Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions

Hyanghee Park
Human-Computer Interaction+Design Lab.,
Seoul National University
hyanghee@snu.ac.kr

Daehwan Ahn
OID Dept., The Wharton School,
University of Pennsylvania
ahndh@wharton.upenn.edu

Kartik Hosanagar
OID Dept., The Wharton School,
University of Pennsylvania
kartikh@wharton.upenn.edu

Joonhwan Lee
Human-Computer Interaction+Design Lab.,
Seoul National University
joonhwan@snu.ac.kr

ABSTRACT

Enterprises have recently adopted AI to human resource management (HRM) to evaluate employees' work performance evaluation. However, in such an HRM context where multiple stakeholders are complexly intertwined with different incentives, it is problematic to design AI reflecting one stakeholder group's needs (e.g., enterprises, HR managers). Our research aims to investigate what tensions surrounding AI in HRM exist among stakeholders and explore design solutions to balance the tensions. By conducting stakeholder-centered participatory workshops with diverse stakeholders (including employees, employers/HR teams, and AI/business experts), we identified five major tensions: 1) divergent perspectives on fairness, 2) the accuracy of AI, 3) the transparency of the algorithm and its decision process, 4) the interpretability of algorithmic decisions, and 5) the trade-off between productivity and inhumanity. We present stakeholder-centered design ideas for solutions to mitigate these tensions and further discuss how to promote harmony among various stakeholders at the workplace.

CCS CONCEPTS

• Human-centered computing (HCC); **• Human-computer interaction (HCI)**; **• HCI design and evaluation methods**; **• User studies**;

KEYWORDS

Artificial intelligence (AI), Fair and responsible AI, Explainable AI (XAI), Future of work, Algorithmic management, Human resource management, Stakeholder-centered design, Interpretability, Transparency, Human Intervention

CHI '22, April 29–May 05, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517672

ACM Reference Format:
Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29–May 05, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3491102.3517672

1 INTRODUCTION

Human resource management (HRM) is an essential part of many workplaces. Many enterprises, ranging from traditional industries to leading IT companies (e.g., IBM and Google), place importance on acquiring, evaluating, and developing employees [29, 62, 83]. Well-established HRM can make a business flourish while poor HRM can cause an enterprise to struggle [55, 64, 65]. Thus, the tasks HRM takes on (i.e., a series of processes ranging from hiring to firing people) is an important matter not only to enterprises but also to the various stakeholders within enterprises. HRM impacts managers, executives, and *employees*. The workplace is where an individual builds personal identity and/or finds purpose, as well as where they earn a living to support their needs (e.g., family support, financial loans).

At the center of HRM, there have always been issues of *fairness* [12, 17, 24, 42, 47, 66, 69, 90, 91] regarding work performance evaluations, such as gender/race discrimination, workplace politics, and subjective peer review systems. For example, managers or HR team members may have too much power and authority regarding their subordinates' work performance evaluations [18]. Such concentration of power often leads to workplace politics, such as when subordinates do extra work in a team project to gain favor with a superior or when a coworker free-rides off the work of others because peer reviews are not reflected in the evaluation. To mitigate these problems, enterprises have adopted peer review systems [18]. However, a horizontal evaluation system can be misused as a tool of workplace bullying or politics. Specifically, a group of workers can make secret pacts with each other to bully the same person or to praise one another lavishly [80]. Some competent employees have been sabotaged by negative comments/scores from unidentified colleagues or have received a severely low evaluation directly reflecting peer reviews [80]. Unfortunately, employees have even committed suicide or uploaded a suicide note on the workplace's online bulletin board [102] after suffering from bullying in the disguise of peer assessments or work performance evaluations [80, 102, 103]. Thus, despite its importance, traditional performance evaluation systems in HRM (i.e., human-oriented evaluation systems) have not yet solved the problem of fairness.

Both researchers and practitioners have hoped that AI might be able to solve the issues of human bias and unfairness present in traditional HRM. Especially when adopted for performance evaluations, people expect AI to be objective, unbiased, and efficient regardless of workplace politics [4, 61]. Similarly, by applying machine heuristics [82] to HR systems, employees tend to perceive AI's work evaluations to be more objective, nonjudgmental, and accurate [61]. Recent surveys [49, 104] also show that people trust AI and perceive it to be fair in comparison to human managers.

Despite these rosy expectations, applications of AI in HRM are causing serious conflicts [40, 61] between firms and employees due to the different purposes and perspectives of AI adoption. This is because the initial adoption of AI in HRM appears to be more focused on *monitoring* and *surveilling* workers rather than solving traditional problems [4] (i.e., human bias, subjectivity, prejudice, and politics in the decision-making process). For example, Amazon has recently implemented AI-powered cameras to monitor delivery drivers, who may lose their job if they decline to consent to AI surveillance [39, 51, 86]. While the firm defended such surveillance as an asset for safety, drivers perceived it as a privacy invasion and micromanagement. Unfortunately, such problems have no definitive or objective solutions. According to an Amazon spokesperson, the cameras are a safety system that has decreased accidents and sign violations. However, such a system could be used as *the panopticon of a new era* to push workers to extreme productivity through supercharged surveillance. For such reasons, workers have resisted the adoption of algorithmic management, and where adopted, serious conflicts between employees and enterprises have arisen. In another Amazon case, workers protested and sued the company for utilizing AI to monitor and fire them without human intervention [105]. Because AI is being used outside of its original purposes to solve existing problems (human-bound unfairness or bias), AI's seamless application to HRM is still far away [67, 69, 83, 87].

Therefore, recent studies [42, 61, 69] have sought to find ways of improving employees' trust in AI and the fairness of its use in HRM through an employee-centered design approach. Notably, Park et al. [61] found that employees are concerned with the surveillance, privacy invasion, or forced meritocracy that AI in HRM could cause. To mitigate these concerns, they suggested designing AI features (i.e., transparency [16, 77, 89], interpretability [30, 57, 58, 60, 88], and human intervention [26, 96]) that reflect employees' needs. While their work helps clarify why employees resist algorithmic evaluation and what may alleviate their antipathy toward it [61], it is still limited because they did not consider the uniquely complicated relationships of stakeholders in the HRM domain. In real-world HRM situations, both employees and multiple other stakeholders (e.g., executives, AI designers) are complexly intertwined with different incentives (e.g., social reputation, profit). This inherently complicated nature of HRM is often characterized by tensions regarding *AI's adoption and its design features*. In this respect, the employee-centered solutions that Park et al. [19] suggested seem unfeasible unless both the firm and its stakeholders agree to them.

To solve this problem, we incorporate a stakeholder-centered design approach [19] for AI in HRM. In sub-domains of AI research in HCI (education [23], healthcare [93], public sectors [71, 74, 75, 85], etc. [25, 79, 98]), scholars have emphasized the importance of including various stakeholders in the AI design process [44, 45, 53]. This is because reflecting the values of only one stakeholder group (e.g., system/policy designer) in the early stage of AI design overlooks the values of other stakeholders (i.e., the people impacted by AI). Aside from the ethical consequences, such decisions can lead to AI system failure that is costly to fix or remove. In this sense, the adoption of AI in HRM is another high-stakes sociotechnical context where fast-evolving technologies, users, stakeholders, and even social systems are complexly intertwined; thus, it is necessary to look at the problems from a bird's-eye view to design a fair system. However, in the field of HRM, research on *stakeholder-centered fair AI design* is extremely lacking. Especially, AI design that reflects *stakeholders' tensions*—which may be the key factor to designing fair AI in HRM—has not been sufficiently explored. Therefore, we first seek to understand the various sociotechnical tensions surrounding AI in HRM through the perspectives (e.g., different goals or scope of AI adoption) of various stakeholders and AI's key design features. Then, while examining the design features for fair AI, we ultimately seek to provide *stakeholder-centered ideas* (e.g., design agenda for AI features). By doing so, we aim to uncover *where issues concerning fair AI emerge in HRM, what kinds of tensions and trade-offs exist in designing key AI features in HRM, and how to design AI for HRM so that all stakeholders find it fair and acceptable.*

In the sense that we stress the role of multiple stakeholder participation, our research aligns with Zhu et al.'s Value Sensitive Algorithm Design [98] and following works [71, 75, 79], which involve diverse stakeholders in the initial algorithm design stages and reflect their values/insights as inputs when creating an algorithm. However, Zhu et al. 1) chose to "remove design options that some stakeholders strongly object[ed] to" [98] and 2) did not use a design process that intentionally promotes an environment to identify and resolve stakeholders' value conflicts in their work. Because our research centers on high-stakes HRM contexts, which are socially complex and involve the conflicting incentives/interests of stakeholders, it may be impossible to eliminate controversial design options as Zhu et al. [98] did.

Thus, we do not discard such design options [98], and instead aim to 1) deliberately foster an environment to detect and appreciate stakeholders' conflicting values and 2) better define and examine these tensions (i.e., value conflicts) surrounding AI in HRM by further incorporating the concept of *wicked problems* [8, 11]. *Wicked problems* refer to problems that cannot be definitively described and have no objective solutions, such as policy problems. Designing fair AI in HRM is also a wicked problem, because there is no objective definition of fairness that satisfies all stakeholders. What is fair to one stakeholder may not be considered fair by another stakeholder. Similarly, there is no way to definitively prove that the proposed solution is optimal and final [68]. Thus, designing fair AI in HRM requires design thinking tailored to wicked problems [8], which begins with a clear understanding of stakeholders' different incentives, values, and beliefs.

To better identify and tackle the wicked problems of designing fair AI in HRM, we adapted Roberts's coping strategies for wicked problems [70] (i.e., authoritative, competitive, and collaborative design) and merged them into the design thinking process [8] (i.e., iterative design). Our approach to modify Roberts's coping strategies is useful for uncovering multiple stakeholders' tensions in advance so that designers and researchers can tackle them. We used this approach in a scenario-based participatory workshop that included both in-depth interviews and codesign sessions with a variety of stakeholders. During the four-stage participatory workshop sessions, the stakeholders actively discussed the fairness of AI evaluating employees' work performance and the sociotechnical tensions surrounding AI in HRM. Then, they brainstormed design ideas to balance the tensions. As a result, the following findings were identified (for details see 4):

- Stakeholders had common interests in 1) the fairness issues of traditional HRM, 2) the goal of AI adoption and the scope of AI tasks, 3) the factors that determine the nature of AI, 4) the impact of AI adoption, and 5) who is responsible for algorithmic decisions. Other interests were specific to a particular group: The employee group emphasized fair and responsible evaluations while HR/employers focused on reducing time and cost. The expert group explored the potential and limitations of AI.
- A total of five major tensions surrounding AI in HRM were identified among stakeholders: 1) the perception of fairness, 2) the accuracy of AI, 3) the transparency of the algorithm and its decision process, 4) the interpretability of algorithmic decisions, and 5) the trade-off between productivity and inhumanity.
- We present AI design insights to balance each tension mentioned above: 1) take organizational and systemic approaches to fairness rather than optimizing fairness itself, 2) complement AI's imperfections in accuracy with human oversight, 3) design partial transparency in accordance with its benefits and costs, 4) carefully design interpretability to avoid imperfect (i.e., harmful) interpretability, and 5) pursue sustainability regarding the trade-offs between productivity and inhumanity.

Our contributions to the HCI field are as follows.

- By viewing the fairness of AI in HRM as a wicked problem [8, 11, 70] and extending the existing employee-centered design to stakeholder-centered design [19], we identified *five tensions among stakeholders* that inherently emerge due to AI's features (e.g., accuracy, interpretability).
- We methodologically integrated Roberts's coping strategies [70] with iterative design thinking. This process facilitated an environment conducive to debate where stakeholders discovered AI features and each other's value differences (i.e., tensions), resulting in design insights.
- Finally, we provide a list of fair AI design insights for both researchers and practitioners to use when designing fair, stakeholder-centered AI for HRM.

2 RELATED WORK

2.1 AI in HRM

In recent years, AI has gradually been adopted in HRM, and tech giants (e.g., Amazon [100, 106], IBM [29, 40, 73]) have begun to hire, train, and even fire their workers based on algorithmic decisions [13, 61, 90]. This adoption of AI in HRM has already created conflicts, but some research [42, 69] has helped to elucidate employees' perceptions and requirements regarding AI in HRM. To answer why employees resist or even protest the adoption of AI in their workplaces (e.g., warehouse workers who sued Amazon when AI was used to fire them [105]), Lee [42] discovered that people are more likely to perceive evaluation by AI as less fair or less trustworthy than humans. Additionally, a recent study by Park et al. [61] suggested that employees resist algorithmic evaluation due to six types of burdens in human-AI interaction (i.e., emotional, mental, bias, manipulation, privacy, and social burdens). To mitigate such negative burdens, employees strongly desired transparency, interpretability, and human-AI collaboration (rather than solely human or solely machine evaluation) [61].

Nevertheless, employee-centered solutions may be limited or even unrealistic. Decision-making in HRM inevitably involves not only employees but also key stakeholders (e.g., CEOs, executives, HR teams, technical experts) [29, 31, 69, 83]. Thus, tensions and/or conflicts among stakeholders can arise due to the differing incentives of each [69, 83]. For example, employees may request a high level of transparency in the algorithmic decision process, but the CEO, HR team, or AI designers may not want the same level of transparency due to the risk of unexpected problems (e.g., increased complaints, breach of HRM know-how, ethical/technical reasons [24]). In other words, if there is no unified consensus among stakeholders, employee-centered solutions may neither mitigate conflicts surrounding AI nor improve its adoption in HRM.

To tackle this problem, a variety of HCI research domains, including but not limited to healthcare [28, 93], high-stakes public sectors [71, 75, 85], and online community design [79, 98], have emphasized the importance of incorporating multiple stakeholders' perspectives [19, 20] in the AI design process. For example, Holstein et al. [23] included both students and teachers when designing AI educational interventions in classrooms. Also, Robertson et al. [71] realized the importance of stakeholders in the early stage of AI design, after finding that the student-school matching algorithm had failed. They determined that the views of the parents and the San Francisco Unified School district policy designers had not been reflected [71]. Similarly, our study builds on prior research that has investigated AI in HRM from an employee-centered perspective [24, 42, 69, 83] by incorporating a stakeholder-centered design approach [19]. We aim to holistically understand AI in HRM and its sociotechnical context from various stakeholders' points of view.

However, as Selbst et al. [77] have argued, algorithmic solutions designed for one social context (e.g., high-stakes public sector) may be misleading, inaccurate, or otherwise do harm when applied to different contexts (e.g., HRM of a private firm). Thus, we differentiate our study from the work of Holstein et al. [24], which focused on AI practitioners' challenges in general AI applications, by deeply exploring the tensions surrounding AI in HRM among *multiple business stakeholders,* including AI practitioners.

2.2 Fair AI in HRM

Two streams of research have emerged that are relevant to the issues of fair and responsible AI. The first stream examines the concept of fairness from a technical perspective—especially *algorithmic (un)fairness*, which includes *algorithmic bias* and *algorithmic discrimination* (i.e. when biased algorithms privilege or discriminate against a group based on their gender, race, ethnicity, etc.) [41, 47, 77]. This stream assesses a range of tasks, such as sorting and eliminating bias in training data, conceptualizing and mathematizing “fairness” as a goal for model optimization, and engineering fairer algorithms/models by optimizing the mathematically defined functions regarding fairness [77]. However, such technical viewpoints are limited in resolving conflicts surrounding AI in the workplace for two reasons. First, designing fair AI in HRM is a wicked problem in which there is no single, objective definition (i.e., mathematical formula) of fairness. Second, existing technical approaches have overlooked the sociotechnical context surrounding the system—where technologies, employees, HR team/executives, technical experts, and even social systems are complexly intertwined with each other.

The second stream focuses on socially fair outcomes (i.e., perceived fairness) rather than algorithmically fair outcomes (i.e., algorithmic fairness) [41–43, 52]. However, due to the “complexity of HR phenomena, data challenges from HR operations, fairness and legal constraints, and employees' antipathy to AI management” [83], there has been little research on AI's application to workplaces. Thus, further research on fair AI design in HRM is even more lacking. In one of the few existing studies, which is based on a literature review of twenty-five design papers, Robert et al. [69] categorized three different types of fairness (i.e., distributive, procedural, and interactional fairness) that could emerge when designing AI for employee management. Park et al. [61] also took a step forward to investigate why employees feel that algorithmic management is unfair, especially in the *algorithmic evaluation of work performance*. Although these recent studies contributed to understanding employees' perceptions toward algorithmic decisions, they are still limited because they only put *employees* at the center of algorithmic management. Since HRM inevitably involves both employees and other social actors (e.g., HR team) to make work performance evaluations, the various perceptions of fairness that multiple stakeholders have concerning AI in HRM should be explored. To fill this gap, we further examine what tensions and challenges exist among various stakeholders in fair AI design and what might resolve them.

3 METHODS

3.1 Coping Strategies and Design Thinking for Wicked Problems

Roberts [70] proposed three coping strategies—authoritative, competitive, and collaborative strategy—that public officials and managers can utilize to deal with wicked problems. The ***authoritative strategy*** is employed when power is concentrated in a small number of stakeholders. The stakeholders are authorized to define the wicked problem(s) and figure out solutions by heavily relying on a few peoples' expertise. When the power is more dispersed and contested among a diverse set of stakeholders, ***competitive strategies*** are suggested. Stakeholder groups battle to win the authority to define the problem(s) and select solutions by assuming it is a zero-sum game. Finally, if power is neither concentrated nor contested, then ***collaborative strategies*** can be utilized. Various stakeholders work together to create win-win solutions by enlarging the pie for all parties involved rather than playing a zero-sum game.

While Roberts's coping strategies [70] have contributed to specific and actionable options (i.e., trichotomy and black-and-white options), the approach relies on categories and irreversible rules. Roberts [70] suggests choosing one of the three coping strategies depending on the characteristics of the given problem. However, in the case of our study, the authoritative, competitive, or collaborative strategies could each differently contribute to designing fair AI in HRM. Specifically, the authoritative strategy is advantageous for shaping stakeholders' initial concepts and needs regarding AI when stakeholders are not yet familiar with it. The competitive strategy could easily highlight the conflicts (i.e., tensions) among stakeholders. The collaborative strategy can be used to derive a win-win solution, which is the ultimate goal when designing fair AI for HRM. Thus, we utilized all three strategies together to better investigate the needs, tensions, and solutions for conflicts surrounding AI in HRM.

The principle of *irreversible rules*, which does not consider the repetitions or iterations of each problem-solving process, may be viable when the social problems (see as examples [70]) are wicked but familiar to the stakeholders, who are experts in their domains. However, since AI has emerged as a *new social actor* at the center of wicked problems, the stakeholders involved in the issues surrounding AI in HRM are not experts; most of them have low AI literacy [50] because they have not seen or used AI in HRM. Thus, every step (i.e., authoritative, competitive, or collaborative design) for fair AI in HRM is inevitably accompanied by much trial and error, which cannot be solved by a linear process.

Regarding these concerns, Buchanan [8] pointed out that, unlike design processes, “categories” have fixed meanings and, once created, can only analyze “what already exists.” However, AI in HRM is a new experience to stakeholders; thus, it does not already have categories (i.e., the fixed meaning accepted within the framework of a theory or a philosophy) to scrutinize. Also, he added, “the actual sequence of design thinking and decision making is not a simple linear process” [8]. By considering Buchanan's perspective, we incorporated Roberts' three coping strategies into an iterative design process. Instead of the category-oriented approach, we use “placements” of signs, things, actions, and thoughts through which patterns can be found in the iterations of authoritative, competitive, and collaborative design. This design process allows the stakeholders to “position and reposition the problems and issues at hand” [8] (i.e., iterative design). By doing so, both researchers and participants could intuitively and deliberately shape a design space for each group's views (e.g., needs and concerns). Ultimately, this process could generate “a new perception of that situation” as well as “a new possibility to be tested” [8].

3.2 Recruitment and Participants

The purpose of our research is to identify tensions among various stakeholders' needs and to elicit design insights from stakeholders through participatory workshops. To obtain various viewpoints

Table 1: Participants' self-reported occupations and roles in the workplace

Stakeholders	Occupations/Roles/Industry Name Description	Participant No.
Employees	Office and sales workers	P01-P08
	Knowledge workers (i.e., product managers, research and development)	P09-P13
	Service and health-related workers	P14-P19
	Menial and physical workers (i.e., logistics, production, and construction)	P20-P24
Employers	AI-based education, Drone/AI-based image processing, Architectural design, Data visualization solutions, and Construction	P25-P30
HR Managers	Multi-industry, Electronics, Gaming, Automobile, Construction, AI-based education, Fashion, and Cosmetics (i.e., seven of them are world-leading companies that have at least one hundred thousand employees worldwide)	P31-P41
AI/Business Experts	AI designers who have experiences in 1) publishing and delivering presentations of first-author papers at top AI conferences (i.e., CVPR, NeurIPS, ICML, etc.) and/or 2) developing ML/DL-based systems in the industry.	P42-P47
	University professors in the intersection of AI, business (marketing, operations, and information systems), and industrial engineering.	P48-P50

from multiple stakeholders (i.e., employee and professional groups), it was important to recruit a diverse and representative set of participants with different demographic and occupational backgrounds. Thus, we distributed our recruitment poster on social media and used the snowballing method to recruit both employees and professional groups.

For the employee group, a total of sixteen participants contacted us to participate in our study. To increase the diversity of the samples, we contacted our personal contacts and repeated the snowballing method to recruit an additional eight participants. As a result, a total of twenty-four participants (P1–P24) with diverse backgrounds registered for our study, and they all successfully completed all of the sessions.

To recruit a diverse group of professionals, the authors emailed a variety of global business leaders and corporate stakeholders through their personal contacts. Some of the initial respondents helped disseminate our recruitment poster to enterprises or other business leaders. Consequently, we recruited nineteen professional participants through our personal contacts and snowball sampling, and seven participants joined from our recruitment via social media. The occupation of the professional participants is comprised of entrepreneurs, employers, HR managers, certified labor consultants, labor attorneys, AI experts, and business/economics scholars, among others.

A total of twenty-six professionals (P25–P50) participated and successfully completed all of the design sessions. All fifty participants were compensated 20 USD each.

To protect the privacy and security of our participants as well as their firms, we let the participants choose between two formats of interview and codesign sessions: a one-on-one or a focus group of three to five people. This hybrid style of research design is widely used for sensitive research in case participants feel uncomfortable sharing private stories with others or when their disclosure could threaten their status or security in society. During our recruitment, several participants requested a one-on-one session because they were concerned about sharing their firms' current system or issues with other corporate stakeholders. Also, due to the sensitive nature of the work, some participants requested not to be put into an interview or codesign focus group with others from their own company. Because some employees may not have been able to directly/freely talk about their opinions in front of other stakeholders from their company (e.g., HR or the CEO), we endeavored to actively reflect the workers' sensitive circumstances in our study design. For example, we decided not to collect any of the participants' private information, such as their affiliation, race, or foreign status. Even if some participants explicitly mentioned sensitive information (e.g., their company name) in the process of our study, we intentionally excluded it from our transcripts. Additionally, we paraphrased parts of the transcripts that could seem like whistle-blowing.

3.3 Procedures

3.3.1 Interviews and Participatory Workshops. As shown below, we conducted four stages of participatory workshop sessions that allowed multiple stakeholders to be actively involved in the research. Participatory workshops enable stakeholders/experts to 1) identify tensions as an informant that gives feedback (as legitimate end-users, direct corporate stakeholders, or domain experts) and 2) envision design ideas or future agendas as equal design partners. The first author also engaged in the design session to play the following essential roles: 1) *a moderator* who creates a free and open space for active brainstorming by asking prepared questions/prompts, 2) *a messenger* who conveys other groups' opinions, and 3) *a mediator* who encourages all of the stakeholders to address tensions through heated debates and to reach a consensus for future design ideas. Regardless of the participation option (i.e., via a one-on-one or focus group), the moderator actively shared other stakeholders' diverse ideas, needs, and opinions in each session. We strictly followed all of the research procedures and protocols that the IRB approved. During the session, the moderator made sure not to disclose anyone's personal or sensitive information. After notifying participants that audio recording was required and receiving their informed consent, we began the following sessions, which lasted approximately 90–120 minutes.

Table 2: Scenarios of AI in HRM presented to participants

Category	Scenarios of AI in HRM	Reference
Evaluation for Menial Labor	A logistics company adopted AI that automatically tracks its delivery drivers through AI-powered cameras. The AI camera monitors and records drivers' behaviors during the delivery time. If the AI judges that the drivers do something dangerous, like checking their phones or yawning, it automatically generates warnings in the middle of driving and suggests that drivers keep their eyes on the road or take a break. According to the company, this AI-powered safety system can help decrease accidents and sign violations.	Amazon [40, 86]
Evaluation for Skilled Labor	An IT company predicts employees with a greater propensity to leave and evaluates employees (e.g., top talent developers, salespersons). Also, the company analyzes unstructured content from their annual and pulse surveys and in-company social media to improve the firm members' communication skills and engagement in their team. Also, AI recommends when to reward strong performance with pay raises or bonuses by including the market rate for employees' skills and how in demand the skills are.	IBM [22, 107]
Evaluation for Emotional Labor	A customer service center adopted AI to analyze and evaluate their employees' hidden efforts (e.g., their warm tone/language, coping capabilities with blacklist customers, content and length of each call) beyond mere customer satisfaction ratings. By combining the AI's evaluation matrix with customers' satisfaction, employees' salary reduction and promotions/bonuses are decided.	Call Center & IBM [42, 63]
Monitoring and Surveillance	In a construction company, AI automatically detects dangers on site and monitors workers via CCTV cameras. Through vision and biological identification techniques, the AI tracks and classifies a variety of loitering and working motions (e.g., smoking vs. hammering) to calculate each worker's actual work time and evaluate their work performance. The firm explains that such surveillance can help decrease accidents on site and provide fairer rewards to dedicated workers.	Construction (CCTV) [101]

3.3.2 Design Sessions. **Authoritative design.** The purpose of this initial stage was to enable each stakeholder group (or an individual representing a stakeholder group) to conceive *individual or group needs*, by brainstorming AI applications and envisioning AI that meets their needs and expectations. This process was essential because, regardless of each stakeholder group's characteristics, most of the participants have not experienced AI evaluating workers in their workplace. Thus, by vividly imagining AI in HRM on their own terms, stakeholders could become more familiar with AI's features and application within the HRM domain. During the process, we encouraged participants to share how AI should be designed to satisfy the needs of the group they belonged to as well as their own needs. The specific interview questions and codesign prompts were as follows: "What do you think about AI performing HRM tasks, such as work performance evaluations?" "What advantages or disadvantages could AI bring to you or the firm?" "Would AI adoption in HRM make performance evaluations fairer?" "What AI features/issues should be considered before adopting AI in HRM?" Based on their responses, we let them design the most ideal AI for HRM according to their own unique needs, requirements, values, etc.

In the beginning, we did not show any real-world use cases or scenarios in which AI is used to evaluate work performance evaluations. This is because showing such scenarios could affect the participants' initial perceptions of AI in HRM. However, when participants expressed difficulty in imagining and designing AI for HRM, we informed them that detailed scenarios would be given later and encouraged them to freely design AI without any specific references in the meantime. After the initial cycle of the authoritative design process was conducted, we showed the detailed scenarios (that we adopted and modified [42, 61]) to help participants better understand how AI currently works in real-world HRM. Then, we repeated the authoritative design process by asking the same interview questions and codesign prompts once more.

Competitive design. In this session, we aimed to have stakeholder groups compete to make AI in HRM work for themselves; they debated AI designs by representing their group/individual incentives. This enabled us to concretely identify most of the important issues that each stakeholder group had discovered through the authoritative design step. While disagreeing with other stakeholder groups' opinions about the definition of normative AI in HRM and its problems, participants could identify what trade-offs exist in AI design and where critical conflicts or tensions emerge. To foster broad and constructive debates, the moderator introduced a set of design agendas, AI requirements, and needs that other people had previously discussed and produced in their design sessions. When the fierce debates concluded, both the moderator and participants summarized the discussion by grouping the AI design agendas and issues into similar categories of tension.

Collaborative design. Centering around the critical issues (e.g., trade-offs and tensions) identified in the competitive design, stakeholders collaboratively discussed future AI design with the goal of mitigating potential conflicts and ultimately reaching an agreement. During the collaborative process, stakeholders could come up with viable design ideas for solutions. Some of the design ideas for solutions reflected the majority of the stakeholders' needs/requirements while others did not draw a consensus or were not feasible for technological or policy-related reasons. Such unresolved issues were provisionally suspended in this stage: simpler issues were readdressed later in the session and more difficult issues were moved to the next session.

Iterative design. The iterative process allowed the stakeholders to go back to a previous stage any time they needed to reinvestigate needs, conflicts, and design ideas for solutions. Such iteration is important because clearly defining wicked problems and giving shape to design ideas for solutions is inherently difficult at each stage. Stakeholders were allowed to return to a stage/idea and repeat the iterative process until final agreements were reached in the collaborative design stage. Thus, our method did not pursue *a linear design process* or solve *rigidly fixed or predeterminate problems* [8].

3.4 Data Analysis

All of the recorded audio from each session was transcribed, and the data was used for thematic analysis. By using open coding procedures, we discovered clusters of similar topics and organized them in an inductive manner according to thematic relationships. We conducted two cycles of reading through the transcripts and (re)categorizing similar themes. Any coded topic that did not have a clear them or relationship was discussed and rearranged until there was no further dispute about it. The themes were finalized after multiple rounds of inductive (re)classification and the addition of new themes, and the reliability of the results was double-checked by screening them once more. Adopting Braun and Clarke's suggestions and justification provided in [7], we did not conduct a specific intercoder reliability testing.

4 RESULTS

In this section, we present our findings from the design sessions. Section 4.1 presents the findings of the authoritative design session, and section 4.2 presents the findings of the competitive and collaborative design sessions.

4.1 Authoritative Design: Stakeholders' Needs

The purpose of authoritative design is to draw out every group's needs and interests by encouraging each group to actively participate in the design process. As a result, we discovered the common interests of all stakeholder groups as well as the distinct interests of each stakeholder group regarding the adoption of AI in HRM.

4.1.1 Common Interests and Incentives. **Fairness issues in traditional HRM.** Many participants expressed that work performance evaluations in traditional HRM are unfair. Surprisingly, participants believed unfairness starts from the very beginning of assigning a task, which amplifies unfairness and affects the work performance evaluation. P35 (HR) said, "It depends on how we define fairness. If fairness means whether the workers perceive it to be fair, I think over a half of our workers would feel our traditional evaluation system is unfair." Similarly, P41 (HR) mentioned,

> "This happens to every firm . . . there are some successful business divisions that naturally bring about good results. If you're put into such a division or a so-called 'promising' new project, the fact that you're assigned to that team leads to good performance. Then, you can get good results with little work or effort, because that business or project by nature has always received a good response from customers in the market. On the other hand, within the market structure, there are minor and dying businesses that barely make a profit in a firm. Then, even if an individual works really hard, there is nothing he can do about his work performance. I believe getting an evaluation based on what *you* [i.e., an individual employee] did is reasonable and fair. But it's hard in the current HRM system."

Similarly, P13 (employee) said, "There are some kinds of construction projects that look so profitable just looking at the blueprint at a glance. If someone is assigned to it, you'll make a profit and get promoted. It's not your performance; it is just luck."

The goal of AI adoption. Most participants expected objectivity and productivity from AI when compared to traditional HRM, regardless of stakeholder group differences. For instance, P35 (HR) said,

> "I would want to see or expect something more objective from AI. Then, I would be able to see the difference in the degree of objectivity between the AI's judgment and my judgment. There always have been points that I missed in that regard. To sum it up, AI should ultimately save time for our [i.e., a human HR team's] decision-making as well as improve the accuracy of our decision-making."

P38 (HR) reported, "Clarifying the purpose of AI adoption in HR is important even though it would be different depending on each stakeholder group. What utility would AI's adoption bring to the evaluated person? Would it be useful to them? I think those who are evaluated would expect an objective performance evaluation. P48 (AI/business expert) mentioned, "the purpose of AI adoption should be the sum of all of the stakeholders' benefits [i.e., gains or interests] by fostering employees' motivation, reducing the firms' agency costs [e.g., inefficiencies, conflicts of interest], and increasing the total productivity of the firm."

The scope of AI tasks. Participants showed a strong interest in narrowing down what specific tasks the AI should intervene in and which part(s) within those tasks the AI should automate or augment. P06 (employee) said, "It depends in what domains AI would be used. I think it's possible to partially use AI for operation or R&D areas." P50 (AI/business expert) mentioned, "I think due to technical limitations, such as lack of data, AI should be used in a way that reduces humans' inefficiency and assists humans' tasks. Critical decisions should be made by humans." Similarly, P42 (AI expert) also said, "To use stock investment as an example, the algorithm's task is close to the concept of stock item recommendation. The final judgment or critical decision should always be made by humans."

Factors determining the nature of AI. A total of twenty-one participants asked about the unique characteristics of AI and/or components that constitute AI. More specifically, participants asked, "What is AI and what does it look like?" or "Is it like a CCTV or computer software system that monitors our behavior?" Explanations of AI's results and inner procedures were also an essential focus of their interest. Over half of the participants (N = 26) asked one of the following questions: "Does AI let me know how it made such a decision?" "Can someone [i.e., a human] operate or manipulate AI?" "How can AI calculate, reason out, and produce results?" and "How can I trust the result is accurate?"

The impact of AI adoption. Especially, participants wanted to know how the adoption of AI for work performance evaluations would influence their life at work. This is because most of the stakeholders' future work—their future employment, ability to work, chance of getting fired, and the type of tasks that they take on—could significantly be affected depending on how the AI is designed and what it requires of them. P32 (HR) asked, "If AI is applied to HRM, a lot of things would change. From the evaluation system itself to the existing HR software system, there would be lots of parts to touch up." P08 (employee) said,

> "There should be infrastructure established to adopt AI for work performance evaluations. Should we use specific software tools like Google Docs so that AI can analyze who contributed to making a report, or something like that? But I guess workers wouldn't like breaking their work customs or disturbing their existing work flows just to try the new AI evaluation."

Responsibility for the algorithmic decisions. Many participants (from AI/business experts, HR managers, and employers) brought up the issue of who would take responsibility when AI makes wrong decisions. P39 (HR) described ambiguous situations in which it is hard to pinpoint the blame on one group:

> "I have two major concerns about the responsibility for the AI. First, who will take responsibility when the AI makes errors? Secondly, let's assume that the AI operates well but we cannot know and explain its mechanism or logic. Who will take responsibility if issues happen in that situation? Those two are a little different but both of them are big problems."

4.1.2 Group-Specific Interests and Incentives. While stakeholder groups had some similar interests in common, distinctive group-based interests and incentives were also identified.

Employees: Fair and responsible evaluations. The major interests and incentives of the employee group fell under the themes of maintaining wellbeing, protecting privacy, boosting capabilities/competence, improving fairness and accuracy in work performance evaluations, and preventing the reduction of communication opportunities. P02 (employee) mentioned, "I can't sacrifice my privacy. My firm just updated the speed gate system so that they can store all of our data, like the time I spend to have a smoke and go downstairs to pick up packages. I already hate the fact that they store it, but now it will be used for work performance evaluations? Workers will suffocate."

HR/Employers: Reducing time and costs. Our findings show that a total of sixteen participants mentioned that AI in HRM should provide benefits for their group and/or the firms in terms of reducing their "time, effort, and costs." Especially, participants emphasized productivity and efficiency (e.g., the simplification of administrative work). For instance, P40 (HR) reported, "We should think clearly about the purpose of the AI evaluations. It may sound very HR-oriented, but increasing efficiency in evaluation, that's what the HR team wants and speaks for." Similarly, P38 (HR) explained,

> "Work performance evaluations are an extremely important task that we carry out annually in the HR department. There are lots of time-consuming and effortful tasks, like making administrative documents. Looking at AI's adoption through the perspective of an HR practitioner like me, I hope the AI helps out a little bit with such HR tasks or reduces the workload. Then, our work could be done more simply or easily."

AI/Business experts: Potentials and limitations of AI. Our findings show that AI designers and business scholars pay attention to AI's technological possibilities. For instance, P49 (AI/business expert) said,

> "AI's role is comprised of three major parts. The first is datafication, meaning AI makes it possible to collect and use new data that was once impossible in the past. The second is prediction, which is a task that anticipates employees' performance based on the collected data. The third is optimization. AI is used to optimize the inefficient parts and achieve efficiency. I think, among the three, AI will first be utilized for optimization tasks and then for datafication tasks. However, in terms of prediction, we have a long way to go technologically. So, I think we should pay special attention to this part [i.e., prediction] and have more careful discussions."

Similarly, P48 (AI/business expert) mentioned, "Although it is possible to automate certain parts of HRM through algorithms, it is essential to monitor, diagnose, debias and even audit the AI through active human interventions." Regarding the potential of AI, P45 (AI expert) said,

> "The details of AI design should be tailored or applied differently after considering the nature of the various businesses or tasks. In such fields where quantitative evaluations are already common, the employees might welcome AI's adoption to HRM. For example, in the case of customer service, work performance is traditionally evaluated based solely on customer satisfaction or call success rates. But if AI datafies more diverse variables, like the customer service agent's warm tone or emotions conveyed during the call, the agent might like that the AI recognizes their hidden efforts. However, in fields where creativity is more appreciated, the AI should be designed to maximize future value by promoting *trial and error* [i.e., creative risk-taking] to workers."

4.2 Competitive and Collaborative Design: Tensions Surrounding AI and Design Insights to Balance the Tensions

We found tensions in five significant areas, each with corresponding design insights: perspectives on fairness (4.2.1), tradeoffs surrounding the accuracy of AI (4.2.2), transparency of the algorithm and the decision process (4.2.3), interpretability of the algorithmic decision (4.2.4), and trade-offs between productivity and inhumanity (4.2.5). By defining and explaining the tensions here, we seek to present the issues at hand that are worth thinking about in advance. Also, we describe the design insights that were discovered in the process of *collaboratively envisioning* solutions that could balance the tensions. Thus, the purpose of providing the design insights in this section is not to guarantee concrete solutions, but to delineate design spaces [99] from which designers can start to untangle the problems and scholars can scientifically approach/test for solutions (see Buchanan's point in 3.1. [8]). Since our research is an early study that directly engaged *diverse stakeholders for AI design in the HRM context*, the design solutions (which we have interpreted as design insights) shaped by our participants could become valuable for future/follow-up research.

4.2.1 Perspectives on Fairness.

4.2.1.1 Tension: "Stakeholders from different planets." As expected, our findings show that not only each stakeholder group's needs, but also the definitions and concepts of fairness held by them are in conflict with one another. While employees desired fairness in adopting AI for work performance evaluations, HR/employers argued that fairness itself is not the ultimate reason for the firm's existence. The AI experts focused on how to (mathematically or economically) measure fairness. We present each group's arguments in the following section.

Employees: "We want a fair decision process." Our results show that employees emphasized *procedural fairness* (i.e., fair decision processes) in AI's application to HRM. Employees expect objectivity in work performance evaluations and believe fairness can be achieved through designing transparent and fair processes. P04 (employee) said,

> "When I hear AI will be adopted by HRM at my work, I naturally expect it will be used for an objective and fair work performance assessment—like getting paid for the work done regardless of gender, or something like that. If AI still makes the same bias or fault we [humans] used to make, why try a new thing? Isn't it more efficient to just stick to the traditional way? There's no need to break the custom unless it's better. If we have to adapt to the new HR environment of AI without anything to gain, it is bothersome or just adds a new type of burden on us."

Individual workers also viewed fairness as receiving fair credit for what they did. P13 (employee) also added, "I overwork and always help other colleagues to learn how to use difficult software tools for architectural design, because I graduated from a top-tier school. But I'm under-appreciated just because I'm female. The AI should help break the glass ceiling and promote equal pay for equal work."

HR managers and employers: "Fairness is not the reason for the firm's existence." Contrary to employees, HR managers and employers argued that fairness is a means of securing a company's sustainable competitiveness and should not be a goal in itself. This group tended to show a negative stance toward centering fairness in HRM and viewed work performance evaluations as a part of increasing the firm's benefits. P41 (HR) explained,

> "The purpose of HRM is neither fairness nor equity, and I haven't ever evaluated workers with the aim of fairness. The purpose of HRM is the enterprise's growth, [so] the fairness of the process can't be the goal of HRM. The nature of HRM and HR evaluation I do is to evaluate workers based on what they have done and *will* do, to assess their suitability to the organization. Many workers would think that they should be evaluated based on what they have done in the past. However, giving a good evaluation means giving opportunities for promotions or priority when negotiating salary. And the firm does so [i.e., gives a good evaluation] because the firm is sure that this employee will bring more profit, such as sales or growth, to the firm. So, making an evaluation strictly based on the sales that a worker has made in the past and giving 100% credit for it isn't sound and desirable HRM."

P49 (AI/business expert) also said,

> "I don't think we should define fairness as employees' satisfaction or increasing it. We should differentiate fairness for enterprise from what we call 'socially fair' and define them differently. This is because there is a difference between the purpose of the enterprise and that of society. So, when trying to directly apply the social definition of fairness to enterprises, conflicts will surely occur and are unavoidable."

AI/Business experts: "If you cannot measure it, you cannot improve it." Similar to the HR/employers group, the expert group showed a negative attitude toward defining fairness but for a different reason—its technical limitations. Experts pointed out that since it is almost impossible to mathematically define and quantitatively measure the fairness of AI in HRM, maximizing fairness itself is difficult. For that reason, many experts said the fairness people pursue is too subjective, and it is necessary to set a more measurable goal. P50 (AI/business expert) explained,

> "Because fairness is an extremely subjective and technical space intertwined with value judgments, it is difficult to precisely define and quantitatively measure the concept of fairness. To solve a certain problem, it is important to set an objective and measurable goal. Without having a systematically clear direction, we cannot achieve the goal. Of course, fairness is important, but it is practically impossible to maximize an index that is difficult to quantitatively measure in the first place. So, I think that fairness shouldn't be the ultimate goal itself."

P49 (AI expert) mentioned,

> "If you really want to measure fairness, you can do that by mathematically defining it. There are so many mathematical ways to define fairness, . . . There are even trade-offs among the varying definitions of fairness. So, it seems impossible to define one measurable

> fairness that everyone agrees on. . . . Also, people's conception of fairness is too subjective. I think people have issues with machine learning algorithms, even if it does the same thing [as humans]. Even people who don't show dissatisfaction with an HR team's imperfect human decisions become fussy about [decisions] when they hear that they were done by ML [i.e., machine learning]. Look at the cases of auto-driving. Even if the car accident rates are way lower, people respond more sensitively to just one accident that algorithms make. This is a major hindrance to AI adoption."

4.2.1.2 Design Insights: "Utilizing organizational and systemic approaches." Our findings—participants' varying views on fairness—reconfirmed a lack of existing literature that differentiates types of fairness in the *employees' position* [69] and *the AI engineers' difficulties* in optimizing (mathematically-defined) fairness [24, 77, 85]. However, by theoretically adopting the definition of wicked problems [8, 70] and applying it to AI adoption in HRM, we further discovered that the stakeholders in HRM (such as HR teams and employers) have needs/incentives regarding fairness that *severely conflict* with each other. Also, most participants could not objectively define or optimize fairness. For this reason, many participants mentioned that each organization, through trial and error, needs to reach a unique ideal where AI and stakeholders coexist harmoniously. In this vein, the expert group suggested firms take a structural and systemic approach (e.g., openness, mutual surveillance, intentional slack) *beyond endeavoring to optimize fairness itself,* and other stakeholders (i.e., employees and HR/employers) agreed.

Openness and mutual surveillance. After all of the stakeholders realized that AI cannot be perfect, they reached a consensus that it is critical to design AI under the assumption that it could have errors. P35 (AI expert) said,

> "From the very initial stage of envisioning the system, we should keep in mind that algorithms could have errors. When designing AI, we must consider some buffers in case AI makes mistakes. Also, to appeal to the employees, giving them some intentional slack would increase their trust toward the evaluation system as well as the AI."

Instead of hiding the imperfections in AI, participants asserted that each stakeholder group has a right to raise and discuss fairness issues from the earliest stage of AI adoption. They believed discussing issues openly and not giving in to the mainstream opinion (e.g., of the HR team) could enable members of the organization to mutually check the power structure. In that respect, participants emphasized that decision-making for the adoption of AI in HRM should be accompanied by enough discussion for consensus and persuasion. They also agreed it is necessary to build a system for mutual surveillance between humans and AI (e.g., external AI audit) that goes beyond human-AI collaboration. P49 (AI/business expert) gave an example of an auditing system: "Although many employees don't know about finance, they perceive it to be fair because an auditing system exists. That means we should set a goal to design a broad system that the members could accept as fair. We could focus on building and maintaining a healthy ecosystem by allowing mutual surveillance through external audits."

Also, for AI to be adopted seamlessly, many participants agreed that it is necessary to design and set the goals of both employees and employers in a similar and symbiotic way. Regarding the AI/business expert group's request to support the firm's sustainable growth with proper compensation as the firm grows, employers (P25, P26, P30) expressed their desire to build the firm's vision, form a social consensus, and provide compensation to dedicated workers. P50 (AI/business expert) said,

> "Sustainable growth should be the ultimate direction we aim for together. The firm exists for its own profit and growth. To the employees, the job could be a means of maximizing their personal growth and [financial or social] benefits in life. They [i.e., the firm and its employees] should align their interests and goals in the same direction so that they can both grow. I think pursuing a sustainable way by satisfying both the firm and employees should be the objective of achieving fairness."

Building an organizational system that allows openness and mutual surveillance may be the first step toward navigating and attempting to resolve the tensions surrounding fairness in HRM. Demand for such a system may be more prevalent/critical in the HRM context compared with, for example, the context of enterprises' AI product development for sales [53]. Madaio et al. [53] have already raised the problems of 1) having no formal process within a firm for utilizing fairness checklists and 2) the typical organizational culture that omits the complex procedures and prioritizes "moving fast." In a similar line of research [10, 79, 85, 98], scholars have also claimed the lack of a mechanism to incorporate fairness feedbacks and to resolve disagreements. Thus, our findings offer academia/industry a good starting point as a critical design agenda: how can firms build an organizational system (e.g., auditing) that deliberately identifies "good tensions [53]" in advance and further adjusts and compensates employees when problems emerge from them?

4.2.2 Trade-offs Surrounding the Accuracy of AI.

4.2.2.1 Tension: "Accurate AI comes at a price." There were conflicting claims about accuracy among participants, because accuracy enhancement often demands sacrifice. Although such findings are broadly in line with prior works [10, 24, 79, 85] that emphasize challenges in meeting and/or communicating model accuracy, our study features the unique characteristics of the HRM context: 1) each stakeholder's purpose and impact of using AI in HRM is unique and 2) the socially-situated HRM context fuels conflicts on the accuracy of AI among stakeholder groups.

Algorithmic unfairness: "It's true that including discriminatory features can increase predictive performance." From the perspective of *AI experts (who are often creators of AI)*, time tracking workers' behavior may be required to collect more data (i.e., privacy invasion), or unfair features may be used to increase predictive power (i.e. algorithmic unfairness). Especially, the expert group pointed out important technical issues that other stakeholders need to be aware of: 1) Methodologically, the imperfection of data often causes algorithms to discriminate against a group based on their gender, race, ethnicity, etc. (i.e., algorithmic unfairness). 2) To solve this problem, it is possible to exclude sensitive features (e.g., gender, race) from predictive models or collect more data. 3) Regardless of its virtue, including sensitive/discriminatory features into a machine learning model when data is lacking could increase its predictive performance (i.e., algorithmic unfairness). P26 (AI expert and employer) and P45 (AI expert) gave an example,

> "When real causal features (e.g., social capital, education) are unavailable and these unavailable causal features are correlated with some available sensitive features (e.g., gender, ethnicity), including the sensitive features can increase predictive performance (e.g., the accuracy of predicting income). Thus, in real-world situations, there are often trade-offs between algorithmic fairness and the performance of AI models."

Privacy invasion: "You want me to go naked for a fair evaluation?" As other participants listened to the experts' explanations, both between and within stakeholder groups, there were disagreements about using sensitive/unfair information for increased accuracy. Especially, most of the *employees* strongly asserted that the firm should not use such sensitive data. This is because such harsh exploitation of data is perceived as a surveillance tool beyond mere privacy invasion that leads to severe inhumanity. Moreover, the employee group strongly refused to accept time tracking by the firm, even if it could allow more objective evaluations. P02 (employee) mentioned,

> "Are you saying I should sacrifice my privacy, just to achieve AI's accuracy? You mean AI will track me to figure out my behavior, like what I'm doing while sitting or standing, or who I'm chatting with or emailing? That's like stripping me naked and forcing me to walk around undressed! That can never happen. That would cause an uproar, and a war would begin. I would rather get an unobjective, inaccurate, bad, and unfair evaluation."

Similarly, P05 (employee) mentioned, "I really don't think that AI could achieve fairness even if it used all of our sensitive information [via] a position tracking system, email, or whatever. The idea that the firm could achieve accuracy or fairness through AI is just an illusion."

In contrast, the *HR/employers group* was more open-minded to micromanagement by advanced technologies that supercharge surveillance. This group believed that such surveillance would benefit hardworking employees by enabling a fairer performance assessment. For instance, P30 (employer) said, "Who would want unfocused workers? Some workers go out to drink water and smoke or go to the restroom every minute, whereas there are excellently capable workers with great work ethic and integrity."

Surprisingly, some HR managers and employers (N = 9) suggested using sensitive information to better predict employees' leave or time off from work or to select the right workers for promotion, disclosing that most firms would already implicitly reflect workers' sensitive information in important decision-making. P33 (HR) said, "Almost every firm has its own confidential list of major talent. It's something like profiling." Also, while some employees (N = 12) were already aware that firms implicitly use workers' sensitive information for critical decision-making, they expressed unpleasant feelings, saying "I really don't think that is necessary" (P08). Regarding this issue, other professional stakeholders gave examples to explain the inevitable reasons for using sensitive features. For instance, P37 (HR) argued,

> "It depends on the purpose of the firm or its situation. For example, a company that has a competitive edge in terms of managing and retaining top talent may want to catch early signals of someone's willingness to leave the company. Isn't it impossible to completely exclude all of the data just because it's sensitive? Deciding whether to include sensitive variables or not could change depending on the situation."

Inefficiency: "I'm not an AI-sitter... Shouldn't adopting AI reduce my time and effort?" We identified that there is a trade-off between efforts in increasing accuracy and the inefficiency it causes. The employee group stressed the importance of accuracy in their performance evaluations, whereas HR/employers were concerned with the inefficiency (i.e., time and effort) of increasing accuracy. For instance, P41 (HR) said, "The weakness of AI is the time spent to modify, improve, and de-bias the algorithms. If the stakeholders have to put too much time and effort in figuring out whether the algorithms' results are meaningful, the AI could just be buried as a result of the countless days spent exploring it." Similarly, P33 (HR) mentioned, "As a metaphor, AI in HR doesn't have to play chess perfectly or make all the right moves. It's enough if it provides a good match for us [i.e., HR teams], because we will judge and make moves in the end."

4.2.2.2 Design Insights: "Does AI have to become evil to be perfect? Can't humans compensate for AI's imperfections?" In the workshops, both the participants and moderator presented several possible situations where sensitive variables may be needed, and all of the participants recognized the problem of using sensitive and controversial variables (especially race, gender, and ethnicity) to increase the model's accuracy for work performance evaluations. In response, the design ideas that participants shared and agreed upon the need for are 1) the *discretion of human decision makers (i.e., HR)* and 2) *informed consent backed by persuasive reasons.*

Human discretion. Instead of using sensitive information in an unethical way to perfect scarce data, both HR teams and employees requested that an imperfect AI be complemented by the HR team's discretion and intervention. P02 (employee) said, "If the firm makes sure that AI doesn't utilize sensitive or personal data and guarantees that HR makes up for the inaccurate results of AI, I think I would accept a trial adoption of AI." Similarly, P48 (AI/business expert) said, "I recommend the controversial data not be used even if it could undermine the model's accuracy. If such sensitive factors need to be reflected in AI models, I think the human HR team should directly validate and readjust AI's results based on clear standards. I believe AI should be used as an assistive tool in this case." Another AI expert (P47) also showed a cautious attitude saying, "Rather than including all of the data just to make a highly accurate model, we need to cautiously approach it by following the process manually and confirming whether it has a causal relationship."

These findings align with prior works [71, 75, 79, 98], which emphasize AI design that allows human discretion. In high-stakes public sectors where the purpose of AI is for common good, stakeholders in practice often utilize sensitive attributes (e.g., race and age [10]) if the benefit outweighs the cost or situationally decide whether to use sensitive attributes by comparing models made with/without the sensitive attributes [85]. However, because using such attributes can lead to unintentional and/or systemic bias, recent research [9, 21, 37] has suggested and legal regulations (e.g., [108]) have required that sensitive attributes be deliberately excluded (i.e., "fairness through unawareness" [9, 38]). Similarly in the context of HRM, workers expect even *more* objectivity and fairness from AI adoption to HRM and have a strong ethical value in common: people should be evaluated by their abilities and not by race/gender/age. Enabling human bias by using sensitive attributes did not make sense to our participants and further aroused their antipathy. For such reasons, our participants agreed that sensitive information should be excluded, preferring human intervention and discretion.

However, recent studies [9, 21, 37, 56] show that the exclusion of sensitive attributes leads to other unintentional biases in algorithms or exacerbates existing biases by introducing a fallacy of "fairness through unawareness"; that is, although the sensitive attributes (e.g., gender and race) are protected, other unprotected variables (e.g., college attended, hometown, or various resume indicators) may still be highly correlated with the protected attributes [14, 84]. This raises important questions for future research: 1) how can we approach such issues that newly emerge due to the deliberate exclusion of sensitive attributes? 2) If the exclusion of such controversial variables leads to a severe lack of data, what other data should be collected to better evaluate employees?

Informed consent. Despite the broad consensus to not use sensitive information, many HR/employers mentioned the possibility that the use of such data may be unavoidably necessary if it results in better predictions for a critical problem. They (P27, P30, P33, P37, P39) presented several examples, such as "predicting leave or time off from work," "promoting someone to an executive position," "choosing to send a representative/resident abroad," etc. P30 (employer) said, "Even we [i.e., humans] suppose that a worker who is pregnant might have some time off or leave from work. It's kind of a tendency. We cannot call everything a bias. The employers should be prepared for the situation in which the worker suddenly leaves." In the exceptional case that sensitive information would be necessary, other AI/business experts suggested that firms obtain informed consent (e.g., from employees) by providing enough persuasive reasons for the use of private information. Additionally, many employees responded that because using such sensitive information is displeasing, the firm should take tremendous responsibility before/while/after using such information. P48 (AI/business expert) agreed and said,

> "If it is essential due to the unique characteristics of the domain, explain the details [i.e., purpose and criteria of data collection] and persuade employees to obtain agreement prior [to its use]. The most important thing is to acquire informed consent in advance. To be persuasive, the firm should give enough reasons with options, such as a legal safety net or financial compensation."

Also, in preparation for biased predictions, the firm should confirm the AI's result with the employee directly. P37 (HR) reported, "When we [i.e., HR] make a critical decision, we always acknowledge that the impact of a wrong prediction could be huge and often irreversible. So, we directly ask the employee's opinion again before carrying out the decision. The AI should do the same when making critical decisions based on sensitive information." Such findings present the following design agenda: What safety net features (e.g., compensation) would positively affect employees' willingness to grant informed consent and adopt AI for HRM?

4.2.3 Transparency of the Algorithm and the Decision Process.

4.2.3.1 Tension: "There is a fine line between a white box and Pandora's box". Most employees strongly asserted that firms should provide transparency for both algorithms and evaluation procedures, believing that transparency could increase their sense of fairness regarding AI's results. Specifically, employees wanted information about how the algorithms work (e.g., a list of features used for prediction, the inner logic of algorithms) and what the evaluation procedure looks like (e.g., the scope of AI adoption, who has the authority to change the final decision). However, in a sociotechnical context, we found that providing transparency can result in a variety of potential problems and create tight tensions among stakeholders. P40 (HR) said, "That's why it's such an uneasy situation. On one side, it seems reasonable to open the algorithm and decision process to some degree. But if critical problems exist, it could put the firm into a very risky position. But if we hide it from them too much, other conflicts also occur."

Gaming AI: "The workers will study how to better abuse AI rather than work hard." While many employees demanded detailed transparency toward AI, HR/employers reported that they could not comply with these demands. The major reason for the rejection was attributed to the inherent trade-off that exists concerning transparency: the more transparent the AI is, the more employees game [i.e., trick] or attack the AI antagonistically. P41 (HR) said,

> "For example, if an HR team reveals that we will check emails, it becomes a kind of game. There are strategies and attacks in the game. To win this game, you might write an email every five minutes and CC people. Or you might intentionally send an email to each of the ten persons by splitting it up, even though you could just send it all at once by CCing all of them. This is not the right direction that the firm wants."

Interestingly, even though most employees demanded a high level of transparency in the evaluation process, they were also concerned about the possibility of other employees gaming the AI. P02 (employee) said,

> "I think giving all of the AI's elements or logic to the employees is unrealistic, and opening everything without carefully thinking could aggravate the situation. Let's say, the firm revealed the fact that AI

> counts how many reports a worker creates and reflects it in the work performance evaluation. Employees might make several [low quality] duplicates rather than make one well-written report. So, the way [elements] are revealed should be carefully designed to encourage workers to work in a valuable and productive way."

Threat to enterprise: "Why should AI disclose details that we don't even reveal in the current evaluation system?" During the sessions, the idea of revealing the AI's details brought a tense atmosphere between employees and employers. While many employees wanted detailed information on the work performance evaluation system, HR/employers were worried about the widespread ramifications of complete transparency (i.e., decreased authority of the enterprise and knowledge leaks). P01 (employee) strongly asserted that "I can agree to adopt AI for work performance evaluations, but there is one thing that I want from my company. The firm must clearly explain how the AI produced the result. If it's unreasonable, I can't accept it. I will leave the company." Regarding such employee's attitudes, P37 (HR) and P40 (HR) explained that workers these days make unprecedented requests for the procedural transparency of HR's evaluations and P37 added,

> "The members of organizations nowadays want more detailed data, like 'Why did I get this grade in my performance evaluation?' 'Why did I receive this amount of compensation as a performance-based bonus?' 'How is the incentive calculated?' and so on. Many employees want very detailed feedback on how the evaluation system is structured. We analyzed the phenomenon, and it is a characteristic of millennials and Generation Z. In the past, we gave workers a grade like A, B, C, or D with simple feedback, saying, 'Oh, you should complement this and that.' But now employees want the process that produced the A, B, C, or D grade."

The employers and HR managers were concerned with a decline in the enterprise's authority. For example, P40 (HR) said,

> "AI's transparency could open up a can of worms. A worker could make harsh complaints, according to their level of receptiveness [toward AI]. Like any other company, our companies are also very sensitive about evaluation systems and completely opening them. If we found critical problems within the algorithms after opening it too much, it could significantly harm the firm."

Many HR participants expressed strong concerns about leaks of the firm's confidential data, because information on their workers as well as performance evaluations are top secret. P31 said, "What if this top-secret information is leaked to other rival companies? . . . If we document the input or logic of our algorithms well and open them, they become vulnerable to security attacks." Similarly, P37 said,

> "The information on our workers is indeed our firm's credentials. Because we are one of the top companies that rank first or second, back and forth, in this field, we have world-class engineers and R&D researchers. Other global companies aggressively fight to dig up our workers' information to scout them by giving them a big pay raise. Actually, there was a hacking attack we found a few months ago."

4.2.3.2 Design Insights: "Is obsessing over algorithm transparency unavoidable?" While transparency may improve users' perception of AI's fairness, being completely transparent can also cause unexpected problems (e.g., confusing employees, gaming the system, unexpected conflicts and costs, breach of know-how). Thus, participants suggested that designers focus on transparency to 1) *untangle the optimal level of transparency and approach it in a social/*systemic *way* and 2) *lead employees' gaming behavior in a positive direction.*

Designing the optimal level of transparency and social/systemic transparency. While being aware of the critical problems that could emerge from full transparency, participants also admitted that most stakeholders (i.e., employers, HR managers, and employees) would not be able to perfectly grasp every detailed aspect of the algorithms. P45 (AI expert) said, "Even if we disclose the detailed algorithms, would other stakeholders be able to understand them? Some of them could be coded in a binary system, and it might be indecipherable to general people." In this regard, the expert group proposed to partially disclose the algorithms' inner logic and complement the lack of transparency with alternatives, such as better process transparency or social and systemic solutions. The HR/employers group (N = 13) approved of disclosing a clear but safe level of process transparency so that the firm can manage its side effects. P40 (HR) reported,

> "Despite the issues that emerge by disclosing the AI's inner process, it would be necessary to clearly share some parts of the AI and its process so that the firm can manage the side effects. At the same time, the firm should prepare to manage a higher level of transparency, because in the future, workers' demand to know the AI's logic or process will eventually grow. Recently, in many other firms, employees demanded that the firm transparently disclose their net profit and reflect it in the performance-based pay system. I recently planned and redesigned our payment system, and we are clarifying what percentage of the firm's net profit will be given to the workers. This is a drastic change compared with tradition, but in the future, there might be more needs for such transparency."

Experts said that "*trustworthy systems are not always transparent systems*" and suggested to find a way to mitigate employees' difficulties in understanding (or negative reaction to a lack of) transparency by considering an organizational and systemic approach. Many participants proposed that an external audit could also resolve transparency issues that cannot be perfectly resolved within the firm. Employees also agreed to the idea of adopting a fair external auditing system as long as the results are explained upon completion. P49 (AI/business expert) said,

> "We should audit AI through an external specialized agency. Instead of directly disclosing the algorithms, this type of indirect but systemic safety net will help

> increase trust among members with fewer side effects. For instance, when we check the firm's financial position, the individuals can't access or examine transaction histories and account books. Individuals judge it through refined financial or audit reports that the firm publishes. Likewise, I think we don't need to judge AI by looking into the raw algorithm itself."

With these perspectives in mind, how can we find the optimal level of transparency that meets employees' desire to know without harming the firm (e.g., knowledge leaks)?

Designing transparency to lead employees' gaming behavior in a positive direction. Previous research in high-stakes public sectors has pointed out the concerns with transparency leading to humans' gaming behavior [71, 75, 79], mentioning there is little work that attempts to solve the problem [85]. In such domains where gaming behavior leads to negative consequence only (e.g., parents game the student-school assignment algorithm to enroll their child in a desirable school [71]), scholars called for future studies that focus on 1) how to hinder such user behavior [71, 85] or 2) how to directly include stakeholders' values [20] instead of making them game the system [10, 75, 98]. However, in the unique HRM context, transparency does not need to cause negative consequences only; while transparency makes the system vulnerable to employees' gaming, it could also be used to shepherd them by publicizing a visible evaluation index. Because motivating workers through right evaluation is stakeholders' unanimous goal, designing transparency that points employees in a positive direction would be a new and important design agenda.

Regardless of the group's characteristics, most participants believed that a proper degree of transparency could ultimately motivate workers and align them to the goals of the company. Surprisingly, employers (i.e., P25 and P26 who own AI-based products as AI experts) as well as many AI/business experts (P42, P45, P49, P50) pointed out, "Don't you think the fact that people game algorithms means that the algorithms are poorly designed?*"* and employees (N = 8) mentioned that gaming itself should be well-designed to motivate them to set and achieve goals. P45 (AI experts) further added,

> "Even establishing evaluation standards or an index can have an incentive effect, by informing or motivating the employees to keep working hard to achieve the standard. I don't think showing transparency itself is a problem; an algorithm that has a hole is the problem. Realistically, there is no such perfect rule or standard in this world, so I'm quite negative about showing the process, logic, standards, or everything completely. It should be disclosed to the degree that the employees' gaming won't produce unexpected and serious side effects."

P49 (AI/business experts) mentioned,

> "Gaming behavior that goes against the rules, that's human nature. We cannot criticize them with moral standards. Also, isn't that behavior already prevalent in the workplace? We try to gain favor with our superiors and want to get good peer reviews from our colleagues. Rather than making it an issue, we need to direct their gaming behavior to benefit the enterprise. In other words, I think it is essential to design transparency well. In that respect, I'm opposed to whitebox AI. I think it will make people behave inefficiently."

4.2.4 Interpretability of the Algorithmic Decision.

4.2.4.1 Tension: "Interpretability is a double-edged sword." Several AI designers mentioned that in the HRM context, the interpretability of AI could increase employee's trust and perception of fairness by explaining why AI derived the evaluation results in *human-understandable terms*. Such explanations could also help HR teams or AI designers identify potential algorithmic biases/errors in the system. When the explanations are unreasonable, people could cast doubt on the algorithm's inner procedures. Despite these benefits, we found that providing interpretability to employees can adversely affect employees' perception of fairness due to the technical limitations (i.e., the interpretation is unreliable and does not guarantee causal insights) and social contexts of HRM (i.e., not only HR but also employees could see the algorithm's biases/errors). Several HR managers (P32, P37, P40, P41) emphasized the existence of "tech-savvy" and "clever" workers in the firm saying, "In a tech company like ours, half of the workers are technology developers. They would instinctively feel something is wrong when interpretability is given." However, even if AI derives correct interpretability, harsh facts could offend the employees and result in a negative algorithmic user experience.

Merely correlation: "Even chocolate consumption is highly correlated with the number of Nobel Prize winners." Most employees strongly asserted that AI should explain in detail the reasons behind the performance evaluation result. Some employees had already heard about the interpretability of AI and demanded it for designing fair AI in HRM. However, the expert group warned that HR/employers should be cautious when providing interpretations to employees. The predictive models are inherently based on correlation; thus, the interpretation of them is also correlative, which does not imply a causal relationship. This means that it is difficult to directly use AI's interpretations to know why employees got a good or bad result or to teach employees how to improve in the future. In this context, P50 (AI/business expert) said,

> "The current interpretability models cannot guarantee the causality of the features. More specifically, to say that through AI's explanation [e.g., coefficient] a certain independent variable has a causal relationship with a dependent variable [e.g., work performance] in predictive models, specific conditions must be satisfied. The independent variable should not be correlated with other independent variables or unobserved confounders, but this is often violated in the reality where data is incomplete. In other words, some variables may have a causal relationship with dependent variables and other variables may not, but we don't know which one has the causal relationship. So, firms shouldn't just disclose interpretability irresponsibly. Let me give a simple example. Do you know the fact that the chocolate consumption of each country is

> highly correlated with its number of Nobel Prize winners? This means if we train an AI model to predict the number of Nobel Prize winners, AI will interpret chocolate consumption as one of the most important factors. But, can we increase the number of Nobel Prize winners by letting people consume more chocolate? The idea is ridiculous, but it is a good example to show the pitfall of [correlation-based] interpretability. If we interpret AI's interpretations without any validation from domain/statistical expertise, we will sometimes draw totally wrong conclusions or future directions based on the evaluation results."

Revealing AI's weaknesses: "Interpretability will betray AI's bias or errors to employees." Similar to the transparency issue, the employee group desired interpretable results as much as possible while both the HR/employer and AI expert groups responded negatively to interpretability. Specifically, both professional groups (i.e., AI/business experts and HR/employers) shared the same idea: just as interpretability helps experts (e.g., data analysts) identify errors and biases in algorithms, it can also help employees find defects in AI's evaluation results. While the AI expert group was more focused on the inherent weaknesses of the current ML/DL/interpretability algorithms, the HR/employer group was more concerned with the aftereffects of even one small error in interpretability. For example, P50 (AI/business expert) mentioned,

> "Methodologically speaking, it is extremely hard for the current algorithms to derive reliable and trustworthy interpretations. Even if we use the same dataset for predictions, different algorithms typically derive different results. Even worse, the interpretation results also vary depending on different interpretability algorithms. Thus, some errors are unavoidable in interpretation, and these errors become larger when more detailed explanations, such as local interpretability—which are explanations [e.g., feature importance] of individual instances—are sought. In this situation, disclosing detailed interpretations can risk undermining employees' trust and perceived fairness of AI decisions."

Regarding HR/employers, P40 (HR) said,

> "If a smart worker raises issues, saying 'Oh, I want to know the reason for the AI's decision.' The HR team might try to explain the inaccurate results by any means—even with their own interpretation. In that case, the smart worker could brilliantly catch the weak points of the AI and squarely confront the HR team and the firm by resisting the evaluation."

Inhumanity: "How dare the AI say that I'm bad, even if I am really bad? It hurts." Many employees mentioned that results with too much objectivity and interpretability could hurt their feelings, and the HR group agreed with those opinions. For example, P17 (employee) said, "I like that AI could explain why I received a certain result, but if the AI makes cutting remarks, I'll feel intimidated. I mean, who would love hyperrealistic [feedback] about their competence, especially when it's bad? I'll feel small." P37 (HR) agreed with the workers' perspective and described her vivid HR experience by giving an example:

> "Generally, most of the firms do not give workers detailed or concrete explanations about work performance evaluations, the same as my firm. We provide rough feedback with a bold summary. There are always a few employees who cannot accept the result and come up to the HR department to request a more detailed explanation. Then, we almost completely disclose why this person received that result and show how well he achieved performance in contrast to the goal that he set at the beginning of the year—the 360-degree feedback from managers, peers, subordinates, and everything. Do you know what's funny? I have never seen a single guy come back to HR to complain again for the rest of his career. This means they got really hurt after seeing the harsh facts."

On the other hand, some employees and employers welcomed AI's interpretability in terms of providing objective feedback that is helpful for employees' improvement. For instance, P13 (employee) said, "Oh, that's what I want. Let's objectively see whether I deserve less than that guy. AI wouldn't inhibit my promotion the way humans discriminate against me just because I'm female. I like more objective data and results, and I want to see how the evaluations made by humans or AI differ."

4.2.4.2 Design Insights: "Be careful when interpreting predictive results." In the results of prior works on AI's application in other domains, interpretability is generally demanded and benefits various stakeholders (e.g., end-users, AI designers, etc.) [30]. However, our findings are uniquely HRM domain-specific and counter-intuitive. Specifically, studies in healthcare have shown that stakeholders (i.e., clinicians and patients) desire clear and detailed explanations of the models from decision support tools [28, 81, 93]. In auto-driving, the interpretability features of an auto-driving decision significantly mitigated the driver's negative auto-driving experience [76]. However, in the peculiar HRM context where the meaningfulness of one's work/identity is considered important, harsh facts (i.e., criticism) of clear interpretability derived from machine evaluation could contradictorily harm the employees. Additionally, because conflicts of interests among large stakeholder groups exist in HRM, interpretation of unfair/biased algorithmic decisions could undermine the trust and perceived fairness of many stakeholders (e.g., employees) at once. Thus, interpretability should be treated more carefully. Our findings offer two important considerations for future research: How can we select and design interpretability to minimize the unintentional disclosure of unfair/biased algorithmic decisions? How can interpretability be achieved without causing offense? As a first step to resolve such issues, participants suggested that a high level of expertise (e.g., domain knowledge) and advanced modeling (e.g., causal inference models) be used along with AI's interpretation. Second, participants agreed that interpretability should be designed to help enhance employees' competence rather than being used to hurt employees with objective results and plain facts.

Selected interpretation. After being informed of such limitations, most stakeholders agreed that AI's interpretation should be provided to the employees at an optimal level, after experts review and confirm it to be reasonable. For example, P48 (AI/business expert) said, "I think the firm should disclose only the validated interpretations after the AI designers verify the models by closely collaborating with HR experts in the firm. It is unavoidable that in many cases, algorithms produce imperfect or wrong interpretations." Similarly, P36 (HR) added, "We should sort out the information types and degrees of disclosure. Although we should give as much information as needed to the managers, it seems safe to inform employees only after selecting finely processed and essential information."

Notably, many employees who wanted exhaustive interpretation in the beginning changed their minds once they understood that model interpretability is not a simple task. Experts also raised concerns regarding model interpretability and suggested two practical design ideas to better convey interpretability: 1) create interpretability tools that can clarify AI's decisions through simple and aggregated explanations (e.g., global interpretability), which are more reliable and easily validated by human experts than complex explanations (e.g., local interpretability); and 2) establish an AI-specific department within the firm that is dedicated to communicating AI's interpretations. Furthermore, if employees request a more detailed interpretation, experts suggested letting the AI-specific department explain and translate it. Indeed, many employees (N = 12) wondered if they could understand the statistical results an AI explains, and they strongly agreed with the experts' suggestion that AI experts within the firm should respond to workers' complaints/inquiries about results.

While the AI experts recommended that such correlation-based interpretations be cautiously used by firms, they insisted that AI designers be devoted to proving causality and applying it to the HRM context. They believed that finding causality is the most important task, because it could give more actionable advice to both employees and employers. P44 (AI expert) said, "Meanwhile, if technologically possible, we should adopt and apply a causal model to better derive actionable strategies from the interpretations of AI. To do so, firms need to invest in causal methodologies for the long term and build know-how, which will greatly increase the firm's competitive advantages."

Interpretability should not be used to attack, but to benefit employees. Many employees (N = 15) envisioned interactive interpretability that relays a worker's current performance level, including weaknesses and strengths, a comparison of their evaluation with other workers, and a suggested direction for future improvement. Among these three features, AI designers and HR managers showed special interest in how to provide beneficial suggestions for future improvement. Employers, employees, and AI/business experts (N = 39) affirmed that it is inhumane or unethical to disclose harsh AI-driven interpretability results to workers; thus, they strongly agreed that interpretability should not be used to criticize employees with purely objective results. Participants thought that the firm should design interpretability as a means of motivating both employers and employees by showing them how to improve in the future. P45 (AI expert) said, "We should disclose with a degree of interpretability that informs workers of their current status and future direction. Technically speaking, we're letting workers know *a gradient of direction*. And the direction should be thick and bold rather than too detailed."

4.2.5 *Trade-off Between Productivity and Inhumanity.*

4.2.5.1 Tension: "You could be the inefficiency that AI wants to improve or eliminate." Most of the participants emphasized the importance of increasing productivity in their workplaces. Employers wanted to improve both their evaluation system and their employees. Both employees and HR managers hoped AI could improve their productivity. However, the more AI is used to raise productivity, the more side-effects there are. For instance, if employers exploit AI in HRM, it could become a means of micromanaging workers or forcing meritocracy. Or, if AI is used as a supportive tool to perform repetitive, dangerous, or difficult tasks, it will replace workers and remove their means to make a living. In this section, we describe each of the tensions that occur due to this trade-off.

Micromanagement: "If you can measure it, you can monitor and surveil it." Regardless of the stakeholder group, most of the participants believed AI evaluation would be most suitable for menial labor (i.e., manufacturing/operational work), physical labor (i.e., construction), and sales. These types of work can easily be quantified in terms of performance goals, actual outcomes, comparisons of records, etc. Such clear performance indexes enable the calculation of objective results in numbers. P49 (AI/business expert) said, "Without considering ethics, if we define productivity as a number of products that a production worker makes per hour, it is indeed possible to automatically or precisely measure that. AI could even squeeze productivity out of the worker to meet the quota." Many employees (N = 14) agreed that AI could easily be used to evaluate manufacturing/production workers.

No matter what work AI is adapted for (i.e., office/production job or knowledge work), most employees (N = 19) insisted that AI should not be used to pressure workers through micromanagement and prioritizing productivity. P02 (employee) added, "The firm might adopt AI like a whip [i.e., to redouble workers' efforts], but that would be like slavery. If a human has a whip, I can take my eye off the ball for one minute. But if a robot has a whip, it will lash me even when I take one second to breathe. Who would like it if they say they will treat me like a slave?" P12 (employee) clearly distinguished micromanagement from enhancing productivity by giving a real-world example,

> "Our firm already has a lot of monitoring systems to increase efficiency or productivity, like a speed gate system at the entrance and an automatic attendance check via my computer. There are CCTVs everywhere inside the building of my firm. These monitoring systems are prepared in case something happens or to make our work efficient. Like nobody wants to manually register their name to prove their attendance or to gain entrance into the building. However, now you're saying all of those systems would be used to evaluate my performance. That's a whole different story. Precisely speaking, people accepted the installation of CCTVs at our workplace just in case something bad happens and we need to check it. We didn't accept them as a surveillance tool."

Meritocracy: "There will inevitably be workers that are naturally selected or eliminated by the harsh competition of work performance." Our findings show that a strong tension exists between employees and employers regarding meritocracy. P30 (employer) said, "The workplace is not a playground. I think workers need competition. Don't employers or enterprises compete with others?" However, P02 (employee) responded,

> "Whether it's intentional or not, performance evaluations are inherently like Pareto's Law. The extraordinary workers in the top ten percent always do well, but if you belong to the low or ordinary ninety percent like me, it becomes a matter of survival. Adopting AI would drive the lower ranks into extreme competition—simply for survival, for a living."

As an example of current evaluation systems, P03 (employee) shared,

> "Within the headquarters of planning, there are four different teams. The firm makes workers who hold the same position in each team compete with each other. If the evaluations were made by AI, people would more likely see that guy's score in the same position and compare the results, assuming that the AI's result is more objective. . . . Also, let's say, more than two people work on a simple task like making a report, and the individuals' efforts were mixed together. How are the participants gonna share the credit with each other and how will AI split it up? Everybody could be happy if it is split evenly, but if not, it becomes really complicated and causes strife."

Role of AI: "Supporter? Or job exterminator?" We found that the adoption of AI could serve to support and/or eliminate workers for two reasons: 1) AI's automation ability could replace existing jobs, and 2) by supercharging surveillance, AI could foster higher demand for employees' productivity or use high standards to fire "low" productivity workers. During the sessions, individual employees showed varying points of view, including ambivalence, toward AI taking on tasks that humans could do. For instance, P18 (employee) said, "Once, I thought it would be good if AI assists with some of the annoying tasks that humans dislike, but that annoying work could be someone's job." P09 (employee) reported, "If AI supports some tedious tasks that HR managers do, one day AI might replace all of the HR jobs, don't you think? It's not just our [i.e., employees] issue but HR should also be worried about it." P33 (HR) was also worried about job loss saying, "Someday, AI might take my HR job too." On the other hand, some employers and experts showed less concern about eliminating a human role. P30 (employer) said, "I mean, I'm not rooting for AI to annihilate all human jobs. But don't you think some jobs will naturally disappear? Can the world go against the trend?" Similarly, P49 (AI/business expert) said, "I assume jobs that are easily monitored and quantified are the ones that will first be replaced by machines. I personally believe such jobs harshly pushed [i.e., surveilled or evaluated] by machines should disappear, and humans can do more creative jobs."

Most of the participants mentioned that AI's adoption to HRM would naturally cause the loss of human jobs. Regarding the issue, the expert group asserted that the firm should not assume an indifferent attitude and fail to prepare. By focusing on how to minimize the impact, the firm should gradually proceed with job dismissals and give enough time to workers to think ahead and be prepared. At the same time, firms have to prepare proactive plans about how to re-train and re-allocate workers who lose their jobs due to AI. P49 (AI/business expert) commented,

> "The dilemma between productivity vs. inhumanity is a way more important agenda and could become metadiscourse that goes beyond the mere fairness issue of AI. For example, Frederick Taylor, who is known as the father of industrial engineering, first quantified the productivity of factory workers through objective figures. He made a starting point by assuring significant productivity through efficiency in manufacturing. And it indeed resulted in unimaginable growth. However, around that time, enormous conflicts emerged as labor unions collapsed. I think that AI may also be following the same path as the past. Although the adoption of AI would eventually bring about enormous productivity to the firm, it is inevitable that there will be workers who are naturally selected or eliminated by the harsh competition for work performance. The firm has to prepare a clear plan for how to retrain and reallocate such eliminated workers."

4.2.5.2 Design Insights: "Toward facilitating the sustainable growth of the organization." During our sessions, many employees were concerned that AI in HRM could be exploited to improve worker's productivity through advanced monitoring and surveillance technologies. Thus, in our design sessions, this focus was actively discussed as a key topic among stakeholders. Although employees and HR/employers fiercely disagreed at the beginning of the discussion, most stakeholders reached the consensus that the adoption of AI should aim to achieve the organization's sustainable growth. Based on this consensus, participants suggested that AI should be used to amplify human ingenuity and requested proactive plans to minimize the negative impacts of AI's adoption in workplaces.

Using AI to amplify human ingenuity instead of micromanaging workers for productivity. Although several employers first admired AI's capabilities to calculate employees' actual work performance, they slightly changed their attitude after listening to other stakeholders (e.g., AI/business experts, HR managers) who were concerned with the unexpected consequences of micromanaging employees. P50 (AI/business expert) mentioned, "Is it sustainable to push workers to extremes to foster productivity? There will be unobserved costs like excellent/dedicated workers' leave, workers' passive attitudes, and reduced well-being, satisfaction, or happiness. The AI adoption model should have a sustainable form by considering the workers' well-being." In response to this opinion, P07 (employee) said, "I like the idea of adopting AI to better compensate workers. If I make more sales, I would get more incentives. However, the AI shouldn't keep notifying me to work harder by showing my poor performance every minute." All of the employers agreed that micromanaging workers through AI could be unethical and that AI should be used for more sustainable values.

Instead of using AI to micromanage workers, many participants suggested using it as a supportive and augmented tool to automate repetitive or difficult tasks. They believed it could enable humans to focus on other, more human-oriented tasks. P48 (AI/business expert) emphasized, "AI should lower workers' workloads and time by playing a supportive role. It should help humans' decision-making so that they can concentrate on where human knowledge is truly needed." Furthermore, the HR group brought up several tedious but time-consuming tasks (e.g., handling evaluation indexes) that AI could replace. Specifically, P41 (HR) said, "I prefer AI monitoring workers' job performance way more than AI judging and making decisions on its own. I'll only refer to it, because I have my experience and intuition. AI shouldn't care too much about producing accurate results but should create a good environment for evaluation." P40 (HR) further described,

> "An evaluation index is comprised of qualitative and quantitative assessments. I think AI could intervene in the process of creating and handling the quantitative assessment index because it's unnecessary for humans to enter the evaluation software system and manually input data. When tangible performance data, like sales, is released, the AI should automatically extract such data, match it to others, and build a concrete evaluation index. I want AI to proactively do some of that work that we have to do."

Such findings echo what roles technology should play to support workers—human ingenuity (e.g., augmented creativity) and productivity [6, 33, 36, 54, 95]—to which the HCI community has long put enormous interest and effort in research and practice. However, while prior works have examined the impact of tracking technologies on productivity enhancement [34], our findings show that new challenges emerge in AI's adoption to the HRM context. Although people expressed that a self-monitoring system would be useful to check their performance and increase productivity on their own [34], they disliked the use of AI by a firm to manage and increase their productivity. Since accurately measuring work performance amidst a complex social context (e.g., micromanagement) can aggravate stakeholders' tensions, how can we optimize the model accuracy without requiring sacrifice (i.e., micromanagement through privacy invasion) from workers? Can we design supportive AI tools that replace tedious/repetitive HR tasks and foster HRM's improved decision-making and human ingenuity?

5 DISCUSSION

5.1 A First Step Toward Stakeholder-Centered Fair AI Design in HRM

Our findings illustrate the common notion that designing AI is inherently difficult [48, 92], but it is messier when it comes to the HRM domain. In part, this is because of *the unique characteristic of HRM* where multiple key stakeholders inevitably engage in one ecosystem, including the consumer who decides to adopt AI for the firm (i.e., employer), the end-users who directly use AI (i.e., HR managers), the creators of AI (i.e., AI designers), and those most impacted by the AI (i.e., employees). Furthermore, *AI's innate imperfections and the trade-offs of AI design features* (e.g., accuracy) worsen the situation when combined with the complex social context where each group's needs and incentives conflict.

As a first step toward approaching such high-stakes AI design in HRM, we suggest the designers identify diverse stakeholders' tensions in advance by utilizing our method and agilely reiterating the process. Although our research is in line with prior works [23, 45, 53, 75, 79, 85, 98] that aimed to reflect diverse stakeholders' values, we focused more on ***deliberately eliciting potential tensions that occur when stakeholders' values conflict***. By identifying these challenges for AI design in HRM, both enterprises and researchers may be able to preemptively find and prepare for issues in the early AI design stage, instead of making hasty/premature design decisions that fail (e.g., the Amazon case and case studies of other domains [71, 85]).

Also, our research (both the list of tensions and method) not only contributes to the uniquely complex HRM domain, but more broadly to other high-stakes AI design research (e.g., public sectors) covered within HCI. Indeed, many HCI scholars clearly stated 1) the lack of research specialized to identify/resolve the tensions [10, 53, 85, 98] and 2) the gaps in existing UX research method for explicitly engaging diverse stakeholders, especially groups who have substantially less power [72], around AI fairness [10, 53]. For example, Zhu et al. [98] left these challenges and questions for future studies: "How do we appropriately aggregate their inputs to ensure algorithm justice and accountability?" and "How can we deal with value conflicts?" Likewise, Madaio et al. [53] called for new ways to "introduce good tensions into the AI development and deployment lifecycle so that they [i.e., stakeholders] can engage deeply with the complex, nuanced concept of fairness, as applied to AI systems" [53]. Our research directly responds to these calls by providing a list of tensions, stakeholder-centered design insights, and the method used to identify them. Importantly, we found that our iterative design method which utilized authoritative, competitive, and collaborative design promoted a good environment for those lacking AI literacy. At the start of the sessions, most participants were not aware of the concept of algorithmic (un)fairness. However, as participants passed through each session, they naturally questioned and learned technical issues regarding AI fairness (e.g., data imperfection such as class imbalance leading to bias/unfairness), and even flexibly changed their attitudes/opinions while debating resolutions for the tensions. Thus, we hope future research that aims to design AI and find tensions by directly engaging diverse groups (e.g., non-experts) could benefit from our research method.

5.2 Sociotechnical Perspective

As is often the case with AI, our results demonstrate that one simple solution does not fit all contexts [78, 93]. One AI design solution (i.e., transparency or interpretability) that works well in a specific domain may not be applicable to a complex HRM environment [78, 83] where various stakeholders exist as the end users, producers, and customers of AI [19]. It may be impossible or even problematic to solely reflect the needs of one arbitrary group (e.g., employees [42, 46, 69]) in designing AI for HRM. For example, many employees initially demanded that AI's inner procedures and logic be fully disclosed. However, after questioning whether it is the best solution and weighing other stakeholders' input, many of the employees changed their minds (i.e., after finding out it could also produce problems, such as gaming the AI and/or cognitive overload for themselves). However, the diverse participants agreed that partial transparency could be balanced by strengthening the social safety net; expanding social transparency was especially important to the employees. Specifically, both employees and employers paid special attention to establishing a social system (i.e., an external AI audit system) for mutual surveillance or to strengthen the social contract.

In recent years, researchers [1, 2, 15, 44, 45, 48, 69] have moved beyond viewing AI design factors (e.g., transparency, interpretability, and accuracy [94]) as a total solution to incorporating the social context (e.g., social transparency [15]) where such AI design factors are socially situated, formed, accepted, and improved. In this vein, Ehsan et al. [15] said, "the humanization of the process can also make the decision explainable to non-primary stakeholders in a way that technical transparency alone cannot achieve." Some scholars, by refuting the notion that showing mere transparency increases fairness, have also argued that such AI design factors are overhyped and/or could be accepted differently depending on the social context [5, 59].

Extending such prior works to HRM, we suggest that both researchers and practitioners expand the concept of social transparency and incorporate sociotechnical/organizational perspectives by 1) acknowledging potential tensions surrounding AI, 2) facilitating open discussions that pursue a consensus on how to meet each organization's characteristics, and 3) preparing a safety net, such as an external audit system. It is noteworthy that, although firms should not completely disclose transparency in the beginning, social transparency could be achieved by building a concrete social system in which multiple safety nets (e.g., external audits and a social contract) are prepared.

Finally, we outlined tensions that many firms could face in the near future and proposed design insights for solutions. By keeping in mind these tensions and design insights in advance, we recommend industry practitioners start designing their own AI, because every company has unique characteristics (e.g., different tasks, organizational culture). Moreover, we advise firms to build a new organizational system that allows members to participate in designing the evaluation index and the AI itself. Since it is impossible to achieve perfection in AI or define and optimize fairness, the key is to find an optimum through discussion, trial and error, and considering the sociotechnical perspective. Based on a deep understanding of the users and social context, we call our HCI community to investigate diverse devices/tools that could expand the environment of social transparency with the aim of mitigating tensions and promoting a safety net.

5.3 Proactive Agreement through Empowerment

Our findings show that many employees came to view AI's adoption in HRM more positively [i.e., were no longer wary of AI's opaqueness] as other stakeholders persuasively shared about the dangers of AI transparency and imperfect interpretability. Especially, employees liked the idea of a proactive contractual agreement by the firm as a part of social transparency, such as the following: "instead of directly using the AI's decision, the HR team will add their own scores," "the firm will adopt an external audit system," "the HR team is obligated to respond to the employee's claims," "the firm will publicize the ratio of human to AI scores," etc. Not only the AI design but also the proactive agreements to inform and empower workers could increase trust and fairness toward AI evaluation systems.

The HCI community has long investigated sociotechnical gaps, in which technology cannot guarantee the complete certainty of privacy or security, but which humans naturally navigate with nuance and flexibility. The adoption of AI to the HRM context has deepened such gaps. Even if we endeavor to define and achieve fairness through discussion, both AI and its application to HRM are inherently new, imperfect, and unpredictable. It is inevitable that many unexpected issues will spring up. Thus, based on our findings, we suggest both researchers and practitioners move beyond algorithm transparency or interpretability itself to focus on *proactive agreements through informed consent*. The need for giving more global control to users [3, 35, 97] or informed consent [27] has been pointed out in the context of direct one-on-one interactions with the technology (e.g., social media [97], recommender system [27]). Similarly, when AI is adopted for HRM, we propose that firms proactively explain the reasons why algorithms and interpretability cannot be completely publicized while endeavoring to create solutions through informed consent [32]. Doing so can give more global control to employees in advance.

5.4 Limitations and Future Research

Both academia and business practices can benefit from what we discovered regarding multiple stakeholders' needs, the emerging tensions between them, and the design insights that could alleviate the tensions. However, we acknowledge the limitations of our scenario-based codesign. Some participants may have been more flexible or held their opinions loosely because the issues are not what they are currently experiencing. If AI adoption in HRM was not a futuristic scenario but was actually happening in their firm, stakeholders may perceive the situation more seriously and pursue their interests more stubbornly. In future research, we hope to conduct case studies or codesign/redesign research for a specific firm that is planning to or has adopted AI for HRM. Indeed, during our study, many employers and HR managers wanted to see diverse case studies of firms that have already utilized AI in HRM. Since such real-world success/failure cases can give future direction for other firms, we invite researchers and industry practitioners to this area of research. Also, although our study involved multiple stakeholders in the design process, our design insights should be implemented and validated through statistical tests with a larger sample.

We found that adopting AI in HRM is not just a matter of completely replacing human HR teams to produce fairer and more objective results—AI has joined the current sociotechnical context by presenting a new important question: Can AI assist humans with difficult and routine tasks that they avoid to make space for them to do what they most need and want? [67]. This implies that the adoption of AI in HRM will greatly change the existing social structures of workplaces, such as the stakeholders' roles, the power structure, and the interactions among them. This phenomenon is entirely new for both scholars and stakeholders in workplaces. Because it is inevitable and will have a great impact, there will be more interesting research questions to pursue. For example, our results show that the adoption of AI calls for newly emerging human roles, which have not been investigated yet, such as *external AI auditors* (similar to accounting auditors), *AI translators* (who help interpret AI's complex explanations), *labor attorneys* (who respond to unfair decisions made by AI), and *in-house auditors* (who validate the robustness/biases of the algorithm and its interpretation). Also, our employees insisted that humans and AI hold each other in check through *mutual surveillance* while maximizing each other's strengths (i.e., collaborate). These new roles and interactions caused by AI provide fruitful research agendas for future work on AI in HRM.

6 CONCLUSION

In this study, we investigated the needs various stakeholders have, the tensions that occur between them, and the design insights that could lead to solutions. If one stakeholder group tries to design AI in a way that benefits their group only, it could cause severe social problems: employees could be surveilled or could suffer from extreme competition, privacy invasions, and loss of humanity. Such social problems already happen at this moment, but AI adoption in HRM may accelerate such issues. Thus, we hope enterprises pay special attention to adopting and designing AI in HRM by providing space within the company for open discussions and codesign sessions among various stakeholders. We can achieve harmony between all of the stakeholders, including AI, through organic collaboration.

ACKNOWLEDGMENTS

We are deeply grateful to the participants for joining our study and the anonymous reviewers for the constructive feedback. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5B8096358) and was financially supported in part by the Institute of Communication Research, Seoul National University. Also, we would especially like to acknowledge Daehwan Ahn's contributions to this study as a corresponding author.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 582:1-582:18. https://doi.org/10.1145/3173574.3174156

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

[3] Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, and Paul N. Bennett. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Human Resources Professionals Association (HRPA). A new age of opportunities: what does artificial intelligence mean for HR professionals?. Retrieved April 20, 2020 from https://www.hrpa.ca/Documents/Public/Thought-Leadership/HRPA-Report-Artificial-Intelligence-20171031.PDF

[5] Boris Babic, Sara Gerke, Theodoros Evgeniou, and I. Glenn Cohen. 2021. Beware explanations from AI in health care. *Science* 373, 6552: 284–286.

[6] Yuki Ban, Sho Sakurai, Takuji Narumi, Tomohiro Tanikawa, and Michitaka Hirose. 2015. Improving work productivity by controlling the time rate displayed by the virtual clock. In *Proceedings of the 6th Augmented Human International Conference*, 25–32. https://doi.org/10.1145/2735711.2735791

[7] Virginia Braun and Victoria Clarke. 2013. *Successful qualitative research: A practical guide for beginners*. sage.

[8] Richard Buchanan. 1992. Wicked problems in design thinking. *Design issues* 8, 2: 5–21.

[9] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, 339–348.

[10] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17.

[11] C. West Churchman. 1967. Guest editorial: Wicked problems.

[12] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. 2020. Multi-Armed Bandits with Fairness Constraints for Distributing Resources to Human Teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (HRI '20), 299–308. https://doi.org/10.1145/3319502.3374806

[13] Bo Cowgill. 2018. Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University* 29.

[14] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies* 2015, 1: 92–112.

[15] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

[16] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300724

[17] Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe. 2020. Fair AI: Challenges and Opportunities. *Business & Information Systems Engineering* 62, 4: 379–384. https://doi.org/10.1007/s12599-020-00650-3

[18] Alessandro Di Fiore and Marcio Souza. 2021. Are Peer Reviews the Future of Performance Evaluations? *Harvard Business Review*. Retrieved July 10, 2021 from https://hbr.org/2021/01/are-peer-reviews-the-future-of-performance-evaluations

[19] Jodi Forlizzi. 2018. Moving beyond user-centered design. *Interactions* 25, 5: 22–23. https://doi.org/10.1145/3239558

[20] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*. Springer, 55–95.

[21] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, 2.

[22] Nigel Guenole. The Business Case for AI in HR. 36.

[23] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2019. Designing for complementarity: Teacher and student needs for orchestration support in ai-enhanced classrooms. In *International Conference on Artificial Intelligence in Education*, 157–171.

[24] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16.

[25] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. Association for Computing Machinery, New York, NY, USA, 1–12. Retrieved November 23, 2021 from https://doi.org/10.1145/3419249.3420149

[26] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2: 119–131.

[27] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S. Ackerman, and Eric Gilbert. 2021. Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18.

[28] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

[29] Qiong Jia, Yue Guo, Rong Li, Yurong Li, and Yuwei Chen. 2018. A conceptual artificial intelligence application framework in human resource management. In *Proceedings of the International Conference on Electronic Business*, 106–114.
[30] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
[31] Katherine C. Kellogg, Melissa A. Valentine, and Angele Christin. 2020. Algorithms at work: The new contested terrain of control. *Academy of Management Annals* 14, 1: 366–410.
[32] Tae Wan Kim and Bryan R. Routledge. 2020. Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly*: 1–28.
[33] Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. 2019. Understanding Personal Productivity: How Knowledge Workers Define, Evaluate, and Reflect on Their Productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300845
[34] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 272–283.
[35] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–14. https://doi.org/10.1145/3290605.3300641
[36] Robert E. Kraut, Susan T. Dumais, and Susan Koch. 1989. Computerization, productivity, and quality of work-life. *Communications of the ACM* 32, 2: 220–238. https://doi.org/10.1145/63342.63347
[37] Matt J. Kusner and Joshua R. Loftus. 2020. The long road to fairer algorithms. *Nature* 578, 7793: 34–36. https://doi.org/10.1038/d41586-020-00274-3
[38] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30.
[39] Lauren Kaori Gurley. 2021. Amazon Delivery Drivers Forced to Sign 'Biometric Consent' Form or Lose Job. Retrieved January 7, 2022 from https://www.vice.com/en/article/dy8n3j/amazon-delivery-drivers-forced-to-sign-biometric-consent-form-or-lose-job
[40] Colin Lecher. 2019. How Amazon automatically tracks and fires warehouse workers for 'productivity.' *The Verge*. Retrieved February 5, 2020 from https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations
[41] Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445261
[42] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1: 205395171875668. https://doi.org/10.1177/2053951718756684
[43] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1035–1048. https://doi.org/10.1145/2998181.2998230
[44] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW: 1–26.
[45] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, and Alexandros Psomas. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW: 1–35.
[46] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1603–1612.
[47] Weiwen Leung, Zheng Zhang, Daviti Jibuti, Jinhao Zhao, Maximilian Klein, Casey Pierce, Lionel Robert, and Haiyi Zhu. 2020. Race, Gender and Beauty: The Effect of Information Provision on Online Hiring Biases. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11.
[48] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.
[49] Jennifer Liu. 2019. A.I. is changing how much workers trust their managers—and that could be a good thing. *CNBC*. Retrieved February 5, 2020 from https://www.cnbc.com/2019/10/15/ai-is-changing-how-much-workers-trust-their-managerswhy-thats-good.html
[50] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–16. https://doi.org/10.1145/3313831.3376727
[51] Louis Stone. 2021. Amazon tells drivers to sign biometric consent form, or lose their job. *AI Business*. Retrieved January 7, 2022 from https://www.aibusiness.com/document.asp?doc_id=768343
[52] Caitlin Lustig, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig. 2016. Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1057–1062.
[53] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
[54] Nikolas Martelaro, Jaime Teevan, and Shamsi T. Iqbal. 2019. An Exploration of Speech-Based Productivity Support in the Car. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300494
[55] Patty McCord. 2014. How netflix reinvented HR. *Harvard Business Review* 92, 1: 71–76.
[56] Mike Teodorescu. Protected Attributes and "Fairness through Unawareness." Retrieved March 21, 2022 from https://ocw.mit.edu/resources/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/module-three-framework/protected-attributes/
[57] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38.
[58] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. 2019. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 1–9.
[59] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11: 501–507.
[60] Christoph Molnar. 2020. *Interpretable Machine Learning*. Lulu. com.
[61] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2021. Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.
[62] Sunyoung Park, Shinhee Jeong, and Boreum Ju. 2021. MOOCs in the workplace: an intervention for strategic human resource development. *Human Resource Development International* 24, 3: 329–340.
[63] Valery Petrushin. 1999. Emotion in speech: Recognition and application to call centers. In *Proceedings of artificial neural networks in engineering*, 22.
[64] Jeffrey Pfeffer. 1994. Competitive advantage through people. *Boston/Mass.*
[65] Jeffrey Pfeffer. 1998. *The human equation: Building profits by putting people first*. Harvard Business Press.
[66] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.
[67] Sebastian Raisch and Sebastian Krakowski. 2021. Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review* 46, 1: 192–210.
[68] Horst WJ Rittel and Melvin M. Webber. 1973. Dilemmas in a general theory of planning. *Policy sciences* 4, 2: 155–169.
[69] Lionel P. Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad. 2020. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction*: 1–31.
[70] Nancy Roberts. 2000. Wicked problems and network approaches to resolution. *International public management review* 1, 1: 1–19.
[71] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.
[72] Samantha Robertson and Niloufar Salehi. 2020. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. *arXiv preprint arXiv:2007.06718*.
[73] Eric Rosenbaum. 2019. IBM artificial intelligence can predict with 95% accuracy which workers are about to quit their jobs. *CNBC*. Retrieved September 15, 2020 from https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html
[74] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms used within the US Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.
[75] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2: 348:1-348:41. https://doi.org/10.1145/3476089

[76] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar R.S. Gerlicher. 2021. ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3411764.3446647

[77] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), 59–68. https://doi.org/10.1145/3287560.3287598

[78] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6: 495–504.

[79] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376783

[80] Madeline Stone. Amazon employees reportedly slam each other through this internal review tool. *Business Insider*. Retrieved July 10, 2021 from https://www.businessinsider.com/amazon-employees-reportedly-slam-each-other-through-this-internal-review-tool-2015-8

[81] Nicole Sultanum, Michael Brudno, Daniel Wigdor, and Fanny Chevalier. 2018. More text please! understanding and supporting the use of visualization for clinical text overview. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

[82] S. Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More Than Humans with Our Personal Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 538:1-538:9. https://doi.org/10.1145/3290605.3300768

[83] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *California Management Review* 61, 4: 15–42. https://doi.org/10.1177/0008125619867910

[84] Mike HM Teodorescu, Lily Morse, Yazeed Awwad, and Gerald C. Kane. 2021. FAILURES OF FAIRNESS IN AUTOMATION REQUIRE A DEEPER UNDERSTANDING OF HUMAN-ML AUGMENTATION. *MIS Quarterly* 45, 3.

[85] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14. https://doi.org/10.1145/3173574.3174014

[86] James Vincent. 2021. Amazon delivery drivers have to consent to AI surveillance in their vans or lose their jobs. *The Verge*. Retrieved September 3, 2021 from https://www.theverge.com/2021/3/24/22347945/amazon-delivery-drivers-ai-surveillance-cameras-vans-consent-form

[87] Demetris Vrontis, Michael Christofi, Vijay Pereira, Shlomo Tarba, Anna Makrides, and Eleni Trichina. 2021. Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. *The International Journal of Human Resource Management*: 1–30.

[88] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–15. https://doi.org/10.1145/3290605.3300831

[89] Adrian Weller. 2017. Challenges for transparency. *arXiv preprint arXiv:1708.01870*.

[90] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 666–677. https://doi.org/10.1145/3442188.3445928

[91] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 1–14. https://doi.org/10.1145/3173574.3174230

[92] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. 13.

[93] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11.

[94] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 279:1-279:12. https://doi.org/10.1145/3290605.3300509

[95] Bowen Yu, Yuqing Ren, Loren Terveen, and Haiyi Zhu. 2017. Predicting Member Productivity and Withdrawal from Pre-Joining Attachments in Online Production Groups. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1775–1784. https://doi.org/10.1145/2998181.2998227

[96] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1245–1257.

[97] Bo Zhang and S. Shyam Sundar. 2019. Proactive vs. reactive personalization: Can customization of privacy enhance user experience? *International Journal of Human-Computer Studies* 128: 86–99. https://doi.org/10.1016/j.ijhcs.2019.03.002

[98] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 1–23.

[99] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07), 493–502. https://doi.org/10.1145/1240624.1240704

[100] 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved April 9, 2020 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[101] 2020. Chinese construction firms using AI to keep tabs on workers. *South China Morning Post*. Retrieved August 15, 2020 from https://www.scmp.com/news/china/science/article/3091738/chinese-construction-firms-using-ai-monitor-workers-safety-also

[102] 2021. Growing pains of Korea's leading internet firms. *koreatimes*. Retrieved April 16, 2021 from http://www.koreatimes.co.kr/www/tech/2021/04/133_304787.html

[103] Inside Amazon: Wrestling Big Ideas in a Bruising Workplace - The New York Times. Retrieved July 10, 2021 from https://www.nytimes.com/2015/08/16/technology/inside-amazon-wrestling-big-ideas-in-a-bruising-workplace.html

[104] From Fear to Enthusiasm. *Oracle & Future Workplace*: 18.

[105] Union Accuses Amazon of Breaking Federal Law by Firing Activist - Bloomberg. Retrieved February 5, 2020 from https://www.bloomberg.com/news/articles/2019-03-20/union-accuses-amazon-of-breaking-federal-law-by-firing-activist

[106] Amazon's sexist AI recruiting tool: how did it go so wrong? Retrieved April 9, 2020 from https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

[107] Your Raise Is Now Based on Next Year's Performance - Bloomberg. Retrieved September 17, 2020 from https://www.bloomberg.com/news/articles/2018-07-09/your-raise-is-now-based-on-next-year-s-performance

[108] Federal Laws Prohibiting Job Discrimination Questions And Answers | U.S. Equal Employment Opportunity Commission. Retrieved March 21, 2022 from https://www.eeoc.gov/fact-sheet/federal-laws-prohibiting-job-discrimination-questions-and-answers