perceived as invading privacy (e.g., use of social media) have been viewed more negatively. Blacksmith et al. (2016) suggested that negative reactions to online interviews may be due to perceptions of unfairness and frustration, because applicants have less ability to manage impressions in online interviews, which are viewed as less personal than face-to-face. Having interviews scored with AI could exacerbate this effect, as applicants could believe they have even less influence over the process (Langer et al., 2019). Perceptions of the AI system being responsible for unfairness could lead to some of the negative behaviors described in the literature, such as forming a negative opinion of the organization and communicating that viewpoint to family and friends (Hausknecht et al., 2004).

Decision Maker Reactions to Selection Processes. Another stream of research has examined the perceptions of the decision makers in the selection process regarding their preferences for expert versus algorithmic decision making (e.g., Diab et al, 2011; Highhouse, 2008; Kuncel et al., 2013; Nolan et al., 2016). In general, standardized selection processes result in more valid organizational decisions than unstandardized (e.g., unstructured interview) processes (Gatewood, Feild, & Barrick, 2011).

However, research and anecdotal evidence routinely show that hiring managers prefer non-standardized or clinical methods for making hiring decisions. They believe that expert decision makers can make better decisions than impersonal algorithms or equations and believe that their own value as experts in the process is reduced when standardized methods are used (Highhouse, 2008; Nolan et al., 2019). These findings have not been specifically interpreted in line with the justice and fairness literatures, but we argue could be extended to suggest decision makers believe it is unfair to hiring managers to be "replaced" with standardized decision-making tools in the selection process. It is also important to note that this research has not been conducted using AI-based decision tools but rather with reference to a mechanical combination of data through a generic algorithm or formula (Kuncel et al., 2013) or a computer program (Nolan, et al., 2016).

Legal Perspectives on HR Decision Making. Within HR and I-O psychology, research and practice on fairness in HR decision making also must consider the legal perspective on fairness. Discriminatory employment decisions, for example, are one example of decisions that many would consider unfair. Within the United States, this is defined by the *Uniform Guidelines on Employee Selection Procedures*, with detailed definitions of adverse impact, or in Germany as defined by The General Equal Treatment Act (Allgemeines Gleichbehandlungsgesetz). Group-level cases of discrimination can be defined statistically, but individual cases cannot. Thus, we highlight that "Fairness is a social rather than a psychometric concept. Its definition depends on what one considers to be fair." (*Principles*, 2018, p. 22). In this chapter, we focus on fairness as defined societally and refer the interested reader to Yankov et al. (2020) who provide some practical suggestions on how to ensure that AI selection methods are responsive to legal requirements as described in the *Principles*.

Summary. The organizational justice and social exchange perspectives offer two broad theories of fairness from which others have developed (e.g., Fairness Theory). They complement each other such that organizational justice establishes the set of desired rules governing fairness judgements and social exchange specifies under what conditions those rules may be acceptably broken, or when a particular set of rules may be seen as fair in one situation but not another. Organizational justice suitably describes circumstances for fairness with respect to job applicants. Social exchange adds nuance that is necessary to also explain fairness judgements with respect to manager selection decision making (where we ask, fair to whom?) and general legal issues in HR Fairness in organizational settings is a complex issue that has been studied for almost a century. From these efforts arose rich, multifocal theories that offer different fairness definitions and situational contingencies for determining what is fair. While the study of AI is relatively new compared to the organizational sciences, many similar themes are already emerging with respect to fairness theories and definitions. We offer below a brief overview of this emerging research.

Fairness from the AI Perspective

Concerns about the intersection between technology, fairness, and legal use of tests are not new in HR, particularly when viewed through a social perspective (Jonas, 1982). Group administered intelligence tests, for example, constitute one of the most notable psychological technologies of the 20th century (Danziger, 1997), using the term technology here in the broader sense rather than representing a specifically digital tool. Once thought to be objective and free from human bias, and therefore ideal for selection and placement, such tests are now recognized as social products subject to the same fairness concerns as any other social process (AERA, 2014; Markus & Borsboom, 2013). Fairness, in other words, was once considered a property of the selection test itself rather than the modern view that fairness is a property of the test's human developers.

To that extent that fallible human discretion is involved its development and implementation, no psychological test can be universally fair (Markus & Borsboom, 2013). Instead, test developers must often sacrifice some aspects of fairness in favor of others. For example, one conceptualization of fairness in employee selection testing is that applicant scores on an intelligence test collected at one point in time should be related to actual performance measured at a future date. Maximizing this view of fairness (most often called the test's validity), however, often selects non-minority candidates at disproportionately high rates, which creates a different type of unfairness known as adverse impact (Pyburn, et al., 2008). Just as human preferences and judgments may conflict with each other in different situation (e.g., Locke, et al., 1994), so too do different definitions of test fairness.

Although one might readily dismiss concerns about different definitions of fairness as unique to psychological and organizational settings, similar concerns have been noted with respect to machine learning and AI decision-making technologies (D'Amour, et al., 2020; Green & Viljoen, 2020; Harrison, et al., 2020; Robert et al., 2020). Historically, "algorithmic fairness is grounded in objectivity and neutrality. Fairness is treated as an objective concept, one that can be articulated and pursued without explicit normative commitments" (Green & Viljoen, 2020, p. 6). For instance, many mathematical definitions of AI fairness rely on what is most often called a confusion matrix, which is a two by two contingency table such as the one shown in Table 1. In such a table, the number of predicted positive and negative outcomes are contrasted against the number of actual positive and negative outcomes, most often with the objective of maximizing the number of true positives and true negatives. A common definition of fairness drawing from the confusion matrix is to maximize accuracy, which is given as the ratio of true positives and true negatives to all observations, given formulaically as $(TP + TN)/(TP + TN + FN + FP)$. For example, in the selection context, we would want to maximize the percentage of correct employment decisions (hired and performed well, not hired and would have performed poorly) relative to all applications. This mathematical approach to assessing fairness does allow some evaluation of the fairness of employment decisions from the distributive justice point of view at the group level (Robert et al., 2020).

Table 1. Typical Confusion Matrix

	Predicted Bad Performer	**Predicted Good Performer**
Actual Good Performer	False Negatives (FN)	True Positives (TP)
Actual Bad Performer	True Negatives (TN)	False Positives (FP)

Although accuracy may initially seem like an adequate fairness metric, its definition through the confusion matrix requires the transformation of continuous variables (the test scores) into a categorical variable (pass/fail, select/reject). In doing so, the actual relationship between the two variables is oversimplified, which obscures meaningful fairness differences in the data, as shown in the figure. This tendency to oversimplify the situation in order to create a generalizable, situation-invariant metric is known as the universalism component of formal algorithmic definitions of fairness (see Green & Viljoen, 2020). There is evidence that people observing decisions made with such models perceive that maximizing accuracy may not be perceived as the fairest solution. For example, Harrison et al. (2020) found that lay individuals prefer machine learning models with equal false-positive rates across demographic groups to models with equal accuracy rates across groups. Fair, in other words, may mean equal access to opportunities across groups more so than equal distribution of outcomes between groups (see also Binns, 2018; Hardt, Price, Srebro, 2016).

Thus, we see that similar to the movement in HR to consider individual perceptions of fairness, the discipline of AI is starting to shift from seeking abstract and context-invariant views of fairness to seeking more concrete and context-sensitive views of fairness (Hutchinson & Mitchell, 2020). Whereas the very concept of an employment test's fairness

was once defined mathematically, contemporary employment testing views acknowledge that the foundational aspect of a test's validity is the test's intended use and the scores' intended interpretation (AERA, 2014; Kane, 2013; Landy, 1986; Markus & Borsboom, 2013). A more general HR shift amounts to moving away from purely conceptualizing different types

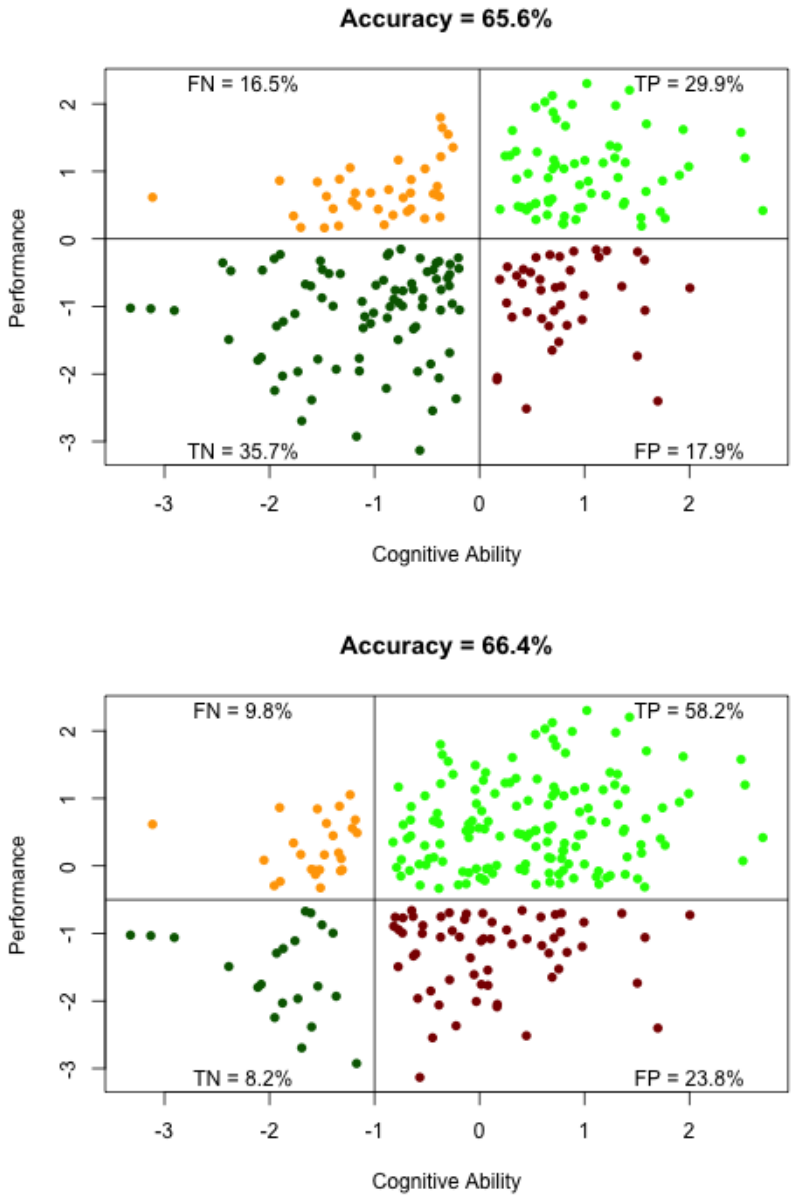

Figure 1. Confusion Matrix Example

(e.g., distributional, procedural) of fairness that should be simultaneously maximized (i.e., the organizational justice view; Colquitt et al., 2001) to understanding the unique combination of individual and situational factors contributing to whether an action is fair, and for whom the consequences of the action are more or less fair (i.e., social exchange view; Rupp et al., 2014). Within the context of AI, this shift amounts to moving away from formal, mathematical definitions of fairness to understanding the particular situations within which certain fairness tradeoffs are acceptable (e.g., increased false positive potential that also increases diversity; Allen, et al., 2006; Green & Viljoen, 2020; Harrison et al., 2020).

Appropriateness of AI-Based Decision Making in HR

As noted above, it can be useful to define fairness broadly as the general appropriateness of HR decisions (Colquitt & Zipay, 2015) rather than equating justice and fairness. Appropriateness takes into account a range of judgements beyond whether or not specific justice rules have been followed. For example, research on stakeholder opinions of AI-based decision-making tools suggests that there is general recognition that they can be useful for making decisions quickly and consistently. However, there is also a widely reported perception of algorithm aversion (Harrison et al., 2020; Logg et al., 2019) that could lead to employees feeling that AI-based tools are inappropriate for use in HR decision making. Such aversions would likely lead to the negative reactions discussed in the justice and fairness literatures such as reduced intention to accept a job offer, lower job satisfaction and organizational commitment, or higher turnover intentions. Algorithm aversion is a broadly negative reaction to algorithms and a desire to avoid using them (Dietvorst et al., 2015). This aversion is aptly described by Edwards and Veale (2017) who noted, "The public has only relatively recently become aware of the ways in which their fortunes may be governed by systems they do not understand, and feel they cannot control; and they do not like it" (p. 19). Algorithmic aversion is partially due to individual characteristics such as the standards for ethics and morality discussed by Leventhal (1980) and affective reactions to specific technologies, such as the construct of creepiness as a reaction to interaction with new technologies (Langer & König, 2018). Below we review several situational and technological characteristics that are outside the specific scope of justice and fairness theories but are related to perceptions of the overall appropriateness of algorithmic decision making in HR.

Reliability and Accuracy. Research on algorithm aversion suggests that people are highly sensitive to the reliability and accuracy of algorithmic decision tools, such that they want to avoid algorithms after they observe the tool making an error (Dietvorst et al., 2015; Glikson & Woolley, 2020). Automated decision/action systems need to be very reliable before they can be highly automated (Parasuraman et al., 2000). Strohmeier and Piazza (2015) argue that AI technology should result in not just good decision quality, but an improvement over prior quality levels in order to justify use of AI in HR decision making.

Opacity. The extent to which people understand (or do not understand) AI-based algorithms is related to their acceptance of decisions made by such tools (Edwards & Veale, 2017). Some of this opacity is due to the technical knowledge required to understand how machine learning algorithms are trained and applied (Kellogg, et al., 2020). Opacity is also the result of companies intentionally obscuring how the algorithms work for purposes of maximizing value and competitive advantage. For example, the vendors that market selection systems based on AI decision tools keep their tools proprietary as part of their business model. Transparency has been found to increase trust across a variety of situations (Glikson & Woolley, 2020). In fact, the movement toward algorithmic auditing is an effort to address opacity and increase trust in AI decisions by offering assurance from an objective third party that the algorithm is valid (Robert et al., 2020).

Control. Managers may lose power in decision making, or be completely removed from a decision process, when AI decision tools are implemented (Kellogg et al., 2020). This concern about control is aligned with the finding in the HR literature that hiring managers prefer clinical or expertise-based decision making over mechanical combination of information, as discussed above. Managers and employees may also make attributions about AI-based decision systems that impact their overall acceptance of such systems. Individuals make attributions about the reasons for why organizations are implementing new HR systems, either commitment-focused attributions of service quality and employee well-being, or control-focused attributions of cost reduction and employee exploitation (Nishii et al., 2008). If managers view that the AI decision system is there to help them improve decision-making quality and perhaps reduce their workload, then they are likely to view it positively and be more accepting of it. In contrast, if they perceive that the system is there to control them for

the benefit of the organization, then they are likely to be less accepting of the system. From the applicant perspective, initial research suggests that interviewees may react negatively to highly automated interviews that incorporate AI because they perceive a lack of control over the situation, with less ability to influence the interviewer through social cues (Langer et al., 2019).

Task-technology Fit. The concept of task-technology fit (Goodhue & Thompson, 1995) is important for the acceptance of AI-based decision agents (Strohmeier & Piazza, 2015). Research has generally shown that people are less likely to accept the decisions made if they perceive that there is a mismatch between the technology and the task. Machines or AI agents are considered more appropriate for technical or mathematical tasks, while humans are more appropriate for social tasks (Glikson & Woolley, 2020). Kim and Duhacheck (2020) found in a series of laboratory studies that people were more accepting of (and persuaded by) AI agents when the system emphasized more concrete suggestions for *how* to perform a task (i.e., how to exercise or how to make soy milk at home) than when the system emphasized more abstract suggestions for *why* someone should perform a task (i.e., why exercise or why make soy milk at home). Thus, there was a match between the persuasive message offered by the AI agent and the perceived appropriateness of that message for an AI agent, resulting in greater intentions to engage in the behavior. However, when the AI was described with more human-like characteristics such as ability to learn or consciousness, this matching effect was reduced and AI agents were able to successfully persuade participants with the more abstract suggestions.

In line with Glikson and Woolley (2020), one explanation for these findings is that providing abstract arguments about *why* someone ought to do something is a purely social task that should be reserved for human (or human-like) agents. Such findings are important to consider in that one major advantage of AI often discussed is autonomous decision making requiring little to no human oversight (Parasuraman et al., 2000; Roberts et al., 2020). As noted above, however, people may consider autonomy a fundamentally human characteristic (e.g., Ryan & Deci, 2000) and reserve such autonomous decision making only for human beings, particularly in social contexts or when stating why someone should make a particular decision or take an action.

One such example in an HR context is an employment interview where applicants are asked a series of questions or to perform a series of tasks that will be used to evaluate their employment qualifications. Many interview questions or tasks are inherently social in nature, for example, how the interviewee would motivate others. Further, the process of interviewing has historically been a social process involving interpersonal interactions and mutual social influence. These factors suggest that interviewees are likely to expect some sort of social interaction during the interview, and that a human evaluator would be involved in determining if the candidate should be selected. In such situations, a fully autonomous AI would likely engender adverse applicant reactions or decision maker reactions by creating task-technology misfit between the social task (employment interview) and the AI's more technical nature (Langer et al., 2019).

User Expertise. Another individual characteristic than can affect perceptions of appropriateness for using algorithms is the level of user expertise in the relevant content domain. Logg et al. (2019) conducted a series of experiments comparing the extent to which people valued advice from an algorithm compared to advice from other people, consistently finding that for low stakes decisions (e.g., determining the weight of someone in a photo, predicting the popularity of a song) that people preferred to take advisement from an algorithm. This effect held until participants were asked to make decisions about a topic area in which they had expertise, when they found that experts discounted both the advice offered by algorithms and other people. Interestingly, in this situation, even the experts had less accurate results in the decision-making task (Logg et al., 2019). This aligns with the research

on managerial decision making for selection (Highhouse, 2018; Nolan et al., 2016), where managers prefer to make their own decisions on hiring rather than accept assistance from algorithms or computerized tools. When people lack expertise or view decisions outside of their own area of expertise, they are more likely to appreciate having an algorithm to assist them (Logg et al., 2019).

Discussion

This chapter has focused on the fairness of HR decisions made by humans compared to the use of AI-based decision tools for the same decisions. We have reviewed literature on justice and fairness in the HR and AI disciplines and examined a wide variety of features relevant to decision making processes that affect perceptions of fairness both for the affected party and the human decision maker. From what we see in practice, it is clearly acknowledged that human decision makers are fallible. They have biases, make errors in judgment, and fail to follow procedures that could enhance perceptions of fairness. And yet, people often prefer the judgments of humans over those of AI-based decision-making tools. It appears that we seem comfortable applying different standards to people and AI when determining what is fair and acceptable in HR decision making.

Fairness in the AI and HR decision making literatures has been approached differently, suggesting that we at a minimum have a *different* standard. In the AI literature, fairness has been defined and operationalized from a mathematical viewpoint, optimizing outcomes given a set of clearly defined assumptions. In the HR literature, fairness is typically rooted in human emotions, reactions, and experiences. There is a focus on how individuals react to their experience and what it means to them, regardless of (and sometimes in spite of) the math. Even when something is mathematically fair, given the stated assumptions, people may perceive a decision or experience as unfair based on their own expectations and individual differences. Creating an algorithm that could maximize the fairness perceptions of the hundreds or thousands of people in the pool for an HR decision is a very high standard indeed. Regardless, the need to broaden definitions of fairness has been recognized in the AI research moving towards more context-sensitive definitions (Green & Viljoen, 2020; Harrison et al., 2020). The concept of algorithmic realism (Green & Viljoen, 2020) encourages computer scientists and engineers to consider the context in which the algorithm will be used to better determine the quality criteria, and even ask if an algorithm is the most appropriate intervention. This is consistent with the notion of task-technology fit (Glikson & Woolley, 2020; Strohmeier & Piazza, 2015), and its importance for the acceptance of AI-based decision agents. As long as applicants and employees believe it is more appropriate for humans to make decisions about hiring, promotion, performance evaluation, and discipline, this viewpoint suggests that AI-based decisions will be considered inferior and unfair even if mathematically the AI produces superior results.

There are some logical reasons for why perhaps we should hold AI to a higher standard (Zerilli et al., 2019), particularly as it reaches higher levels of automation or autonomy. First, autonomous AI decision agents can make decisions about many more people at one time than a human decision maker can. Humans possess relatively narrow or bounded information processing capabilities that limit the impact of any single decision (e.g., Simon, 2000). The bounds on AI information processing capabilities, however, remain an open question. Second, autonomous AI decision agents cannot be held legally responsible or accountable for their decisions. Managers can be disciplined or fired if they consistently make unfair personnel decisions (e.g., who to hire, who to fire). Finally, there are often procedures or committees in place to reduce human bias. Larger group of managers and HR professionals, for example, often review and discuss managerial decisions about whom to

promote or whom should receive a raise. This may not be the case for large numbers of decisions made by autonomous AI decision agents.

Some of this higher standard for AI is likely related to broader societal concerns for the impact of technology on our lives in unknown and unexpected ways (Orben, 2020). Orben calls this phenomenon technology panic, "in which the general population is gripped by intense worry and concern about a certain technology" (2020, p. 1144). One possibility for why we seem to be holding AI to a higher standard at this point in time is availability of information as a result of this intense concern about the impact of AI in general, and on workplace decisions specifically. A Google search of "AI decision making" results in hundreds of thousands of articles about how AI is being used in decision making in a variety of disciplines, with potentially inflammatory headlines such as "Is AI taking over[1]?" There may simply be more information available right now to the average person about concerns regarding AI decision making than unfair human decision making. Press coverage and social media coverage of AI certainly exceeds coverage of other potentially unfair HR decisions such as legal cases or US Supreme Court cases.

Another factor strongly related to appropriateness, and therefore perceived fairness, of AI decision making is opacity. Research has been demonstrating the positive effects on fairness perceptions of applicants receiving an explanation and therefore understanding why a decision was made (McCarthy, et al., 2017). For example, if a person was not selected for a position for which they interviewed, providing some feedback about why they did not get the job offer often helps increase fairness perceptions. Guidelines for ethical practice recognize "that applicants deserve to know the meaning of the information collected about them and how the information was used to make decisions about them" (London & McFarland, 2017, p. 409). Employers do not always offer such feedback. However, it seems plausible that people imagine that they could understand how another person made a hiring decision, and they could receive an explanation from a person if they asked for one, where this is less probable for AI-based decisions. Thus, even if AI-based decisions are more accurate and more consistent, which should be associated with fairness, we may be weighting understandability (or opacity) and explainability over accuracy and consistency (see Langer, this volume, for a detailed discussion of explainability). For example, it is quite possible that nobody in the hiring organization really understands how a decision is reached using AI tools such as automated analysis of video interviews. As noted by Burrell (2016, p. 10), "The workings of machine learning algorithms can escape full understanding and interpretation by humans, even for those with specialized training, even for computer scientists conceptualizing different types" (quoted in Kellogg et al. 2020, p. 372). Quite simply, if organizational representatives (managers, human resource professionals, or IS professionals) do not understand how a decision was made, this will limit the extent to which an explanation can be provided (see Edwards & Veale, 2017) and will make it more difficult for the applicant or employee to understand the decision and rationally evaluate its fairness. With limited potential for opacity, we have to look at other cues for judging fairness.

Fairness, Intentionality, and the Social Context of AI in HR

Early views of fairness in both HR (e.g., organizational justice) and AI (e.g., algorithmic formalism) sought consistent, de-contextualized sets of rules that could be applied in any situation. Such rules, it was assumed, would exist independent of any particular social context, or would be separate from any particular person or set of people. In the context of HR, such views began fading from scientific favor due, in part, to the humanistic shift away

[1] We found this phrase in over 5,000 Google search result items on 8 September, 2020.

from the more mechanical and asocial views of scientific management (e.g., Jaffee, 2001). Perhaps more than in any other area of HR, this shift is evident in employment testing and personnel selection where fairness became defined as a property of the test users' intentions rather than of the test itself (Kane, 2013; Markus & Borsboom, 2013). A test's fairness, in other words, could not be fully separated from the social context within which the test would be used (*Principles,* 2018).

More generally, the idea of socially de-contextualized rules encourages societal views of technology as a fixed property of the physical environment rather than the as the product of social action (Jonas, 1992). While such technological views have generally fallen out of favor in the organizational sciences (e.g., DeSanctis & Poole, 1994; Orlikowski, 1992), such views remain prevalent in society at large and periodically re-enter the realm of scientific study (Orben, 2020). While one could argue that concerns about AI decision making are the most recent emergence of such views, the AI field has only recently begun recognizing the need to consider more socially contextualized views of fairness (Hutchinson & Mitchell, 2020) where an AI and the society that generated it are necessarily intertwined (i.e., algorithmic realism; Green & Viljoen, 2020). Viewing AI as a part of human social processes, however, likely contributes to unrealistic expectations of the technology (Hew, 2014), and hence the higher standard we have described.

A core component of the more social exchange theory-like views of fairness noted above (e.g., Fairness Theory, Folger & Cropanzano, 2001) is the role of an actor's intentions. For example, adverse treatment in selection is determined by the extent to which the organization knowingly and intentionally treats members of a protected class differently than members of the non-protected class. Adverse impact, on the other hand, is not necessarily intentional and thus may not be considered as unfair as adverse treatment. Adverse impact is legally defensible if using a validated selection instrument, but adverse treatment is never legally defensible.

The critical question with respect to AI and fairness then becomes, "is AI capable of intentionality?" Although fully answering such a grand question is well beyond the purview of our efforts here (interested readers may wish to see, e.g., Floridi & Sanders, 2004 and Sun, 2016), there is reason to believe that intentionality-possessing AI is, if at all possible, not likely in the foreseeable future (Hew, 2014). Within HR contexts specifically, we offer the more immediate question, "To whom should we ascribe an AI's intentions when making HR decisions, and how should we determine whether those decisions are fair?" Fortunately, we argue that theoretical foundations to answer such questions in HR have already been established by the complementary streams of organizational justice (e.g., Colquitt et al., 2001) and social exchange (e.g., Rupp et al., 2014) theories described above.

A good first step for the design and use of fair AI decision tools in HR is to ensure that the decision rules presented in organizational justice theory are followed. Some attention is already being paid to this as noted by Robert et al. (2020), with 72% of the papers in their review of the AI fairness literature addressing distributive justice, 48% addressing procedural justice, and 24% addressing interactional justice (the umbrella term for informational and interpersonal justice), even if they did not address these concepts explicitly. Thus, we could imagine that when a company is designing an AI decision-making tool such as a chatbot in a selection pre-screening process, they could design it to follow specific procedural rules in a consistent fashion, provide information about how the employment screening decisions are made, and respond to applicant requests with respect. For example, the chatbot could consistently ask each applicant about their prior work experience, tell them that this is an important part of the screening process, and respectfully inform those who do not meet the criterion that their application will not move forward in the process.

In contrast, the social exchange perspective on fairness does not appear to have been applied to AI decision tools. Social exchange helps us identify when it is acceptable to break

the justice rules. These judgements largely depend on the decision maker's intentions. For example, if humans were making employment screening decisions, they might decide to ignore the requirement for prior job experience if the applicant has other desirable characteristics such as belonging to an underrepresented demographic group or having graduated from the same university as the decision maker. The decision to keep the member of the underrepresented group in the pool could be viewed positively if the intention is to bring greater diversity into the work unit. The decision based on university affiliation could be viewed more negatively if the intention is to provide help to a friend's child who does not have the requisite job skills. But if AI does not have intentionality, if it is "designed and programmed by humans to serve" (Kim & Duhachek, 2020, p. 363), then how would it break the stated justice rules? One example of breaking justice rules is when ML algorithms, such as the one developed by Amazon, include decision rules biased against women or other underrepresented demographic groups. Because of opacity, an observer may judge this feature to be the fault of the AI rather than the fault of the human programmers or designers.

A different situation occurs when AI decision tools are advisory to the human decision maker. In this case, the human decision maker is still the relevant target for the fairness judgement and things would proceed as described by the social exchange theories. However, as AI agents become more autonomous, this will become correspondingly more difficult. When an employee is rejected for a promotion because of the results of an AI-based analysis of their recorded interview, we do not yet know how that employee would respond. One alternative is to assign blame to the manager or HR staff person who designed the assessment process. Counterfactual thinking could include thoughts that they should not be using these unproven and potentially biased computer programs, and this decision is unfair because if the assessment process had proceeded normally (with human assessors) then they would have gotten the promotion. Alternatively, they could blame the organization as a whole, as it developed or purchased the AI agent. Regardless of the level of autonomy, it seems inappropriate to blame the AI analysis tool because it has no intentionality. More research is needed in this area using samples of working adults using real AI tools to build on the foundation of early research in the HR field that has largely used vignette and simulation studies.

In conclusion, we see a potential paradox between attempting to enhance trust and potentially adoption of AI based decision agents by giving them human characteristics and the limited evidence about task-technology fit and social exchange expectations for AI regarding fairness. Kim and Duhachek (2020) found that people were more accepting of abstract recommendation messages by AI when they believed it had more human characteristics such as ability to learn. This suggests that task-technology fit can be manipulated with how the AI is described to users. However, we do not know if these findings would extend to situations of HR decision making. There is much research to be done in this area. For example, Robert et al. (2020) raised the question of the potential effectiveness of adding expression of emotion or regret in AI decision tools. Portraying AI decision tools as human may be contributing to the higher standards for AI that cannot be met, as "high expectations of anthropomorphic characters are designed to fail" (Gilkson & Woolley, 2020, p. 645). It would be unwise to design AI tools for HR such that they enter into the "uncanny valley" that increases perceptions of creepiness (Caballar, 2019) and rejection of the technology. However, at least with AI in its current form and in the foreseeable future (Hew, 2014), AI agents will not have the intentionality needed to be truly held accountable for their actions for quite some time. Even if they do reach the stage of having intentionality, it may not be possible within current social and legal frameworks to hold them accountable (Robert et al., 2020, Zerilli et al., 2019), which may have negative impacts on fairness perceptions. More work is needed to determine legal accountability for HR decisions made on the basis of AI tools.

AI-based decision tools in HR are being implemented for operational efficiencies and often with the promise of reduced bias and enhanced fairness. Evidence suggests there are still many challenges in moving forward with the use of such technologies and in facilitating perceptions of fairness on the part of people about whom decisions are being made, with AI tools currently being held to a higher standard than human decision makers. For practice and future research in this area, a definition of fairness as appropriateness that integrates aspects of organizational justice theory and social exchange theory will likely help level the playing field.

References

AERA (2014), *The standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Allen, C., Wallach, W., and Smit, I. (2006), Why machine ethics?. *IEEE Intelligent Systems, 21(4*), 12-17.

Binns, R. (2018), Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research, 81*, 149-159.

Blacksmith, N., Willford, J. C., and Behrend, T. S. (2016), Technology in the employment interview: A meta-analysis and future research agenda. *Personnel Assessment and Decisions, 2(1*), 12-20. DOI: 10.25035/pad.2016.002.

Caballar, R. D. (2019). What is the uncanny valley? *IEEE Spectrum*. Retrieved from https://spectrum.ieee.org/automaton/robotics/humanoids/what-is-the-uncanny-valley.

Carmody, C. (2015), Fairness as appropriateness: Some reflections on procedural fairness in WTO law. In A. Savarian & F. Fontanelli (eds.), *Procedural Fairness in International Courts and Tribunals*. British Institute for International and Comparative Law.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., and Ng, K. Y. (2001), Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425-445.

Colquitt, J. A., and Rodell, J. B. (2015), Measuring justice and fairness. In R. Cropanzano and M.L. Ambrose (Eds.) *The Oxford Handbook of Justice in the Workplace* (pp. 187-202). Oxford University Press.

Colquitt, J. A., Scott, B. A., Rodell, J. B., Long, D. M., Zapata, C. P., Conlon, D. E., and Wesson, M. J. (2013), Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. *Journal of Applied Psychology*, *98*(2), 199-236.

Colquitt, J. A., and Zipay, K. P. (2015), Justice, fairness, and employee reactions. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 75-99.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. (2020, January), Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 525-534).

Danziger, K. (1997), *Naming the mind: How psychology found its language*. Thousand Oaks, CA: Sage.

DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science, 5*(2), 121-147.

Diab, D. L., Pui, S., Yankelevich, M., and Highhouse, S. (2011), Lay perceptions of selection decision aids in U.S. and non-U.S. samples. *International Journal of Selection and Assessment, 19*(2), 209–216.

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.

Edwards, L., and Veale, M. (2017), Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law & Technology Review, 16(1*), 18-84.

Eurostat (2017). Persons killed in road accidents by type of vehicle (CARE data). Retrieved from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=tran_sf_roadve&lang=en

Floridi, L. & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machine, 14*, 349-379.

Folger, R., and Cropanzano, R. (2001), Fairness theory: Justice as accountability. In J. Greenberg & R. Cropanzano (Eds.), *Advances in Organizational Justice*, 1-55, Stanford University Press.

Gatewood, R. D., Field, H. S., & Barrick, M. (2011). *Human resource selection: Seventh edition.* Mason, OH: Cengage.

Gilliland, S. W. (1993), The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18(4*), 694-734.

Glikson, E., and Woolley, A. W. (2020), Human trust in Artificial Intelligence: Review of empirical research. *Academy of Management Annals, 14(2),* 627-660, https://doi.org/10.5465/annals.2018.0057.

Goodhue, D. L., and Thompson, R. L. (1995), Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-236.

Green, B., and Viljoen, S. (2020, January), Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 19-31).

Gonzalez, M.F., Capman, J.F., Oswald, F.L., Theys, E.R., and Tomczak, D.L. (2019), "Where's the I-O?" Artificial Intelligence and Machine Learning in Talent Management Systems, *Personnel Assessment and Decisions*, 5(3), Article 5. DOI: 10.25035/pad.2019.03.005

Hardt, M., Price, E., and Srebro, N. (2016), Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315-3323).

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., and Ur, B. (2020, January), An empirical study on the perceived fairness of realistic, imperfect machine learning models.

In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 392-402).

Hausknecht, J. P., Day, D. V., and Thomas, S. C. (2004), Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*(3), 639-683.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*, 197-206.

Highhouse, S. (2008), Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1(3)*, 333-342.

Hutchinson, B., and Mitchell, M. (2019, January), 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (pp. 49-58).

Jaffee, D. (2001). *Organization theory: Tension and change*. New York, NY: McGraw-Hill.

Jonas, H. (1982), Technology as a subject for ethics. *Social Research, 49*(4), 891-898.

Kane, M. T. (2013), Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

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

Kim, T. W., & Duhachek, A. (2020), Artificial intelligence and persuasion: A construal-level account. *Psychological Science, 31(4*), 363-380.

Konradt, U., Garbers, Y., Böge, M., Erdogan, B., and Bauer, T. N. (2017), Antecedents and consequences of fairness perceptions in personnel selection: A 3-year longitudinal study. *Group & Organization Management, 42(1*), 113-146.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., and Ones, D. S. (2013), Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*(6), 1060.

Landy, F. J. (1986), Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*(11), 1183-1192. doi: 10.1037/0003-066X.41.11.1183

Langer, M., and König, C. J. (2018), Introducing and testing the Creepiness of Situation Scale (CRoSS). *Frontiers in Psychology, 9*, 1-17.

Langer, M., König, C. J., and Papathanasiou, M. (2019), Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), 217-234.

Leventhal, G. S. (1980), What should be done with equity theory? New approaches to the study of fairness in social relationships. In K. Gergen, M. Greenberg, & R. Willis

(Eds.), *Social exchange: Advances in theory and research* (pp. 27-55). New York: Plenum.

Locke, E. A., Smith, K. G., Erez, M., Chah, D. O., and Schaffer, A. (1994), The effects of intra-individual goal conflict on performance. *Journal of Management, 20(1*), 67-91.

London, M., and McFarland, L. A. (2017), Assessment feedback. In J.L. Farr and N.T. Tippins (Eds.) *Handbook of Employee Selection* (pp. 406-425). Routledge.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019), Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90-103.

Markus, K. A., and Borsboom, D. (2013), *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., and Ahmed, S. M. (2017, Applicant perspectives during selection: A review addressing "So what?,""What's new?," and "Where to next?". *Journal of Management*, *43*(6), 1693-1725.

National Highway Traffic Safety Administration (n.d.). Fatality Analysis Reporting System / Summary. Retrieved from https://www-fars.nhtsa.dot.gov/Main/index.aspx

Nishii, L. H., Lepak, D. P., and Schneider, B. (2008), Employee attributions of the "why" of HR practices: Their effects on employee attitudes and behaviors, and customer satisfaction. *Personnel Psychology*, *61*(3), 503-545.

Nolan, K. P., Carter, N. T., and Dalal, D. K. (2016), Threat of technological unemployment: Are hiring managers discounted for using standardized employee selection practices?," *Personnel Assessment and Decisions, 2(1*), DOI: 10.25035/pad.2016.004

Orben, A. (2020), The Sisyphean cycle of technology panics. *Perspectives on Psychological Science, 15*(5), 1143-1157.

Orlikowski, W. J. (1992), The duality of technology: Rethinking the concept of technology in organizations. *Organization Science, 3*(3), 398-427.

Ötting, S. K., and Maier, G. W. (2018), The importance of procedural justice in human–machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior, 89*, 27-39.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000), A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Principles for the Validation and Use of Personnel Selection Procedures (2018), *Industrial and Organizational Psychology: Perspectives on Science and Practice, 11*(Supl 1), 2–97. https://doi.org/10.1017/iop.2018.195

Pyburn Jr, K. M., Ployhart, R. E., and Kravitz, D. A. (2008), The diversity–validity dilemma: Overview and legal context. *Personnel Psychology*, *61*(1), 143-151.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., and Alahmad, R. (2020), Designing fair AI for managing employees in organizations: A review, critique, and design agenda. *Human–Computer Interaction*, 1-31.

Rupp D. E., Shao R., Jones K. S., and Liao H. (2014), The utility of a multifoci approach to the study of organizational justice: A meta-analytic investigation into the consideration of normative rules, moral accountability, bandwidth-fidelity, and social exchange. *Organizational Behavior and Human Decision Processes, 123*, 159–85.

Ryan, R. M., & Deci, E. L. (2000), The darker and brighter sides of human existence: Basic psychological needs as a unifying concept. *Psychological Inquiry, 11(4*), 319-338.

Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society, 1*(1), 25-39.

Strohmeier, S. (2020), Algorithmic decision making in HR. In T. Bondarouk and S. Fisher (Eds.) *Encyclopedia of Electronic HRM*, pp. 54-59, DeGruyter Publishing.

Strohmeier, S., and Piazza, F. (2015), Artificial intelligence techniques in human resource management—a conceptual exploration. In C. Kahraman & S. Ç. Onar (Eds.) *Intelligent Techniques in Engineering Management* (pp. 149-172). Springer.

Sun, R. (2016), *Anatomy of the mind: Exploring psychological mechanisms and processes with the Clarion cognitive architecture.* New York, NY: Oxford University Press.

Yankov, G. P., Wexler, B., Haidar, S., Kumar, S., Zheng, J., and Li, A. (2020), *Algorithmic Justice*. SIOP White Paper Series. Society for Industrial and Organizational Psychology.

World Health Organization (2020), World Health Data Platform: Road Safety. https://www.who.int/gho/road_safety/mortality/en/

Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2019), Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32,* 661–683. https://doi.org/10.1007/s13347-018-0330-6

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
28-07-2022	Final	June 2020 – March 2022

4. TITLE AND SUBTITLE

Held to a Higher Standard? Fairness Perceptions of Artificial Intelligence in HR Decision Making

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
622785

6. AUTHOR(S)

Fisher, S. L.
Howardson, G. N. H

5d. PROJECT NUMBER
A790

5e. TASK NUMBER
1101

5f. WORK UNIT NUMBER
PAMU

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U. S. Army Research Institute
for the Behavioral & Social Sciences
6000 6TH Street (Bldg. 1464 / Mail Stop 5610)
Fort Belvoir, VA 22060-5610

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Institute
for the Behavioral & Social Sciences
6000 6TH Street (Bldg. 1464 / Mail Stop 5610)
Fort Belvoir, VA 22060-5610

10. SPONSOR/MONITOR'S ACRONYM(S)
ARI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT: Distribution Statement A: Approved for public release;

13. SUPPLEMENTARY NOTES: Book chapter published in Handbook of Research on Artificial Intelligence in Human Resource Management (ISBN 978 1 83910 752 8)

14. ABSTRACT

Implementation of decision-making tools based on artificial intelligence (AI) has introduced a new set of concerns related to fairness in organizations. It is clear that human decision makers demonstrate biases and make errors in decisions related to hiring, compensation, discipline, and other human resource (HR) topics. Regardless, applicants and employees are expressing high levels of concern about the fair and appropriate use of AI in HR decision making. This chapter reviews research and theoretical perspectives about fairness in HR, including organizational justice and social exchange. We then compare these perspectives with research from the AI and information systems literatures to examine different viewpoints on fairness and why the fairness of decisions made by humans and AI systems might be judged differently. We highlight a potential paradox in thinking about automation and intentionality of AI decision tools, and discuss the potential impact on the use of these systems within HR.

15. SUBJECT TERMS
talent management, personnel decisions, technology, sociotechnical systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Charles Keil
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	21	19b. TELEPHONE NUMBER 703-851-7711