Hiring Algorithms: An Ethnography of Fairness in Practice

Elmira van den Broek
VU University Amsterdam, e.p.h.vanden.broek@vu.nl

Anastasia Sergeeva
VU University Amsterdam, a.sergeeva@vu.nl

Marleen Huysman
VU University Amsterdam, m.h.huysman@vu.nl

Hiring Algorithms: An Ethnography of Fairness in Practice

Short Paper

Elmira van den Broek
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
e.p.h.vanden.broek@vu.nl

Anastasia Sergeeva
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
a.sergeeva@vu.nl

Marleen Huysman
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
m.h.huysman@vu.nl

Abstract

While increasing attention in society is given to the role of AI in affording and threatening ethical values, such as fairness, little is known about how ethical values and AI are played out in organizations. Building on an ethnographic in-depth study of a large multinational company that recently implemented AI to enable a fair recruitment process, we show that AI brings to the fore the role of fairness in decision-making in several ways. We reveal that the development and use of AI does not necessarily improve nor degrade ethical values, but instead shapes what comes to be understood as ethical in the first place. We extend the conversations on AI by showing that it may not be enough to focus on changes in work practices, occupational boundaries, and power relations, but that research should take into account the role of AI in shaping what we consider ethical.

Keywords: Artificial Intelligence, ethical values, fairness, ethnography, hiring algorithms

Introduction

Artificial Intelligence (AI) is currently an object of conversations of two highly contrasting debates: one group says that AI will help to create a more ethical world by bringing objectivity and fairness, and the other accuses AI of committing ethical violations such as racism, sexism, and harming privacy. There is, however, still little insight into what actually happens to ethical values when AI is rolled out in organizations. In this paper we reveal that the development and use of AI does not necessarily improve nor degrade ethical values but instead shapes what comes to be understood as ethical in the first place. This paper draws on a study on the implementation of an AI application at a large multinational company that brings to the fore one specific ethical value: fairness in hiring employees.

Emerging studies on fairness and AI have highlighted the importance of considering the societal context that surrounds AI applications for achieving fair decision-making (Hoffmann 2019; Selbst et al. 2019; Madden et al. 2017). Research has identified several traps AI applications can fall into, such as a failure to model the entire system over which fairness will be enforced ("the framing trap"), or a failure to account for the full meaning of fairness ("the formalism trap") (Selbst et al. 2019). To account for the different meanings

of the concept of fairness within AI, we borrow ideas from research on organizational justice. Organizational justice deals with the role of fairness in the workplace (Greenberg 1990), and has suggested that in order for decisions to be perceived fairly they should be made consistently, represent the interests of affected individuals, suppress personal bias, use as much accurate information as possible, be correctable, and be compatible with ethical values (Leventhal 1980). We aim to extend the conversation on AI by showing that it may not be enough to focus on changes in work practices, occupational boundaries, and power relations (Barbour et al. 2018; Faraj et al. 2018; Newell and Marabelli 2015; Orlikowski and Scott 2014). Instead, research should take into account the role of AI in shaping what we consider ethical.

We are currently at the stage of iteratively developing theory based on our preliminary findings of this phenomenon. We find that AI in recruitment brings to the fore the role of fairness in organizational decision-making in several ways. First, when AI is not yet used, notions of fairness are not yet questioned and its meaning is shared between stakeholder groups. As groups actually implement and inscribe fairness into the tool, however, AI exposes different means to achieve fairness which were previously latent. Later, as groups start to work with AI across a range of situations, AI triggers clashes between diverging perceptions about how fairness should be achieved. We illustrate these findings by our process analysis of the specific practices and interactions of multiple stakeholders in the workplace involved in working with AI.

Method

We conducted an ethnographic in-depth study at the HR department of a large multinational company in Europe, "MultiCo" (pseudonym), that recently implemented AI to enable a fair recruitment process. At the time of data collection, MultiCo belonged to one of the world's largest Fast-Moving Consumer Good (FMCG) companies, with annual revenues exceeding $50 billion. The company had almost 200,000 employees in more than 50 countries worldwide. In September 2018, MultiCo launched an AI application for the recruitment process of all its graduate trainee programs in Europe. Candidates could apply for four different trainee programs in ten different European locations. The trainee programs were highly competitive: yearly more than 10,000 candidates applied for 100 open positions. The AI application was delivered by an external vendor, "NeuroYou" (pseudonym), that promised the organization to remove subjectivity and bias from workforce decisions, by drawing on data science, neuroscience, and machine learning. The AI application replaced the standardized online tests MultiCo used before (e.g. logical reasoning test) by neuroscience games, and added the possibility of automated video analysis. Three stakeholder groups were involved in working with the AI application at MultiCo: the HR team, which consisted of multiple recruiters and two HR managers, the AI team, which consisted of several data scientists and a People Analytics (PA) manager, and managers from the different business units of the company who served as assessors during selection events.

We have conducted 7 months (726 hours) of non-participant observation of the work around the AI application in graduate recruitment - including 110 meetings and 27 selection events - in the period between October 2018 and April 2019. During our observations, we focused on the design and development of the AI application, including the development and analysis activities of the AI team around the hiring algorithm. Moreover, we focused on the activities of the HR team and managers around the selection of candidates, such as their use of criteria and AI output during selection events. In addition, we also conducted 32 formal interviews and 18 informal interviews with candidates, the AI team, the HR team, and employees of NeuroYou. The interviews with the decision-makers around the AI project served to gain an understanding of the objectives, context, developments and implications of AI for work. In the interviews with candidates, we asked them to walk us through the specific steps of their selection process, and share their experiences with the AI and human assessment.

We followed a process research approach (Langley 1999) to track the flow of events and understand the unfolding work with AI in practice. In analyzing our data, we started with creating a list of the events at the organization to represent and organize the evidence across phases and identify patterns. We then organized the key events, separating those representing the perspectives of the HR team, AI team, managers, and candidates by systematically going through the field evidence for each group. We zoomed in on the various notions of fairness across the groups of people over time, and finally constructed our narrative of how fairness evolved following the introduction of AI as shown in Table 1. As aspects of procedural justice turned out to have an accentuated role, we focused on this dimension of fairness.

Findings

In the following sections we discuss how notions of fairness evolved over time, and the key actions related to AI and fairness of the HR team, the AI team, managers, and candidates. This is summarized in Table 1.

Table 1. Evolving notions of fairness, and the key actions of stakeholders around AI and fairness.

	Actions around AI	Actions around fairness	Fairness themes
Phase 1: pre-implementation	• HR team develops a digital agenda	• HR team guards fairness by suppressing bias in the process • HR team considers AI as means to enhance fairness	• Guarding fairness
Phase 2: implementation	• HR team collects training data from top performers • AI team performs comparative analysis • AI team builds hiring algorithm • HR uses fixed rejection threshold for algorithmic recommendations	• Pilot confirms issue of bias • AI application promises accuracy with hiring algorithm, and bias suppression and consistency with algorithmic recommendations • HR team adds correctability by keeping a "human in the loop" • HR team communicates and educates groups about how AI enhances fairness	• Confirming fairness concerns with AI • Inscribing fairness into AI • Enrolling actors into AI
Phase 3: post-implementation	• HR team deviates from fixed threshold • Managers overrule AI assessment • Candidates complain and game the system • AI team blames users for incorrect use	• HR team contests consistency and representativeness • Candidates contest accuracy and opportunity to perform • Managers contest consistency accuracy • AI team contests consistency and correctability	• Contesting notions of fairness

Phase 1. Pre-Implementation of AI

Guarding Fairness

Before the implementation of the AI application, the HR team was the main actor responsible for guarding fairness in the recruitment process. They considered *personal bias*, or the influence of one's personal interest on the decisions (Leventhal 1980), as a main threat to fairness, and perceived it as their task to limit the potential impact of this ethical violation in the workplace. The reasoning of the HR team was that an unbiased recruitment process would allow for a wider pool of talent to hire from, and aligned with the company's global mission of diversity and inclusion in the workplace. As explained by an HR manager: "So why are we looking for objectivity and fairness in the recruitment process? It is part of a wider strategy to make sure that we have diversity within our company. And not just diversity like gender and nationality, but actually diversity of thought" (HR manager 1). A recruitment process without personal biases, the HR manager reasoned, would result in a more diverse group of employees in terms of skill set and traits.

The HR team engaged in multiple activities to suppress bias in the recruitment process. During selection events in which managers would assess candidates, HR professionals removed resumes to facilitate "blind"

assessment, gave "unconscious-bias trainings", and coached managers how to assess fairly. For instance, an HR professional would emphasize before the start of a selection event: "Candidates need to feel they had as much time as the other candidates. That they have been assessed fairly" (Field notes group panel). HR professionals also corrected managers when they engaged in biased practices, such as using "lack of culture fit" as an excuse to reject candidates for which they could not explain their reasoning. As illustrated by a comment of an HR manager: "We often use cultural fit as argument [to reject a candidate]. But why? We are going to push you a bit to explain why you think someone does not fit with the company culture" (Field notes in-house day). Managers were aware of the guarding role of HR and sometimes joked about the HR team acting as "the bias-police". However, managers seemed to generally agree with the importance of suppressing bias, by learning about sources of bias, correcting each other on subjective judgment, and admitting own potential biases. For example, a manager openly admitted his bias about a stereotypical polite, old-fashioned British candidate, and asked about the opinion of the other managers: "I was 100% biased. When the guy walked in I shut him out. I am very open here to you guys. Just the whole appearance didn't work. The way he walked, the way he spoke. But I come biased! So, feel free to challenge" (Field notes group panel). Managers reasoned that hiring candidates on the basis of ability instead of job-irrelevant characteristics such as appearance would eventually benefit the performance of their teams.

In Spring 2017, the HR team started to consider new technological solutions as a means to enhance fairness in the recruitment process, driven by the company's larger strategy of digitizing HR. The HR team was convinced that the use of AI would offer a solution to perceived threats to fairness, by enabling them to "objectively measure soft skills and human traits, expose diversity, and compare candidates without bias" (Company documents). In sum, when AI was not yet used, the HR team was the main guardian of fairness in the organization. The notion of fairness was considered as being settled and shared by other actors, and mostly focused on suppressing bias in the recruitment process. However, once the HR team considered AI as a means to enhance fairness, other notions of fairness which were previously latent, came to the fore.

Phase 2. Implementation of AI

Confirming Fairness Concerns With AI

In the second phase, the HR team decided that AI was an appropriate solution for further suppressing bias in their recruitment process. In Summer 2017, the HR director spearheaded the agenda for the use of AI in recruitment by bringing on board a People Analytics (PA) manager, a newly established role for the purpose of the AI project. Eventually, they decided to enter the pilot stage with the vendor "NeuroYou". NeuroYou offered an AI application assessing over 150 skills and traits, emphasizing their superior ability to "eliminate bias and guarantee objective and role-specific evaluations" and "ensure consistency and accuracy at every stage" (Company documents). In Autumn 2017, the HR team piloted the AI application. The aim of the pilot was to assess the hiring practices of the current selection process, and to collect the data necessary to build the algorithm for the new selection process. HR managers were responsible for collecting the data, by asking more than 350 current graduate trainee candidates and 40 top-performing employees to complete the neuroscience gamified assessments and asynchronous video interview via a provided link. Top-performers were identified based on their performance appraisal scores. The neuroscience games aimed to measure cognitive (e.g. task-switching), social (e.g. assertiveness) and emotional (e.g. expression recognition) skills and traits of participants in an easy-to-use interactive environment. In the video interviews candidates were asked to record short answers to questions about themselves, such as: "What is your most significant and challenging achievement to date?". Video interviews helped to extract data on verbal and nonverbal communication, intonation, and facial expressions.

The results of the pilot were interpreted as confirming concerns that the selection process was not entirely fair. Specifically, it was revealed that assessors tended to choose more extravert candidates, while the top-performers were found to be less extroverted than an average employee. The HR managers concluded: "This suggests bias and shows we are not always assessing candidates on what actually makes them successful" (Field notes selection event). Therefore, the AI pilot confirmed the concerns of the HR team about the presence of bias in the selection process, and the need for a technical solution to constrain this bias.

Inscribing Fairness Into AI

When running the pilot, the HR team mainly referred to fairness as suppressing bias. However, the development of the AI application exposed other means to achieve fairness which were previously latent.

This resulted in additional interpretations of fairness, including the importance of accuracy of information related to performance on the job, and making decisions consistently across candidates and time (Leventhal 1980). Specifically, NeuroYou promised that their AI application would make the selection process highly accurate, by using hiring algorithms to predict future performance. The AI team was responsible for building the hiring algorithm based on the training data taken from the assessments completed by participants of the pilot. The hiring algorithm extracted the traits of top-performing employees from the training data, and matched these traits against candidates to identify those candidates with the highest overlap in traits. The algorithm would then calculate an algorithmic recommendation per candidate. This was represented as a "match percentage score" with the top-performer profile of MultiCo. For example, a score of 95 would imply that the candidate's traits matched the top-performer's traits for 95 percent, indicating a highly predicted successful employee at MultiCo. Thus, the hiring algorithm appeared to be a valid tool for evaluating candidates' performance on the job. Moreover, the algorithmic recommendations were considered to suppress personal bias by constraining the influence of human interests on the evaluation of the candidate. Finally, the algorithmic recommendations promised to offer consistency, by allowing for an equivalent comparison of candidates across different programs, locations, and time. This would facilitate easy and consistent comparison of candidates for decision-makers.

In Summer 2018, the AI application went live for the recruitment of all trainee programs in Europe. Candidates were now completing a four-step selection process, consisting of neuroscience games, a video interview, an in-house day, and a final group panel. Although the AI application offered new means to ensure fairness, the HR team explicitly decided on involving human decision-makers in every step of the process, as they considered retaining a "human in the loop" as another important factor for a fair process. The decision about human oversight signaled yet another important fairness condition, previously not considered before, i.e. correctability of decisions (Leventhal 1980). Human oversight was ensured in several ways. In the first step, in which candidates had to complete the neuroscience games, the HR team manually selected candidates based on the algorithmic recommendation, instead of automatic rejection of candidates. In the second step, in which candidates recorded their answers to video questions, the HR team did not employ the algorithmic recommendation because they considered it not yet reliable. They chose instead to have HR professionals assessing candidates' communication skills using a standardized scoring form. In the third step, in which candidates visited the in-house assessment center of the company and went through a series of role-play assessments and structured interview, junior managers assessed candidates' social skills using a standardized scoring form. Finally, candidates who reached the final step were invited to a group panel where senior managers assessed their performance and made the final hiring decision.

Thus, in developing the AI application, the groups were confronted with new means to achieve fairness. This forced both teams to consider new notions of fairness, including accuracy by the use of a hiring algorithm, consistency by acting upon algorithmic recommendations, and correctability by retaining human oversight in every step of the selection process.

Enrolling Actors Into AI

When the AI application went live, the HR team started to actively enroll stakeholders into their project, by communicating and educating them about the use of AI for enhancing fairness. First, the HR team focused their attention on managers, in which they communicated the benefits of the AI application for the accuracy of assessment and the diversity aims of the company. This was illustrated by an HR manager during a selection event:

> "The beauty of this is that for the first time it can help to objectively assess [soft skills and traits]. It is more or less a proxy for diversity of thought. [AI can help us to answer the question of] how can we start building teams that are more diverse?" (Field notes group panel)

The HR team often referred to the pilot results to support their claims. For example, during every group panel, HR managers highlighted the exposed differences in extraversion between top-performers and candidates as an example illustrating bias that needs to be suppressed. These insights were enthusiastically received by senior managers, who wanted to test their own traits with neuroscience games, and were impressed by the ability of AI to expose bias in the hiring process. As commented by a senior manager: "This brings interesting insights, to challenge our own biases on it [the assessment of candidates]" (Field notes group panel). Moreover, as managers were required to use the algorithmic output in their assessment

to suppress personal bias, the HR team started to educate managers on how to interpret and use the output. For example, senior managers were shown "word clouds" representing dominant candidate traits. The HR manager would help to formulate specific questions about traits that senior managers could ask, such as: "Have you ever experienced a situation where your emotions where in the way?" for the trait "emotionality".

Moreover, the HR team debriefed candidates about the use of AI in the selection process and its benefits for accuracy and bias suppression during selection events. Candidates attending recruitment events were encouraged to take the neuroscience assessments as natural as possible, "because it is impossible to fake" (Field notes informal talk in-house day). The HR team also had internal discussions on how to use the tool as consistently as possible, since the HR team would work with the tool on a daily basis. The HR team decided to achieve consistency by using a fixed threshold for rejecting candidates, which was calculated by the AI team based on the pilot results. During weekly team meetings, the HR team reviewed the algorithmic recommendations of candidates, and rejected candidates based on the fixed threshold. As explained by an HR manager to a new colleague: "We need a very structured process, because we are dealing with so many candidates. And everyone is assessed in the same way. We need to be objective" (Field notes weekly HR team meeting). In sum, the HR team aimed to enroll stakeholder groups by emphasizing notions of fairness that linked most closely to their interaction with the AI application, including accuracy and bias suppression for managers and candidates, and consistency for the HR team.

Phase 3. Post-Implementation of AI

As different stakeholder groups started to work with AI across a range of different situations, they experienced clashes between those notions of fairness that were inscribed in the tool and those implicit understandings of fairness that were important for their daily work. We discuss how the different groups contested different notions of fairness below.

HR Team Contesting Notions of Fairness

The first mismatch between notions of fairness was concerned with having consistency and representing the interests of different parties, yet a newly surfaced notion of fairness referred to as "representativeness" (Leventhal 1980). For example, the HR team assumed that using a fixed threshold to reject candidates would enable a consistent process. However, during their work with the AI application on a daily basis, HR professionals experienced that the fixed threshold did not allow for differentiation between the situated contexts of the programs, locations, and temporary changes in supply and demand. This turned out to be problematic for hiring managers who needed to meet hiring targets for programs that were less popular and thus received fewer applications. HR professionals therefore increasingly had to make exceptions and changes to the fixed threshold. As explained by an HR professional:

> "The threshold depends on the number of candidates that we have. So, in a simple way, how picky we can be. We know for the management program in Germany that we can really hand-pick every single candidate. We have a great brand name there. So, there we decided that the threshold is at least, let's say, 85%. For the video interviews, at least 86%. And so on. But for example, for the supply program in the UK we know that we cannot be so picky."
> (Field notes weekly HR team meeting)

HR professionals deviated from the threshold based on the supply and demand of candidates, specific candidate requirements (e.g. language, educational background), or personal arguments for selecting candidates. For example, when a candidate had a score slightly below the threshold, HR professionals would often say: "Can we give him a chance?" or "If it's one percent difference, you can just move her forward". These deviating practices in turn raised concerns of the HR manager, who feared the changes undermined consistency in the selection process. The HR manager aimed to favor consistency over representativeness or at least find a solution to achieve both. She approached the AI team for advice on how to set new thresholds that would take the situated context in mind. In her explanation to the AI team she repeatedly emphasized how paramount consistency of process was in her idea of a fair process, implying that representativeness was undermining its goals:

> "So, what we started implementing as an approach this year was differentiating thresholds, and I really want to look into the details of thresholds to have the right ones for the entire [recruitment]

> cycle, with no changes. So, that's a priority for me, to have a super fair process during the [recruitment] cycle." (Field notes HR and AI team meeting)

Candidates Contesting Notions of Fairness

The second mismatch between notions of fairness had to do with accuracy that did not coincide with the perceived validity of the assessments and opportunity to demonstrate one's skills, through the eyes of candidates. The HR team was confronted with several candidates who expressed during recruitment events, selection events, or via email that they were not given a fair chance to prove themselves. For example, candidates complained that they did not recognize themselves in the AI output they received upon completing the games. During one specific instance, the HR manager received an email of an upset candidate, stressing that his results were inaccurate and demanding that his data was deleted. The HR manager complied with this request. Other candidates complained they felt unfairly treated because they lacked the opportunity to perform (Gilliland 1993), as they had the impression they were just "playing a game". This resulted in HR managers worrying about how candidates experienced fairness of the selection process.

In contrast, the HR team was also confronted with candidates who aimed to gain an unfair advantage over other candidates in the selection process by "gaming the system". For example, HR professionals found out that several candidates bypassed the system by creating a new account with a different email address, in the hope to improve their AI scores. An HR professional addressed this issue during a weekly team meeting: "So, he was applying for the leadership program, I think, and he had a really low match score. So, we rejected him. And he was so cheeky that he started a second account with another email!" (Field notes weekly HR team meeting). Although the candidate showed high scores the second time he took the test, the HR team decided to reject the candidate as he had an unfair advantage over other candidates. The possibility to "game the system" was confirmed by candidates during the interviews, in which they explained that it was possible to cheat on several neuroscience games as well. For example, a candidate expressed about a specific game in which you had to memorize changing figures: "You could actually cheat on those games. If you would do the game with two people, hold your phone in your hand, and both make a picture [of the figure you have to memorize], I am sure you would pass the game" (Candidate 1). Thus, the HR team was confronted with candidates who found it difficult to demonstrate their skills and perceived the AI output as inaccurate, and started to use strategies to game the system.

Managers Contesting Notions of Fairness

The third mismatch was between consistency and accuracy, in which managers aimed to overrule algorithmic recommendations which they perceived as inaccurate, while the HR team aimed to guard consistency. In several instances, the HR team was confronted with frustrated managers who could not hire their preferred candidate because the candidate was rejected by the AI application. For example, a sales manager could not hire his current intern, because the intern "failed" the AI assessments. The intern commented on the incident: "I had a score of 30%, so basically I cannot concentrate. Which is kind of true. But I am lucky that I can still do my internship, because for this internship the test was not required". The sales manager was furious, and felt the algorithm had judged the intern on inaccurate information, as the intern had already proved himself successful during internship. This resulted in a conflict between the HR team and the sales manager, in which the sales manager argued that human assessment should be preferred over AI assessment: "Our human assessment should be leading. If the person is super great and nails that but doesn't pass the AI assessment, how do you explain that?" (Field notes work floor). The HR team responded to critiques of AI assessment by defending the importance of consistency, however, this was often unsatisfactory to the managers.

In addition, the HR team was confronted with managers who criticized the whole underlying notion of consistency brought by the AI application by claiming that it would result in less diversity. For example, one of the senior managers objected: "We will have less diversity because we will hire more of the same profile, right?" (Field notes group panel). Another manager expressed the fear of "cloning people", by selecting candidates on the same set of traits. As a consequence of fears of what consistency may result in, several managers started to direct their assessment on traits and skills not measured by AI. This was illustrated by the following discussion between two managers:

> "Manager 1: We have to watch out that we don't get only leaders.
> Manager 2: Yes, but with AI we are only selecting similar profiles. We basically don't have any diversity.
> Manager 1: I always assess if people take the lead, but also if they let others speak."
> (Field notes in-house day)

In sum, the HR team started to experience critiques of managers who had an alternative understanding of fairness, and started to criticize or oppose the accuracy and consistency of the AI assessment.

AI Team Contesting Notions of Fairness

A final mismatch between notions of fairness had to do with ensuring consistency in selection and retaining "humans in the loop" in order to ensure correctable decisions. Several months after the launch of the AI application, the AI team discovered that the hiring algorithm was not used in the way they imagined: not all candidates were assessed and selected based on the algorithmic recommendation, with one group automatically passing to the next step of the selection process, thereby escaping the algorithmic filter. As explained by the PA manager during an AI team meeting:

> "Realistically, the selection of the people for the stages was done by people. By the recruiters. So, not all of them started to use the hiring algorithm at the same moment, and one of the recruiters was using the wrong algorithm for another week." (Field notes AI team meeting)

This inconsistency in the use of the algorithm was problematic for the PA manager as it implied that the obtained data was in fact contaminated, compromising a possibility of rigorous comparative analysis on the candidates before and after the use of the algorithm. During an HR and AI team meeting, the PA manager shared her concerns: "This is a challenge for me, because the assessment was not fair there" (Field notes meeting HR team and AI team). In turn, the inconsistent use of the algorithm revealed to the HR team that surprisingly some candidates did in fact receive hiring offers, despite the algorithm predicting the opposite. This resulted in the HR team doubting the validity of the AI application, as illustrated by an HR professional:

> "So, Alex, this is the only one guy who got the offer for the sales program in Spain. He declined it, but he got it. With a video score of 69, and a game score of 83. But actually, he was not assessed by the algorithm. […] And actually, right now, he would be rejected. And he is the only one guy that got the offer. So, candidate's success is unpredictable actually, I would say."
> (Field notes weekly HR team meeting)

During a team meeting in Spring 2019, the doubts of the HR team culminated: while the PA manager proudly presented the latest analytic results, the HR manager criticized the validity of the results. The results showed that top-scoring candidates (i.e. candidates with match scores of 92%) were rejected by human assessors and did not receive any hiring offers. This resulted in a fierce discussion between the HR manager and PA manager:

> "HR manager: Maybe it is also a question about whether this [AI assessment] actually works?
> AI manager: You mean, does the human assessment work? [..]
> HR manager: But [we can criticize either] the human assessment or the AI assessment, right? It can be one or the other failing. They are clearly not matching."
> (Field notes meeting HR and AI team)

The groups thus experienced a mismatch between their different notions of fairness: while the HR manager considered unfairness to be AI's problem due to the invalid assessment of candidates, the PA manager blamed human assessors due to their inconsistency and potentially biased assessment.

Conclusion

In sum, implementing and using AI in recruitment practices brings to the fore the role of fairness in organizational decision-making. Our analysis of the specific practices and interactions of multiple stakeholders in the workplace shows that enabling fairness with AI can take a very different shape from what it promised, when put into practice. In particular, before the use of the AI application, the meaning of fairness was considered unproblematic and shared between stakeholder groups. However, as the various

groups started working with AI in practice, they experienced mismatches between those notions of fairness that were inscribed and those implicit understandings of fairness that were important for daily work. Our preliminary findings thus illustrate that while stakeholder groups can discursively agree upon AI as ways to enact ethical values, these values become reconsidered and negotiated once people start to interact with AI and put it into practice.

This study adds to research on technology and work by furthering our understanding of the crucial role that ethical values in general, and the ethical value of fairness in particular, play in the development and use of algorithmic technologies in practice. Previous research has emphasized how work gets reconfigured as a result of algorithmic technologies increasing people's dependencies on machines (Newell and Marabelli 2015), transforming the standards for valuation of work (Orlikowski and Scott 2014), or enabling power shifts in work relations (Barbour et al. 2018). Our findings extend this line of research by demonstrating the importance of AI in triggering the negotiation of ethical values. Our ethnographic story shows that we cannot understand the role of AI and ethics without studying this in practice. This study contributes as well to research on sociotechnical systems that highlight the importance of considering the societal context that surrounds AI applications for achieving fair decision-making. By showing how inscribed fairness ideals can fail to account for the contextual, temporal, and constable nature of fairness, we empirically illustrate what happens after organizations fall into the "formalism trap", i.e. fail to account for the full meaning of social concepts (Selbst et al. 2019). Beyond implications for future research, our preliminary findings also provide an in-depth case of AI in practice that serves as a caution to those who have too high expectations of AI as ways to achieve fairness in organizations.

References

Barbour, J. B., Treem, J. W., and Kolar, B. 2018. "Analytics and Expert Collaboration: How Individuals Navigate Relationships When Working with Organizational Data," *Human Relations* (71:2), pp. 256-284.

Faraj, S., Pachidi, S., and Sayegh, K. 2018. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28:1), pp. 62-70.

Gilliland, S. W. 1993. "The Perceived Fairness of Selection Systems: An Organizational Justice Perspective," *Academy of management review* (18:4), pp. 694-734.

Greenberg, J. 1990. "Organizational Justice: Yesterday, Today, and Tomorrow," *Journal of management* (16:2), pp. 399-432.

Hoffmann, A. L. 2019. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse," *Information, Communication & Society* (22:7), pp. 900-915.

Langley, A. 1999. "Strategies for Theorizing from Process Data," *Academy of Management review* (24:4), pp. 691-710.

Leventhal, G. S. 1980. "What Should Be Done with Equity Theory?," in *Social Exchange*. Springer, pp. 27-55.

Madden, M., Gilman, M., Levy, K., and Marwick, A. 2017. "Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans," *Wash. UL Rev.* (95), p. 53.

Newell, S., and Marabelli, M. 2015. "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification'," *The Journal of Strategic Information Systems* (24:1), pp. 3-14.

Orlikowski, W. J., and Scott, S. V. 2014. "What Happens When Evaluation Goes Online? Exploring Apparatuses of Valuation in the Travel Sector," *Organization Science* (25:3), pp. 868-891.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. 2019. "Fairness and Abstraction in Sociotechnical Systems," *Proceedings of the Conference on Fairness, Accountability, and Transparency*: ACM, pp. 59-68.