Special Issue: New Technologies for Business and Society: Achieving Multiple Goals With Multiple Types of Organizations

Business & Society
1–43

DOI: 10.1177/00076503261448488
journals.sagepub.com/home/bas
Sage

Is It Fair to be Accurate? Moral-Emotional Responses to Organizations' AI Orientation Choices

Flore Vancompernolle Vromman[1], Corentin Hericher[1], Corentin Vande Kerckhove[2], and Nicolas Raineri[3]

Abstract

While research on algorithmic decision-making has grown substantially, little is known about people's moral reactions to organizations' artificial intelligence (AI) orientation choices. Drawing on deonance theory, we hypothesize that an organization's choice between an algorithm maximizing accuracy at the expense of fairness and one prioritizing fairness over accuracy triggers distinct moral-emotional responses among third-party observers. We conducted three vignette-based experiments comparing accuracy- and fairness-oriented algorithms in hiring (Studies 1 and 3) and dismissal (Study 2), with different degrees of accuracy loss (Study 3). Results indicate that moral emotions (i.e., other-condemning and other-praising) mediate the effects of this choice on observers' behavioral responses (i.e., negative and positive word-of-mouth) toward the organization. By highlighting how accuracy–fairness trade-offs shape observers' moral appraisals of organizations, this

[1]UCLouvain – Louvain Research Institute in Management and Organizations, Louvain-la-Neuve, Belgium
[2]UCLouvain – Louvain Research Institute in Management and Organizations, Mons, Belgium
[3]ICN Business School, Université de Lorraine, CEREFIGE, Nancy, France

Corresponding Author:
Flore Vancompernolle Vromman, UCLouvain – Louvain Research Institute in Management and Organizations, Place des Doyens, 1, Louvain-la-Neuve 1348, Belgium.
Email: flore.vromman@gmail.com

article advances management research on algorithmic decision-making and extends deonance theory to algorithmic human resource management, establishing AI orientation choices as a moral context informing observers' approval or disapproval of organizations.

Keywords
algorithmic fairness, accuracy–fairness trade-off, moral emotions, ethical AI, deonance theory

Introduction

Artificial intelligence (AI), defined as the use of algorithms that enable machines to learn from data, reason, and act in ways that replicate or supplement human intelligence (Russell & Norvig, 2021), is increasingly impacting decision-making in organizations and society (Correani et al., 2020; van den Broek et al., 2025). AI is used in employment (Hunkenschroer & Luetge, 2022), healthcare (Kaur et al., 2022), credit markets (Bansak & Paulson, 2024), law enforcement (Meijer et al., 2021), and the judicial system (Martin, 2019b), raising ethical concerns in terms of fairness and discrimination (Bankins et al., 2024; Chhillar & Aguilera, 2022; Kordzadeh & Ghasemaghaei, 2022). A well-known example is Amazon's recruitment AI, which systematically disadvantaged female candidates after being trained on historical company data (Dastin, 2022), consistent with broader evidence that AI algorithms, although designed for efficiency and accuracy, may both err and reproduce societal biases and prejudices (Barocas et al., 2023; Martin, 2019a; Mehrabi et al., 2022; Selbst et al., 2019).

In response to these concerns, research on algorithmic fairness has primarily developed in computer science, notably around the accuracy–fairness trade-off (Kleinberg et al., 2016; Menon & Williamson, 2018; Zhao & Gordon, 2022) and algorithmic biases and discrimination (Mehrabi et al., 2022; S. Mitchell et al., 2021). In parallel, management research has examined how individuals perceive and respond to algorithmic outcomes, mainly by contrasting human and algorithmic decision-making in organizational uses of selection algorithms (Bonezzi & Ostinelli, 2021; Kern et al., 2022; Lee, 2018; Newman et al., 2020; Yan et al., 2024). However, this stream of research has rarely considered how algorithms themselves may differ in the values they embody (Martin, 2019a, 2019b; van den Broek et al., 2025), limiting our understanding of how individuals evaluate algorithmic decision-making, particularly when organizations face choices between alternative algorithmic orientations with moral implications. Building on this stream of

management research, our article extends this line of work by connecting the accuracy–fairness trade-off with individuals' responses to selection algorithms, an area that remains largely unexplored (Yan et al., 2024).

To date, empirical management studies have yielded fragmented insights into how individuals evaluate algorithmic decisions, often focusing on performance or fairness in isolation (Bansak & Paulson, 2024). Research highlights that individuals may exhibit greater aversion toward accurate algorithms than toward less accurate human counterparts (Dietvorst et al., 2015), often resulting in lower trust and acceptance among users (Efendić et al., 2020), applicants (McGuire & De Cremer, 2023; Wesche et al., 2024), and the general public (Lee, 2018; Yan et al., 2024). Regarding fairness, findings remain mixed: algorithms are sometimes viewed as fairer due to their consistent application of decision rules (e.g., Bonezzi & Ostinelli, 2021; Wesche et al., 2024), less fair due to their limited sensitivity to context (e.g., Lee, 2018; Newman et al., 2020; Yan et al., 2024), or not meaningfully different from human decision-makers (e.g., Kern et al., 2022; Ötting & Maier, 2018). While this body of research advances our understanding of how individuals evaluate algorithmic decisions, it offers only limited insight into the moral and social implications of selection algorithms that embody different value orientations, such as accuracy or fairness.

As AI progressively replaces human decision-making in organizations (Lee, 2018; Moser et al., 2022), understanding reactions to such value-laden algorithms becomes increasingly salient, particularly in human resource management (HRM), where algorithmic selection affects employment-related decisions (Martin, 2019a; McGuire & De Cremer, 2023; Wesche et al., 2024). In this regard, our study moves beyond prior debates over whether AI is more or less fair than humans by examining how individuals respond when organizations adopt accuracy- versus fairness-oriented algorithms in HRM decision-making.

To explain these reactions, we draw on deonance theory (Folger, 2001; Folger & Glerum, 2015), which helps account for the moral emotions elicited by actions that affirm or undermine fairness norms in response to organizations' choices between algorithmic orientations. Moral emotions, which Haidt (2003) characterizes as emotions that concern the common good of society or the well-being of others, are a core motivational force of justice in deonance theory shaping individual behavior (Folger & Shukla, 2019; Greenbaum et al., 2020). This perspective aligns with a third-party observer view, whereby individuals morally appraise organizational actions that affect others rather than themselves (Lee, 2018; Yan et al., 2024).

Specifically, we advance that accuracy-oriented algorithms that prioritize performance over fairness may evoke more negative other-condemning

moral emotions, reducing people's support for the organization through expressions of disapproval, such as negative word-of-mouth. In contrast, fairness-oriented algorithms that allocate outcomes (e.g., being hired) equitably across demographic groups (e.g., by gender) may evoke more positive other-praising moral emotions, thereby prompting positive word-of-mouth as an expression of approval toward the organization. We therefore ask whether organizations' choices between accuracy- and fairness-oriented algorithms elicit distinct moral emotions that translate into individuals' approval or disapproval toward the organization.

Accordingly, we examine individual moral-emotional responses and subsequent behavioral reactions to a company's choices between alternative algorithmic orientations in three vignette-based experiments. Consistent with computer science research showing that accuracy and fairness rest on different algorithmic constraints and cannot be simultaneously maximized (Kleinberg et al., 2016; Menon & Williamson, 2018; Zhao & Gordon, 2022), our three experiments contrasted an accurate but gender-biased algorithm with a fair, unbiased yet less accurate algorithm across different contexts: hiring in Studies 1 and 3 and dismissal in Study 2. Study 3 further varies the accuracy–fairness trade-off across two contrasting accuracy levels of the fairness-oriented algorithm (Kleinberg et al., 2016; Nussberger et al., 2022).

Our research makes three main contributions. First, we contribute to the management literature on algorithmic decision-making from a psychological perspective by extending deonance theory to the context of algorithmic HRM. Specifically, we provide insights into how organizations' choices between alternative, value-laden algorithms constitute a moral context reflecting adherence to or deviation from fairness-related norms, which elicit moral-emotional responses from third-party observers. Second, we advance research on algorithmic fairness by focusing on individuals' moral evaluations of organizational choices between accuracy- and fairness-oriented algorithms. We demonstrate that such choices give rise to morally meaningful tensions and that approval of fairness-oriented algorithms depends on the level of accuracy loss. Third, we extend understanding of the organizational implications of algorithmic decision-making by linking moral appraisals of algorithmic orientation choices to approval or disapproval toward the organization.

Theoretical Background and Research Hypotheses

Morality and Fairness of Algorithmic Decision-Making

Although ethical concerns related to the use of AI are not new, the rapid development of AI in organizations and society is relatively recent, and

research on its effects and risks remains relatively limited (Bigman & Gray, 2018; da Motta Veiga et al., 2023; Mehrabi et al., 2022). According to Martin (2019a), two aspects of this development point to key issues related to ethical AI, which has been defined as "the fair and just development, use, and management of AI technologies" (Bankins & Formosa, 2021, p. 60). First, the replacement of human judgment with algorithmic decisions (McGuire & De Cremer, 2023) is often accompanied by insufficient organizational reflection on this substitution and on the governance of algorithmic decision-making (Martin, 2019a; Moser et al., 2022). Second, the growing adoption by organizations of algorithmic decision-making creates broader, potentially systemic impacts on multiple stakeholders, such as employees, consumers, and society at large (Bankins et al., 2024; Breidbach, 2024; Marabelli et al., 2021).

Against this backdrop, ethical AI emphasizes core issues associated with algorithm governance, with a specific focus on accountability (responsibility for algorithmic decisions) and fairness (mitigating discriminatory outcomes), alongside other dimensions such as explainability (making algorithms understandable to stakeholders) and privacy (protecting individual data) (Bankins & Formosa, 2021; Dignum, 2019; Kaur et al., 2022). Existing contributions regarding the attribution of responsibility for algorithmic outcomes are largely theoretical, highlighting that algorithms cannot make moral judgments (Dignum, 2019; Martin, 2019a). Unlike humans, who can show empathy and take values and context into account, algorithms simply follow programmed rules and perform calculations (Moser et al., 2022; Russell & Norvig, 2021). Therefore, delegating decisions with moral implications to algorithms, including employment-related decisions (e.g., hiring, promotion, and dismissal), raises questions about the moral responsibility of organizations adopting them (Chhillar & Aguilera, 2022; da Motta Veiga et al., 2023; Moser et al., 2022). The few existing empirical studies corroborate this concern, showing that people tend to perceive algorithmic decisions in moral contexts as less fair and less legitimate compared with human decisions (Bigman & Gray, 2018; McGuire & De Cremer, 2023). Overall, this supports the view that choices about algorithms are not only technical but also moral, shaping how organizations are judged by others (Yan et al., 2024).

A key moral challenge is that algorithmic outcomes can create unfairness (S. Mitchell et al., 2021; Pessach & Shmueli, 2020). Since AI relies heavily on data to train and make decisions, incomplete or biased data can yield less accurate results for some groups and disadvantage people based on sensitive attributes such as gender, ethnicity, age, or disability (Chhillar & Aguilera, 2022; Kleinberg et al., 2018; Mehrabi et al., 2022), leading to potential restrictions of rights (Martin, 2019a). These impacts are pervasive, but they are also more likely to be noticed and reported as awareness of the risks

associated with AI failures and biases has increased (Bankins et al., 2024). In HRM contexts, where decisions affect people's careers and livelihoods, such issues of (un)fairness are especially salient (McGuire & De Cremer, 2023; Newman et al., 2020), contributing to the moral standing of organizations that adopt algorithm-based decision systems (Kordzadeh & Ghasemaghaei, 2022).

Yet research findings on the use of algorithms in HRM remain contested. On the one hand, some studies show that people may prioritize efficiency over fairness (Bansak & Paulson, 2024) and value the consistency of algorithmic rules that promise objectivity and accuracy (Bonezzi & Ostinelli, 2021; Kern et al., 2022). On the other hand, other studies suggest that algorithmic decision-making is often perceived as less fair because it lacks sensitivity to subjective and contextual factors (Lee, 2018; Newman et al., 2020). Adding nuance, Yan et al. (2024) showed that unless algorithm adoption is employee-oriented (i.e., motivated by fairness) rather than company-oriented (i.e., motivated by efficiency), associated HRM decisions undermine deontological principles of respectful employee treatment. Taken together, these findings highlight the need to examine how individuals respond when organizations adopt algorithms oriented toward either accuracy or fairness in employment-related decisions.

The Accuracy–Fairness Trade-off

Fairness has been approached from the organizational justice perspective in management research, emphasizing distributive, procedural, and interactional dimensions (Colquitt et al., 2001). In computer science, by contrast, scholars have advanced technical definitions of fairness grounded in different mathematical formalisms and normative assumptions (Barocas et al., 2023; Castelnovo et al., 2022; Pessach & Shmueli, 2020). In this tradition, fairness is approached from the perspective of algorithmic bias (Kordzadeh & Ghasemaghaei, 2022) and quantified by statistical measures such as "demographic parity," which requires an algorithm to assign the same proportion of positive outcomes (e.g., being hired, receiving a loan, or gaining access to treatment) across demographic categories (e.g., gender, ethnicity, or age) (Castelnovo et al., 2022; Corbett-Davies et al., 2024; van den Broek et al., 2025). In line with our focus on accuracy- versus fairness-oriented algorithms, we adopt the dominant definition in the algorithmic literature, which conceptualizes fairness as the reduction of bias against protected groups, that is, demographic groups recognized as vulnerable to discrimination (Barocas et al., 2023; Castelnovo et al., 2022; Kleinberg et al., 2018; Pessach & Shmueli, 2020).

Within this literature, researchers have developed fairness-oriented algorithms (also referred to as fairness-aware algorithms; Fabris et al., 2022), defined as algorithms that explicitly integrate fairness criteria into their design (Le Quy et al., 2022). Fairness can be integrated into algorithmic design at different stages of the process: by adjusting the data before training (pre-processing), by incorporating fairness constraints into the model during training (in-processing), or by modifying the predictions after training (post-processing) (Pessach & Shmueli, 2020). Fairness-oriented algorithms differ from more standard accuracy-oriented algorithms that optimize performance without explicitly considering fairness (Barocas et al., 2023; Green & Hu, 2018).

In practice, algorithmic performance is primarily assessed based on predictive accuracy, measured as the proportion of predictions that correctly reflect actual outcomes (Hastie et al., 2001). Consequently, integrating fairness into algorithm design often comes at the expense of accuracy because ensuring equitable outcomes across groups (e.g., demographic parity) requires relaxing the statistical criteria that typically maximize predictive performance, which rely on real-world data that often reflect unbalanced distributions (Kleinberg et al., 2016; Zhao & Gordon, 2022). The accuracy loss of incorporating fairness can vary in magnitude, meaning that fairness-oriented choices may involve smaller or larger performance sacrifices (Menon & Williamson, 2018).

As a result, organizations seeking to implement algorithmic decision-making in their management practices may face difficult choices: prioritizing fairness can not only reduce unequal access to opportunities but also undermine efficiency, while focusing on predictive accuracy risks reproducing existing inequalities. Despite this dilemma, most research has conceptualized the trade-off mathematically, with limited attention to its social and moral implications (Barocas et al., 2023; Kleinberg et al., 2016). Scholars have emphasized that the literature largely abstracts away from real-world settings (Selbst et al., 2019) and provides little understanding of how such trade-offs are perceived in practice, particularly in consequential employment decisions (Kordzadeh & Ghasemaghaei, 2022).

With the increasing use of algorithms that can perpetuate or mitigate discrimination in HRM decisions, it is therefore important to understand how individuals react to organizations' algorithmic choices. Individuals may respond to algorithm design not only because it shapes their own work experience but also because it reflects the organization's moral stance toward fairness and its implications for others. Algorithmic biases are often framed in relation to moral values and ethical standards; therefore, individuals' moral perceptions can inform the assessment of such biases (Kordzadeh &

Ghasemaghaei, 2022). Yet little is known about how algorithm design choices that prioritize either accuracy or fairness shape people's moral responses toward organizations (Yan et al., 2024). To address this gap in knowledge, we draw on deonance theory (Folger, 2001; Folger & Glerum, 2015), which helps account for how individuals react when organizational practices are perceived as affirming or undermining moral norms.

Deonance Theory

According to deonance theory (Folger, 2001; Folger & Glerum, 2015; Folger & Shukla, 2019), people have an innate tendency to make moral appraisals of the actions of social actors (whether others or themselves) that either affirm or undermine fairness-related norms (i.e., what is perceived as the right thing to do). Such appraisals are typically made from an observer standpoint rather than as a direct recipient of the actions, reflecting a disinterested concern for fairness and the well-being of others (Folger, 2001; Haidt, 2003). As a result, even when they are not directly affected, individuals experience positive or negative moral emotions, such as other-praising or other-condemning emotions, that motivate efforts to restore moral balance (Folger & Glerum, 2015). Deonance theory builds on Haidt's work on moral emotions, defined as "emotions that are linked to the interests or welfare either of society as a whole or at least of persons other than the judge or agent" (Haidt, 2003, p. 853). From this perspective, actions perceived as unjustifiably harming others are seen as undermining moral norms related to fairness, whereas actions that do good for others are viewed as affirming these norms (Folger & Shukla, 2019; Haidt, 2003).

Empirical studies drawing on deonance theory have thus far placed great emphasis on negative moral emotions (Hericher & Bridoux, 2023; M. S. Mitchell et al., 2015). Yet, management research shows that individuals also respond positively to actors' actions that benefit others and are perceived as doing the right thing (Hericher et al., 2023; Rupp et al., 2006). Greenbaum et al. (2020) classified moral emotions into four categories depending on whether an observed action upholds or undermines moral norms, and whether the target of moral judgment is the actor or the recipient of the action. These four categories differ in their moral focus: other-condemning, self-condemning, and other-suffering emotions arise from perceived disregard of moral norms, whereas other-praising emotions emerge when others' actions uphold such norms (Haidt, 2003).

Furthermore, deonance theory suggests that, because they stem from moral emotions, behavioral responses to reprehensible actions are often spontaneous and sometimes pursued as an end in themselves (Folger et al.,

2005). The punitive reactions of individuals who observe a failure of fairness are central to deonance theory (Colquitt & Zipay, 2015) and play an important role in sustaining fairness-related norms and restoring moral balance (Hericher & Bridoux, 2023). Conversely, when actors' actions are perceived as fulfilling fairness norms, individuals experience positive moral emotions and feel motivated to endorse or support the actor responsible (Folger & Glerum, 2015; Haidt, 2003).

Building on deonance theory, we argue that choices between value-laden algorithmic orientations (i.e., accuracy vs. fairness) embody moral norms concerning fair and respectful treatment of employees (Yan et al., 2024). In this respect, prior research shows that algorithmic selection decisions can elicit moral-emotional responses from observers, such as pride or anger, arising from perceptions of fair or unfair treatment (Lee, 2018). Accordingly, extending deonance theory to the context of algorithmic HRM, we conceptualize organizations' choices between accuracy- and fairness-oriented algorithms as forming a moral context for observers, reflecting adherence to or deviation from fairness-related norms, with implications for moral emotions and associated approval or disapproval toward the organization (i.e., positive or negative word-of-mouth).

Accuracy- Versus Fairness-Oriented Algorithms and Moral Emotions

Accuracy- and fairness-oriented algorithms have different implications for individuals and society, notably regarding their moral significance in relation to algorithmic bias and discrimination (Kordzadeh & Ghasemaghaei, 2022). Accuracy-oriented algorithms can inadvertently reinforce existing inequalities and prejudices, whereas fairness-oriented algorithms aim to mitigate unequal treatment between demographic groups (Barocas et al., 2023; Martin, 2019a; Mehrabi et al., 2022; Selbst et al., 2019). Relatedly, recent work in the psychology of algorithms, particularly in HRM and recruitment contexts, shows that people interpret algorithmic decisions through moral and fairness-related lenses (Bigman & Gray, 2018; Lee, 2018; Newman et al., 2020) and assess their fairness depending on whether these decisions reflect concern for how candidates and employees are treated rather than a focus on organizational efficiency (Yan et al., 2024). That is, organizational choices in algorithmic design, and especially the primacy given to accuracy or fairness, attest to the values of the organization. Building on deonance theory (Folger, 2001), we argue that when organizations delegate employment-related decisions such as hiring or dismissal to accuracy- versus fairness-oriented algorithms,

these choices either affirm or undermine fairness-related norms, thereby signaling whether the organization fulfills its obligation to treat others fairly (Yan et al., 2024).

When an actor's actions fail to adhere to fairness-related norms, observers experience other-condemning moral emotions (Greenbaum et al., 2020; Haidt, 2003), characterized by negative moral feelings toward those who disregard such norms. These emotions include anger, contempt, and disgust, collectively referred to as the hostility triad (Rozin et al., 1999). Empirical research shows that decisions made by selection algorithms, compared with those made by humans, elicit stronger negative emotions such as anger, as their computational nature typically lacks sensitivity to qualitative and contextual aspects (Lee, 2018; see also Newman et al., 2020). In line with this, studies show that individuals morally condemn organizations when algorithmic decisions are perceived as unfair or discriminatory (McGuire & De Cremer, 2023; Yan et al., 2024). Evidence also shows that observers hold organizations accountable for adhering to fairness-related norms and experience other-condemning moral emotions when organizational actions cause harm or negative consequences to others (Antonetti & Maklan, 2016; Hericher & Bridoux, 2023). Thus, when organizations rely on accuracy-oriented algorithms that disregard fairness principles to make decisions that affect people's lives, such as determining who is hired or who loses their job, these choices are likely to provoke moral condemnation among observers.

Conversely, organizations that adopt fairness-oriented algorithms can be viewed as fulfilling their moral obligation to do the right thing, which is likely to evoke positive moral emotions toward them. Other-praising moral emotions are the counterpart of other-condemning ones and refer to "positive feelings that occur when another person upholds moral standards" (Greenbaum et al., 2020, p. 96). These emotions, comprising gratitude and elevation, arise when observers perceive others' actions as fair and morally appropriate. Empirical research highlights that witnessing positive or prosocial actions generates both feelings of gratitude (Ford et al., 2018; Xie et al., 2015) and moral elevation (Hericher et al., 2023; Vianello et al., 2010). Accordingly, when organizations adopt fairness-oriented algorithms in hiring and dismissal decisions, they are likely to elicit other-praising moral emotions among individuals, as this conveys the organization's effort to counter potential algorithmic bias and promote equitable treatment (Martin, 2019a; Yan et al., 2024).

Hypothesis 1a: Accuracy-oriented algorithms elicit stronger other-condemning moral emotions than fairness-oriented algorithms.

Hypothesis 1b: Fairness-oriented algorithms elicit stronger other-praising moral emotions than accuracy-oriented algorithms.

From Moral Emotions to Behavioral Responses

A key characteristic of moral emotions is that they act as a motivational force driving behavioral responses that reflect judgments about right and wrong (Haidt, 2003). Deonance theory posits that moral emotions mediate the relationship between a situation and subsequent behaviors directed toward the actor whose actions affirm or undermine fairness-related norms (Folger & Shukla, 2019). Other-condemning emotions lead to reparative actions or behaviors that express disapproval of perceived unfairness (Greenbaum et al., 2020; Haidt, 2003). Prior research shows that experiencing negative moral emotions toward an organization triggers reactions such as negative word-of-mouth, that is, speaking unfavorably about the organization to others (Hericher & Bridoux, 2023), as well as other behaviors that symbolically sanction the organization, such as moral protest, reflecting disapproval and the desire to restore fairness (Antonetti & Maklan, 2016).

By contrast, other-praising emotions lead to supportive behaviors toward the actor or organization perceived as affirming fairness-related norms (Ford et al., 2018; Greenbaum et al., 2020; Haidt, 2003; Vianello et al., 2010). Research provides evidence that when observers experience other-praising moral emotions in response to an organization's actions, they are more likely to engage in discretionary behaviors that express support for the organization, such as positive word-of-mouth, a widespread and low-cost way for individuals to convey moral support, much like negative word-of-mouth expresses moral disapproval (Antonetti & Maklan, 2016; Hericher & Bridoux, 2023).

Hypothesis 2a: Accuracy-oriented algorithms lead to stronger negative word-of-mouth through other-condemning moral emotions than fairness-oriented algorithms.

Hypothesis 2b: Fairness-oriented algorithms lead to stronger positive word-of-mouth through other-praising moral emotions than accuracy-oriented algorithms.

Research Design and Results

Following ethical clearance from our research institute's ethics committee, we tested our research hypotheses using three vignette-based experiments, each conducted with an independent sample. Vignette-based experiments

allow researchers to establish causal relationships by manipulating independent variables in written scenarios to which participants are randomly and evenly assigned (Atzmüller & Steiner, 2010). This method is particularly well suited for studies on ethics and morality, as it enables the quantitative investigation of sensitive issues that are often difficult to examine in real-world settings using other approaches, such as field surveys (Aguinis & Bradley, 2014). Moreover, consistent with deonance theory, participants evaluated organizational actions that affected others rather than themselves, allowing us to capture moral emotions based on disinterested concern for others (Lee, 2018; Yan et al., 2024). Accordingly, we used Prolific Academic, a crowdsourcing research platform that provides access to a wider range of participants and higher data quality compared to other online platforms (Peer et al., 2017), to recruit a more diverse and thus theoretically appropriate sample for our study (Aguinis & Bradley, 2014).

Each vignette-based experiment tested all four research hypotheses. Studies 1 and 2 examined participants' moral emotions and subsequent behavioral responses using a two-level manipulation of a company's decision to adopt either an accuracy-oriented algorithm (i.e., highly accurate but gender-biased) or a fairness-oriented algorithm (i.e., unbiased but less accurate) in the contexts of hiring and dismissal, respectively. In Study 1, the accuracy-oriented algorithm accurately detected curricula vitae (CVs) matching the set of skills required for the job but favored male over female candidates with equal qualifications, whereas the fairness-oriented algorithm was less accurate but ensured equal opportunities. Study 2 replicated the design of Study 1 in a dismissal context, and Study 3 extended it by testing whether an accuracy-oriented algorithm achieving 100% performance would elicit stronger other-condemning and weaker other-praising emotional and behavioral responses than a fairness-oriented algorithm performing at two empirically representative and contrasting accuracy levels (90% and 60%; see Kleinberg et al., 2016; Nussberger et al., 2022), within a four-level manipulation design. Appendix A presents the experimental vignettes used in each study.

We selected scenarios from HRM focused on gender discrimination for two main reasons. First, employment decisions such as hiring and dismissal carry significant consequences for both individuals and society and have been shown to perpetuate gender inequalities in access to opportunities (Hunkenschroer & Luetge, 2022; Kordzadeh & Ghasemaghaei, 2022). Second, research and practice have further shown that AI-driven biases are especially acute in hiring, where algorithmic decision-making often reproduces or amplifies human biases against women and other minority groups (Dastin, 2022; Mehrabi et al., 2022; Nyberg et al., 2025).

We assessed the internal validity of all our manipulations by pretesting each one on independent samples (Podsakoff & Podsakoff, 2019) prior to the main studies (Studies 1–3). We conducted all analyses with R (version 4.4.0) using two-tailed significance tests. The data and analysis scripts supporting this article are publicly available. The dataset is available on Dataverse at https://doi.org/10.14428/DVN/ZGD5DR, and the analysis scripts are available on GitHub at https://github.com/cvandekerckh/flocos.git.

Study 1

Method

Participants and Procedure. We recruited 140 participants online via Prolific Academic, restricting eligibility to native English speakers born and residing in the United Kingdom who had a high approval rate history on Prolific based on prior researchers' evaluations, to minimize the risk of low-quality responses (Aguinis et al., 2021; Chandler et al., 2014). Participants received an invitation to complete our questionnaire through the crowdsourcing platform, read information about the study's purpose and ethical approval, and provided informed consent. After being randomly and evenly assigned to one of the two experimental conditions, they read the scenario vignette and then completed the questionnaire.

Prior to analysis, we screened for careless responses (Aguinis et al., 2021) using an attention check and participants' response time. The attention check consisted of an open-ended question asking participants to recall the name of the company mentioned in the vignette they had just read. We excluded respondents who failed the attention check or had an average response time below one second (Wood et al., 2017) or above 15 seconds per item, which likely indicates that they were busy with other tasks while responding (Hericher & Bridoux, 2023), resulting in a final sample of 113 participants[1]. This sample size is consistent with the recommendation of Lonati et al. (2018), who identified 50 participants per condition as the minimum required to avoid randomization failure in simple empirical settings.

Among the 113 participants, 50.4% identified as female, with a mean age of 39.3 years ($SD = 11.1$) and an average work experience of 20.1 years ($SD = 11.3$). With respect to educational level, 35.3% had post-secondary qualifications (i.e., A-levels), 42.5% held an undergraduate degree, and 22.2% a graduate degree or higher.

Design and manipulation. To test our hypotheses, Study 1 used a pretested, between-subjects vignette-based experiment with two conditions, in which

participants read that a company, named FloCos, calibrated its AI for hiring by choosing between two options (Appendix A presents the full text of the experimental vignettes):

> With option A, the AI is very accurate at detecting good CVs. It reduces the risk that FloCos hires candidates with a mismatching set of skills. However, at equal skills, male candidates have more chances to be hired than other candidates.
>
> With option B, the AI is less accurate at detecting good CVs than in option A. It increases the risk that FloCos hires candidates with a mismatching set of skills. However, at equal skills, male and female candidates have equal chances to be hired.

We randomly assigned participants to one of the two scenarios in which the company selected option A (accuracy-oriented algorithm) or option B (fairness-oriented algorithm). They then completed randomized items of moral emotions (other-praising and other-condemning), word-of-mouth (positive and negative), and the control and demographic variables.

Measures. We measured other-praising moral emotions (Cronbach's $\alpha = .93$) on a 6-item scale (Romani & Grappi, 2014), and other-condemning moral emotions ($\alpha = .97$) on a 9-item scale (Xie et al., 2015). We administered the randomized items together with fillers to minimize potential biases. For word-of-mouth, we used the 4-item negative word-of-mouth scale ($\alpha = .95$) by Hericher and Bridoux (2023), rephrasing the items in positive form to measure positive word-of-mouth ($\alpha = .93$). To control for theoretically plausible alternative explanations (Becker et al., 2016) of morality-based reactions, we measured moral identity ($\alpha = .79$) using the 5-item scale by Aquino and Reed (2002). Responses to all items were rated on a 7-point Likert-type scale. Appendix B presents all scales' items and anchors, including those for the manipulation check described below, and Table 1 reports the descriptive statistics and zero-order correlations among the study variables.

Results

Manipulation check. To assess the internal validity of our manipulation, we pretested the scenarios (Podsakoff & Podsakoff, 2019) with an independent sample of 70 participants recruited via Prolific Academic from the same population as in Study 1, applying the same filtering criteria, which resulted in a final sample of 55 participants. We used two *ad hoc* items to measure AI accuracy ("The option that FloCos chose is accurate at detecting good CVs")

Table 1. Descriptive Statistics and Zero-order Correlations.

Variables	α	*M*	*SD*	1.	2.	3.	4.	5.
Study 1								
1. Other-praising moral emotions	.93	2.02	1.25	—				
2. Other-condemning moral emotions	.97	2.17	1.47	−.34***	—			
3. Positive word-of-mouth	.93	3.23	1.44	.34***	−.56***	—		
4. Negative word-of-mouth	.95	3.22	1.71	−.40***	.78***	−.59***	—	
5. Moral identity	.79	6.13	.77	−.06	−.02	.05	.02	—
Study 2								
1. Other-praising moral emotions	.94	2.01	1.27	—				
2. Other-condemning moral emotions	.98	2.50	1.65	−.25**	—			
3. Positive word-of-mouth	.92	3.22	1.37	.75***	−.50***	—		
4. Negative word-of-mouth	.94	3.60	1.64	−.43***	.67***	−.64***	—	
5. Moral identity	.87	6.11	.89	−.06	.12	−.10	.03	—
Study 3								
1. Other-praising moral emotions	.95	2.13	1.40	—				
2. Other-condemning moral emotions	.96	1.94	1.32	−.23***	—			
3. Positive word-of-mouth	.92	3.79	1.31	.59***	−.53***	—		
4. Negative word-of-mouth	.95	3.06	1.58	−.42***	.67***	−.63***	—	
5. Moral identity	.74	6.08	.80	.01	−.06	.06	−.07	—

***$p < .001$. **$= p < .01$.

and AI fairness ("The option that FloCos chose gives equal chances to male and female candidates to be hired").

To check the manipulation, we conducted independent-samples *t*-tests using the R package "psych" (Revelle, 2024), which applies Welch's *t*-tests when the assumption of equal variances across groups is not met. Participants in the accuracy-oriented condition rated the AI as significantly more accurate ($M = 6.11$, $SD = 1.15$) than those in the fairness-oriented condition ($M = 3.00$, $SD = 1.31$, $t_{(52.62)} = 9.37$, $p < .001$, Cohen's $d = 2.52$). Conversely, participants in the fairness-oriented condition rated the AI as significantly fairer ($M = 6.36$, $SD = 1.16$) than those in the accuracy-oriented condition ($M = 1.59$, $SD = 1.37$, $t_{(51.01)} = 13.91$, $p < .001$, $d = 3.83$). These results indicate that the manipulation was successful.

Measurement Model. We conducted confirmatory factor analyses (CFAs) to assess the hypothesized measurement model and the distinctiveness of five latent constructs: other-praising moral emotions, other-condemning moral emotions, positive word-of-mouth, negative word-of-mouth, and moral identity. The goodness of fit was assessed using cutoff values of 0.90 (or higher) for the comparative fit index (CFI) and Tucker–Lewis index (TLI), and 0.08 (or lower) for the standardized root mean square residual (SRMR) (Marsh et al., 2004). The five-factor measurement model demonstrated a good fit ($\chi^2_{(340)} = 610.05$, $TLI = .92$, $CFI = .91$, $SRMR = .07$) and yielded a significantly better fit than a one-factor model in which all items loaded on a single factor ($\Delta\chi^2/\Delta df = 1248.78/10$, $p < .001$; $\chi^2_{(350)} = 1858.83$, $TLI = .54$, $CFI = .50$, $SRMR = .17$), supporting both the convergent and discriminant validity of the study's variables. We report additional model fit analyses based on hierarchical models in Table C1 of Appendix C.

Hypothesis testing. To test the hypothesized direct effects of the company's choice of algorithmic orientation on other-condemning (H1a) and other-praising (H1b) moral emotions, we conducted independent-samples *t*-tests using the R package "psych." The main effect means reported in Figure 1 show that participants reported significantly higher other-condemning moral emotions when the company adopted the accuracy-oriented ($M = 2.58$, $SD = 1.70$) rather than the fairness-oriented algorithm ($M = 1.74$, $SD = 1.02$, $t_{(94.40)} = 3.19$, $p < .01$, $d = 0.60$), supporting H1a. In contrast, participants reported significantly higher other-praising moral emotions in the fairness-oriented ($M = 2.32$, $SD = 1.48$) than in the accuracy-oriented condition ($M = 1.73$, $SD = 0.90$, $t_{(88.15)} = 2.53$, $p < .05$, $d = 0.49$), supporting H1b.

Then, we tested the hypothesized indirect effects of the company's choice of algorithmic orientation on (H2a) negative word-of-mouth via other-condemning

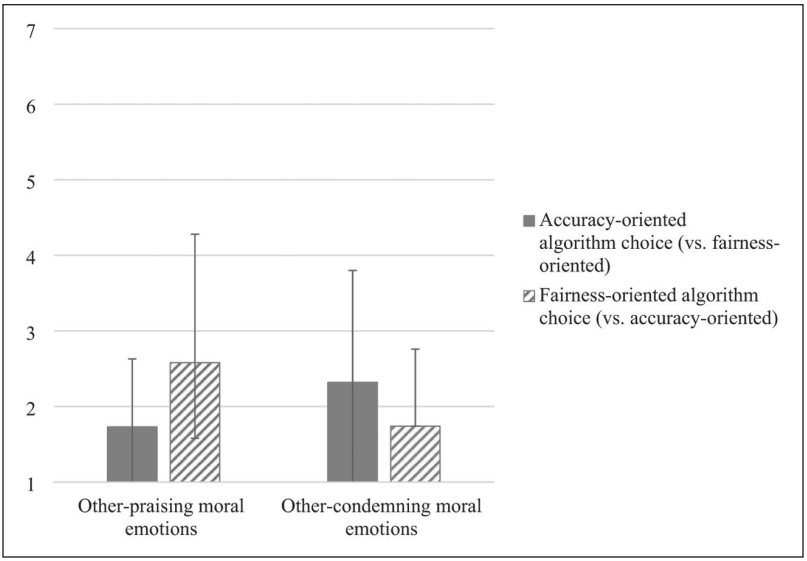

Figure 1. Study 1: Direct Effects of Algorithm Choice on Moral Emotions.

moral emotions, and (H2b) positive word-of-mouth via other-praising moral emotions. We conducted structural equation modeling (SEM) using the R package "lavaan" (Rosseel, 2012) with 10,000 bootstrap resamples to estimate confidence intervals (CIs). In support of H2a, we find positive and significant effects between the choice of accuracy-oriented algorithm and other-condemning ($\beta = .94$, $SE = .27$, $p < .001$) and between other-condemning and negative word-of-mouth ($\beta = .77$, $SE = .08$, $p < .001$). When the company prioritized accuracy, participants reported greater negative word-of-mouth via other-condemning moral emotions (indirect effect = 0.72, $SE = 0.21$, 95% CI [0.33, 1.09]).

Similarly, we find positive and significant effects between the choice of fairness-oriented algorithm and other-praising ($\beta = .54$, $SE = .24$, $p < .05$) and between other-praising and positive word-of-mouth ($\beta = .54$, $SE = .11$, $p < .001$). When the company prioritized fairness, participants reported significantly greater positive word-of-mouth via other-praising moral emotions (indirect effect = 0.30, $SE = 0.13$, 95% CI [0.04, 0.57]), supporting H2b. Taken together, these results provide preliminary evidence that organizations' algorithmic orientation choices are interpreted through a moral lens, eliciting divergent moral-emotional responses toward the organization.

Study 2

Method

Participants. We followed the same procedure and applied the same eligibility criteria as in Study 1 to recruit 140 new participants on Prolific Academic. After screening for careless responding using the same attention check and response-time data as in Study 1, we obtained 110 valid responses for analysis.[1] The final sample consisted of 48.2% female participants, with a mean age of 38.9 years ($SD = 11.1$) and an average of 19.0 years of work experience ($SD = 11.2$). Additionally, 27.1% reported a post-secondary qualification, 50.9% held an undergraduate degree, and 20.0% a graduate degree or higher.

Manipulation. Study 2 replicated the design of Study 1 using a pretested, between-subjects vignette experiment with two conditions. We only modified the HRM context: while Study 1 focused on hiring decisions, Study 2 addressed dismissal decisions. The vignette presented a company, FloCos, that relied on AI to analyze employees' performance reviews and suggest whether to retain or dismiss them. To calibrate the AI, the company had to choose between two options (Appendix A presents the full text of the vignettes):

> With option A, the AI is very accurate at detecting lower performing employees. However, at equal performance reviews, male employees have more chances to keep their job than other employees.
>
> With option B, the AI is less accurate at detecting lower performing employees than in option A. However, at equal performance reviews, male and female employees have equal chances to keep their job.

As in Study 1, we randomly assigned participants to one of the two scenarios in which the company chose option A (accuracy-oriented algorithm) or option B (fairness-oriented algorithm). After reading the vignette, participants completed randomized items measuring moral emotions and word-of-mouth, followed by control and demographic variables.

Measures. We used the same measures and scale anchors as in Study 1 for all variables (see Appendix B), namely other-praising moral emotions ($\alpha = .94$), other-condemning moral emotions ($\alpha = .98$), positive word-of-mouth ($\alpha = .92$), negative word-of-mouth ($\alpha = .94$), and moral identity ($\alpha = .87$).

Results

Manipulation check. We checked the internal validity of our manipulation by pretesting the scenarios with 70 participants recruited on Prolific Academic from the same population as in the main studies (i.e., Studies 1 and 2). We randomly assigned participants to the two conditions and applied the same filtering criteria as the main studies, resulting in a final sample of 60 participants. Study 2 used the same two *ad hoc* items as in Study 1 to assess AI accuracy and fairness (see Appendix B).

Independent-samples *t*-tests confirmed the success of the manipulation. Participants in the accuracy-oriented condition rated the AI as significantly more accurate ($M = 5.50$, $SD = 1.94$) than those in the fairness-oriented condition ($M = 3.10$, $SD = 1.75$, $t_{(57.37)} = 5.03$, $p < .001$, $d = 1.32$). Conversely, participants in the fairness-oriented condition rated the AI as significantly fairer ($M = 6.23$, $SD = 1.38$) than those in the accuracy-oriented condition ($M = 1.63$, $SD = 1.27$, $t_{(57.61)=}13.41$, $p < .001$, $d = 3.52$).

Measurement model. We estimated the latent measurement model using CFAs. The five-factor model showed a good fit ($\chi^2_{(340)} = 569.50$, $TLI = .93$, $CFI = .93$, $SRMR = .06$) and fit the data significantly better than a one-factor model ($\Delta\chi^2/\Delta df = 1363.17/10$, $p < .001$; $\chi^2_{(350)} = 1932.67$, $TLI = .53$, $CFI = .50$, $SRMR = .21$), supporting the convergent and discriminant validity of the measures. We report the hierarchical models in Table C2 of Appendix C.

Hypothesis testing. Supporting H1a, independent-samples *t*-tests show that participants reported significantly higher other-condemning moral emotions when the company opted for the accuracy-oriented algorithm ($M = 2.87$, $SD = 1.65$) rather than the fairness-oriented algorithm ($M = 2.16$, $SD = 1.58$, $t_{(105.34)} = 2.33$, $p < .05$, $d = .70$). Consistent with H1b, participants reported significantly higher other-praising moral emotions in the fairness-oriented condition ($M = 2.40$, $SD = 1.48$) than in the accuracy-oriented condition ($M = 1.57$, $SD = .79$, $t_{(88.82)} = 3.72$, $p < .001$, $d = .70$). These findings in the context of dismissal decisions replicate the pattern observed in Study 1 in the context of hiring decisions. We plot the main effect means in Figure 2.

In support of H2a, results from SEM using the R package "lavaan" show positive and significant effects between the choice of accuracy-oriented algorithm and other-condemning ($\beta = .81$, $SE = .37$, $p < .05$) and between other-condemning and negative word-of-mouth ($\beta = .58$, $SE = .05$, $p < .001$). When the company prioritized accuracy, participants reported greater negative word-of-mouth via other-condemning moral emotions (indirect effect = 0.47, $SE = 0.19$, 95% CI [0.31, 0.85]).

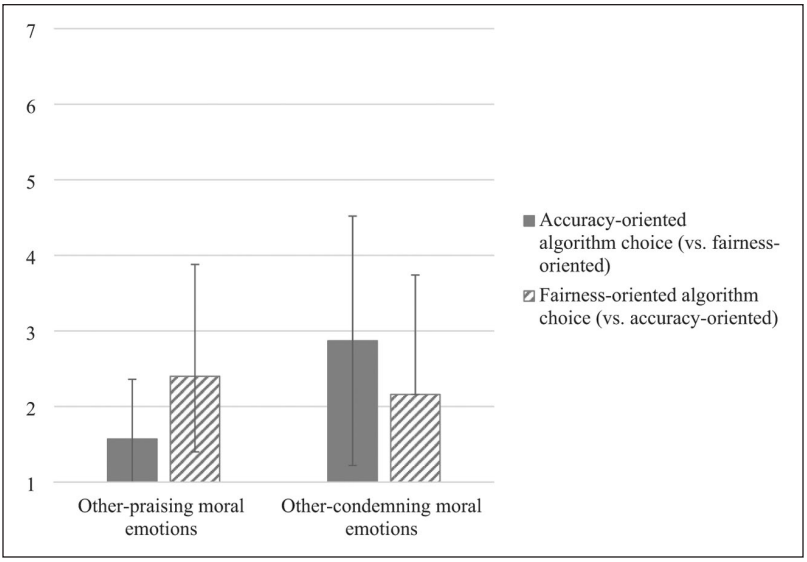

Figure 2. Study 2: Direct Effects of Algorithm Choice on Moral Emotions.

We also found support for H2b, with positive and significant effects between the choice of fairness-oriented algorithm and other-praising ($\beta = .61$, $SE = .18$, $p < .001$) and between other-praising and positive word-of-mouth ($\beta = .84$, $SE = .18$, $p < .001$). When the company prioritized fairness, participants reported greater positive word-of-mouth via other-praising moral emotions (indirect effect = 0.52, $SE = 0.14$, 95% CI [0.30, 0.69]). Consistent with Study 1, these findings indicate that algorithmic orientation choices in a dismissal context elicit similar moral emotions and corresponding approval or disapproval toward organizations.

Study 3

Method

Participants. The purpose of Study 3 was to replicate and extend Study 1 through a two-scenario design comparing an accuracy-oriented algorithm described as achieving 100% performance with a fairness-oriented algorithm displaying either 90% (Scenario 1) or 60% accuracy (Scenario 2). We recruited 300 participants through Prolific Academic to address the four experimental conditions, using the same procedure and eligibility criteria as

in Studies 1 and 2. After screening, we obtained 261 valid responses for analysis.[1] Participants were 49.4% female, with a mean age of 38.2 years ($SD = 10.0$) and an average of 18.6 years of work experience ($SD = 10.0$). Regarding education, 35.6% had post-secondary qualifications, 41.0% an undergraduate degree, and 23.4% a graduate or higher degree.

Manipulations. Study 3 replicated the two-level design of Study 1 and added a second manipulation whereby the fairness-oriented algorithm was described as achieving 90% accuracy in Scenario 1 and 60% in Scenario 2. The accuracy-oriented algorithm was described as always accurate (i.e., 100%) in both scenarios. Consistent with the notion of classification accuracy (Hastie et al., 2001), the percentages presented in the vignettes indicated the proportion of hired candidates possessing the required job skills (e.g., 90% accuracy means that 9 out of 10 hired candidates have the required skills).

We randomly assigned participants to four experimental groups ($N = 75$ per condition). In Scenario 1, the company "FloCos" decided between the accuracy-oriented (option 1A) and the 90%-accuracy fairness-oriented algorithm (option 1B). In Scenario 2, there was a greater trade-off between the accuracy-oriented (option 2A) and the 60%-accuracy fairness-oriented algorithm (option 2B). As in Studies 1 and 2, participants read the vignette and then completed randomized items assessing moral emotions and word-of-mouth, followed by control and demographic variables. Appendix A provides the full vignette scenarios for each condition.

Measures. We used the same measures and scale anchors as in Studies 1 and 2 for all variables (see Appendix B): other-praising moral emotions ($\alpha = .95$), other-condemning moral emotions ($\alpha = .96$), positive word-of-mouth ($\alpha = .92$), negative word-of-mouth ($\alpha = .95$), and moral identity ($\alpha = .74$).

Results

Manipulation check. Following the procedure adopted in Studies 1 and 2, we pretested the vignette scenarios for Study 3 with 200 participants randomly assigned to each of the four conditions, resulting in 181 eligible responses. Study 3 used the same two *ad hoc* items as in Studies 1 and 2 to assess AI accuracy and fairness in all four conditions (see Appendix B).

Independent-samples *t*-tests confirmed the effectiveness of all manipulations. In Scenario 1, participants in the 90%-accuracy fairness-oriented condition rated the AI as significantly less accurate ($M = 4.67$, $SD = 1.13$) and fairer ($M = 5.82$, $SD = 1.76$) than those in the accuracy-oriented condition (accuracy $M = 6.29$, $SD = 1.10$, $t_{(85.35)} = 6.35$, $p < .001$, $d = 1.35$; fairness

$M=2.02$, $SD=1.44$, $t_{(84.60)}=11.21$, $p<.001$, $d=2.39$). Similarly, in Scenario 2, the 60%-accuracy fairness-oriented condition was rated as significantly less accurate ($M=3.47$, $SD=1.40$) and fairer ($M=5.92$, $SD=1.51$) than the accuracy-oriented condition (accuracy: $M=6.19$, $SD=1.22$, $t_{(88.99)}=9.92$, $p<.001$, $d=2.09$; fairness: $M=2.14$, $SD=1.51$, $t_{(86.96)}=11.90$, $p<.001$, $d=2.53$).

We found no significant differences between the accuracy-oriented algorithm conditions across Scenarios 1 and 2 for either accuracy ($t_{(82.66)}=0.40$, $p=.69$, $d=0.09$) or fairness ($t_{(83.86)}=0.38$, $p=.70$, $d=-.08$). Similarly, participants in the 90%-accuracy condition rated the AI as significantly more accurate ($M=4.67$, $SD=1.31$) than those in the 60%-accuracy condition ($M=3.47$, $SD=1.40$, $t_{(91.96)}=4.28$, $p<.001$, $d=0.89$). Furthermore, we found no significant difference in fairness between the 90%- and 60%-accuracy fairness-oriented conditions ($t_{(87.11)}=0.29$, $p=.78$, $d=-0.06$). We thus successfully manipulated the accuracy of the fairness-oriented algorithm across scenarios while holding fairness constant.

Measurement model. The hypothesized five-factor model fit the data well ($\chi^2_{(340)}=770.43$, *TLI*=.94, *CFI*=.93, *SRMR*=.06), significantly better than a one-factor model ($\Delta\chi^2/\Delta df=3002.57/10$, $p<.001$; $\chi^2_{(350)=}3773.00$, *TLI*=.51, *CFI*=.47, *SRMR*=.18), supporting the convergent and discriminant validity of the measures. Tables C3 and C4 in Appendix C provide the hierarchical models.

Hypothesis testing. Results from independent-samples *t*-tests are consistent with those of Studies 1 and 2 and support H1a and H1b. In both scenarios, participants reported significantly higher other-condemning moral emotions in the accuracy-oriented condition (Scenario 1: $M=2.27$, $SD=1.45$; Scenario 2: $M=2.33$, $SD=1.50$) than in the fairness-oriented condition (Scenario 1: $M=1.43$, $SD=0.96$, $t_{(115.25)}=3.95$, $p<.001$, $d=0.69$; Scenario 2: $M=1.75$, $SD=1.10$, $t_{(113.36)}=2.54$, $p<.05$, $d=0.45$).

Conversely, participants reported higher other-praising moral emotions in the fairness-oriented condition (Scenario 1: $M=2.73$, $SD=1.7$; Scenario 2: $M=2.23$, $SD=1.48$) than in the accuracy-oriented condition (Scenario 1: $M=1.74$, $SD=1.11$, $t_{(106.90)}=3.91$, $p<.001$, $d=0.69$; Scenario 2: $M=1.83$, $SD=1.12$, $t_{(125.09)}=1.81$, $p<.10$, $d=0.45$). The effect in Scenario 2 is weaker, as it reaches significance only at the 10% level ($p=.07$). Figure 3 shows the main effect means.

The mediation analysis results are in line with the pattern observed in Studies 1 and 2, but only partially support H2a and H2b. Participants reported significantly greater negative word-of-mouth via other-condemning moral

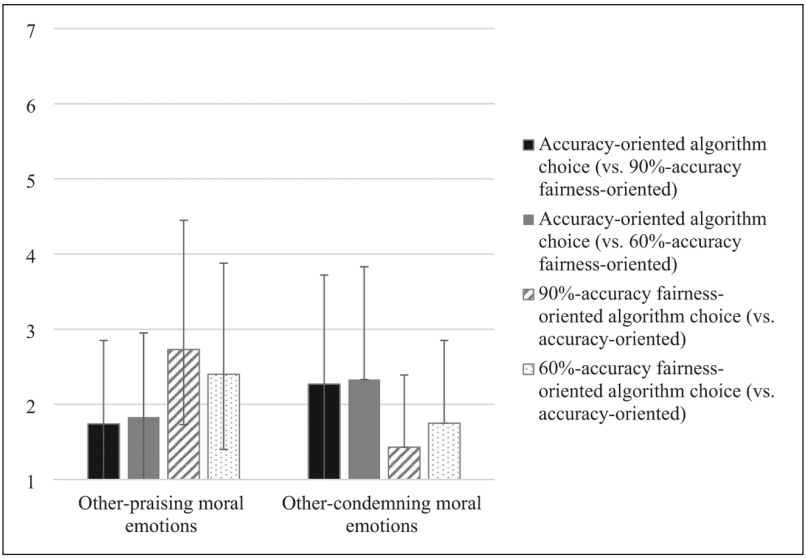

Figure 3. Study 3: Direct Effects of Algorithm Choice on Moral Emotions.

emotions in the accuracy-oriented condition, both in Scenario 1 (indirect effect = 0.60, *SE* = 0.17, 95% CI [0.27, 0.92]) and in Scenario 2 (indirect effect = 0.42, *SE* = 0.17, 95% CI [0.12, 0.77]). However, participants reported greater positive word-of-mouth via other-praising moral emotions in Scenario 1 (indirect effect = 0.38, *SE* = 0.13, 95% CI [0.15, 0.69]) but not in Scenario 2 (indirect effect = 0.14, *SE* = 0.07, 95% CI [−0.02, 0.28]). In contrast to the direct effect on other-praising, the indirect effect in Scenario 2 (which compared the accuracy-oriented condition to the 60%-accuracy fairness-oriented condition) remains nonsignificant at the 10% level (90% CI [−0.0005, 0.25]). Figure 3 shows the main effect means. Findings from Study 3 suggest that while fairness-oriented algorithmic choices elicit positive moral-emotional responses aligned with fairness-related norms, these responses are contingent on the algorithm's level of accuracy.

Discussion

Interpretation of Findings

Our research examines how individuals respond to organizations' AI orientation choices regarding accuracy versus fairness, thereby addressing the limited research attention devoted to how observers perceive algorithmic

decision-making in organizations and society (Lee, 2018; Yan et al., 2024). Drawing on deonance theory, we conceptualize organizations' AI orientation choices as value-laden decisions with moral implications. Consistent with our predictions, the results provide empirical evidence for arguments central to current debates on the ethics of AI (Bankins & Formosa, 2021; Dignum, 2019; Martin, 2019a), demonstrating that individuals do not perceive such choices as morally neutral, thereby extending emerging work suggesting that algorithmic HRM decisions motivated by organizational efficiency may be viewed as conflicting with principles of respectful employee treatment (Yan et al., 2024). More specifically, across our studies, our findings establish that organizational choices between distinct algorithmic orientations—accuracy versus fairness—elicit other-condemning and other-praising moral emotions, respectively leading to expressions of disapproval or approval toward the organization, consistent with prior work on moral protest and endorsement (Antonetti & Maklan, 2016; Hericher et al., 2023).

Importantly, results from Study 3 refine these insights by showing that the positive moral-emotional and behavioral responses observed in Studies 1 and 2 held when the fairness-oriented algorithm achieved high accuracy (90%), but not when its accuracy was substantially lower (60%). Conversely, the negative responses associated with prioritizing accuracy over fairness proved robust regardless of the fairness-oriented algorithm's level of accuracy. Consistent with deonance theory, these results suggest that people morally value fairness-oriented choices as long as they do not entail an excessive compromise on performance. That is, fairness and accuracy may be co-constitutive of perceived morality, as previous research emphasizing the consistency of algorithmic rules has alluded to (Bonezzi & Ostinelli, 2021; Kern et al., 2022; Wesche et al., 2024). When accuracy decreases substantially, the same fairness-oriented decision no longer elicits other-praising emotions strong enough to motivate supportive behavior. Even though the indirect effect was not statistically significant, Study 3 indicates that when organizations adopt fairness-oriented algorithms with substantially reduced predictive accuracy, the moral significance of such choices diminishes. Overall, these findings advance understanding of how moral approval for fairness-oriented algorithms is shaped, showing that people weigh both fairness and effectiveness in their moral appraisals.

Altogether, the three studies therefore affirm that algorithmic orientation is not a mere technical choice (Martin, 2019a; Selbst et al., 2019), and people do not construe the accuracy–fairness trade-off in binary terms. People appear willing to tolerate some loss of algorithmic accuracy when it reflects fair treatment in HRM, but their willingness to praise the organization for it diminishes when the loss becomes too substantial. These findings indicate that moral appraisals of fairness-oriented algorithms may rest on a threshold

of acceptable accuracy, below which fairness no longer offsets the loss in performance. In the context of our research question, these results show how moral emotions shape approval or disapproval of organizations' algorithmic choices in HRM decision-making.

Theoretical Contributions

Our study advances understanding of how individuals morally evaluate organizations' AI orientation choices in HRM and makes three contributions to theory and research. First, we contribute to the management literature on algorithmic decision-making from a psychological perspective by extending deonance theory to the context of algorithmic HRM, revealing how value-laden algorithms may affirm or undermine fairness-related norms and thereby elicit moral reactions toward organizations that adopt them. Previous management research has mainly compared algorithmic and human decision-making, but has rarely examined how people respond to different algorithmic orientations (Bansak & Paulson, 2024). By conceptualizing an organization's choice of algorithmic orientation as a normatively charged context, our study extends deonance theory to people's moral reactions to organizational decisions involving algorithms that affect others. Our findings support this theoretical argument by indicating that organizational efforts to prioritize fairness through algorithmic choices, even at some cost to accuracy, foster moral approval from third-party observers. In doing so, we highlight how algorithmic orientation constitutes a moral context through which organizations are perceived as adhering to or deviating from fairness-related norms.

Second, we advance research on algorithmic fairness by providing insight into the accuracy–fairness trade-off (Kleinberg et al., 2016; Zhao & Gordon, 2022), demonstrating that the choice between accuracy and fairness constitutes a moral tension. While prior research has established this trade-off as a technical property of algorithms (Menon & Williamson, 2018), we move beyond this technical focus by highlighting that individuals morally balance fairness with acceptable levels of accuracy loss. This indicates that accuracy, and not only fairness, matters for moral appraisal, revealing that the trade-off is socially and morally co-constituted. The moral significance of fairness-oriented choices thus depends on their level of accuracy, as they elicit moral approval only when the associated accuracy loss is not perceived as excessive. We thereby refine the understanding of the accuracy–fairness trade-off by showing that it bears moral relevance (Martin, 2019a; Selbst et al., 2019), and moral approval for fairness-oriented algorithms is contingent on the magnitude of the accuracy loss observers deem acceptable.

Third, we contribute to linking deonance theory with research on how moral perceptions of algorithmic practices shape third-party behavioral

intentions toward the organization (Lee, 2018; Yan et al., 2024). Specifically, we extend understanding of the organizational implications of algorithmic decision-making by showing that moral appraisals of algorithmic choices translate into approval or disapproval toward the organization. This is supported by our findings, which demonstrate that other-condemning and other-praising moral emotions elicited by an organization's choice of algorithmic orientation are associated with negative or positive word-of-mouth, respectively. In doing so, we highlight that these choices embody values that shape external moral appraisals of organizations, thereby subjecting them to moral endorsement or sanctioning.

Conclusion

This research demonstrates that people, as third-party observers, do not interpret organizations' algorithmic choices as morally neutral. Across three vignette-based experiments in hiring and dismissal contexts, we found that an accurate but gender-biased algorithm elicits other-condemning moral emotions and behaviors, whereas a less accurate but unbiased algorithm elicits other-praising moral emotions and behaviors. However, people's moral approval of fairness-oriented algorithmic choices also depends on algorithmic performance, revealing a tension between fairness and accuracy goals. Organizations should therefore approach algorithmic governance responsibly to find the right trade-off between accuracy and fairness, thereby fostering both effectiveness and equality in significant employment decisions, which ultimately reflect how they balance technological progress with moral responsibility.

Managerial Implications

Our findings have distinct managerial and practical implications, depending on whether observers are internal or external to the organization. For internal observers such as employees, organizations that choose a fairness-oriented algorithm can foster moral emotions toward the organization, thereby promoting constructive behaviors. When individuals perceive that fairness is prioritized in HRM decision-making processes, they are more likely to exhibit behaviors such as cooperation, trust, and commitment to organizational goals (Armstrong et al., 2010). These positive social dynamics can enhance team cohesion, productivity, and overall organizational effectiveness. In the context of dismissal, reducing the risk of discrimination against employees by using fairness-oriented algorithms could also help foster a stronger sense of fairness among surviving employees, limiting their negative reactions toward the organization (Skarlicki et al., 1998).

For external observers such as job applicants, customers, and the general public, the implications of algorithmic orientation are more closely related to the organization's perceived ethicality and legitimacy. When organizations prioritize fairness and inclusivity in their algorithmic design choices, they are more likely to be seen as ethical, which can earn them the approval of external stakeholders (Heath et al., 2023). Moreover, such choices can help shield them from reputational threats arising from potential discrimination scandals that standard accuracy-oriented algorithms may trigger (Hunkenschroer & Luetge, 2022). By contrast, organizations that maximize efficiency at the expense of fairness run a greater risk of public backlash, as illustrated by Amazon's experience with its AI recruitment tool, which triggered widespread moral disapproval and sanctioning (Dastin, 2022).

Limitations and Future Research

Although our research employed vignette-based experiments to investigate the impact of algorithmic orientation choices on observers' moral-emotional responses, it is important to acknowledge the inherent constraints of this methodology. Vignette-based experiments have proven useful in the study of moral issues in causal designs (Aguinis & Bradley, 2014), but the use of specific vignettes may limit the generalizability of our findings to particular cases of algorithm use in hiring and dismissal contexts. Future research could examine how accuracy- versus fairness-oriented algorithms shape moral reactions in other managerial decisions and test whether our findings generalize beyond hiring and dismissal contexts.

Finally, our research focused on observers rather than direct recipients of algorithmic decisions. Although examining observers' moral responses to algorithmic decision-making provided valuable insights, their perspectives may not fully capture the nuanced experiences and moral reactions of specific stakeholders directly affected by organizational choices. Future research could address this limitation by studying specific stakeholder groups who experience algorithmic decisions firsthand, thereby enriching our understanding of the moral implications of algorithm use from the perspective of those most directly affected by algorithmic decisions. For instance, examining direct recipients' responses to accuracy- versus fairness-oriented algorithms in hiring and dismissal situations could deepen understanding of how such technologies satisfy or undermine employees' instrumental and moral needs (Cropanzano et al., 2001).

Acknowledgment

The authors thank the participants of the 2023 Business & Society Research Seminar, 2023 EBEN research conference, and 2024 PRME chapter France-Benelux meeting for their insightful comments on earlier versions of the work.

ORCID iDs

Flore Vancompernolle Vromman https://orcid.org/0000-0002-4756-8467
Corentin Hericher https://orcid.org/0000-0002-1193-842X
Corentin Vande Kerckhove https://orcid.org/0000-0003-3297-0151
Nicolas Raineri https://orcid.org/0000-0002-0336-4550

Ethical Considerations

This article does not contain any studies with animals performed by any of the authors. All procedures performed in studies involving human participants were in accordance with the ethical standards of the first author's research institute and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Participants were anonymous and informed about the use of the information they provided on a voluntary basis.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the authors' own research budget.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

Data and scripts are publicly available online in, respectively, a Dataverse: https://doi.org/10.14428/DVN/ZGD5DR, and a GitHub repository: https://github.com/cvande-kerckh/flocos.git.

Note

1. Running correlation analyses with the non-filtered and filtered samples yielded a similar pattern of results.

References

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Antonetti, P., & Maklan, S. (2016). An extended model of moral outrage at corporate social irresponsibility. *Journal of Business Ethics*, *135*(3), 429–444. https://doi.org/10.1007/s10551-014-2487-y

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Armstrong, C., Flood, P. C., Guthrie, J. P., Liu, W., MacCurtain, S., & Mkamwa, T. (2010). The impact of diversity and equality management on firm performance: Beyond high performance work systems. *Human Resource Management*, *49*(6), 977–998. https://doi.org/10.1002/hrm.20391

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology*, *6*(3), 128–138. https://doi.org/10.1027/1614-2241/a000014

Bankins, S., & Formosa, P. (2021). Ethical AI at work: The social contract for artificial intelligence and its implications for the workplace psychological contract. In M. Coetzee & A. Deas (Eds.), *Redefining the psychological contract in the digital era* (pp. 55–72). Springer International Publishing. https://doi.org/10.1007/978-3-030-63864-1_4

Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2024). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, *45*(2), 159–182. https://doi.org/10.1002/job.2735

Bansak, K., & Paulson, E. (2024). Public attitudes on performance for algorithmic and human decision-makers. *PNAS Nexus*, *3*(12), 1–13. https://doi.org/10.1093/pnasnexus/pgae520

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning*. The MIT Press.

Becker, T. E., Atinc, G., Breaugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2016). Statistical control in correlational studies: 10 essential recommendations for organizational researchers. *Journal of Organizational Behavior*, *37*(2), 157–167. https://doi.org/10.1002/job.2053

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied*, *27*(2), 447–459. https://doi.org/10.1037/xap0000294

Breidbach, C. F. (2024). Responsible algorithmic decision-making. *Organizational Dynamics*, *53*, Article 101031. https://doi.org/10.1016/j.orgdyn.2024.101031

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-07939-1

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Chhillar, D., & Aguilera, R. V. (2022). An eye for artificial intelligence: Insights into the governance of artificial intelligence and vision for future research. *Business & Society*, *61*(5), 1197–1241. https://doi.org/10.1177/00076503221080959

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425–445. https://doi.org/10.1037/0021-9010.86.3.425

Colquitt, J. A., & Zipay, K. P. (2015). Justice, fairness, and employee reactions. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 75–99. https://doi.org/10.1146/annurev-orgpsych-032414-111457

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2024). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, *24*(1), 1–117.

Correani, A., De Massis, A., Frattini, F., Petruzzelli, A. M., & Natalicchio, A. (2020). Implementing a digital strategy: Learning from the experience of three digital transformation projects. *California Management Review*, *62*(4), 37–56. https://doi.org/10.1177/0008125620934864

Cropanzano, R. S., Byrne, Z. S., Bobocel, D. R., & Rupp, D. E. (2001). Moral virtues, fairness heuristics, social entities, and other denizens of organizational justice. *Journal of Vocational Behavior*, *58*(2), 164–209. https://doi.org/10.1006/jvbe.2001.1791

da Motta Veiga, S. P., Figueroa-Armijos, M., & Clark, B. B. (2023). Ethical perceptions of AI in hiring and organizational trust: The role of performance expectancy and social influence. *Journal of Business Ethics*, *186*(1), 179–197. https://doi.org/10.1007/s10551-022-05166-2

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In K. Martin (Ed.), *Ethics of data and analytics* (pp. 269–299). Auerbach Publications.

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

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6

Efendić, E., van de Calseyde, P. P. F. M., & Evans, A. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, *157*, 103–114. https://doi.org/10.1016/j.obhdp.2020.01.008

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery*, *36*(6), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

Folger, R. G. (2001). Fairness as deonance. In S. Gilliland, D. Steiner, & D. Skarlicki (Eds.), *Theoretical and cultural perspectives on organizational justice* (pp. 3–33). Information Age.

Folger, R. G., Cropanzano, R., & Goldman, B. (2005). What is the relationship between justice and morality? In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 215–245). Psychology Press.

Folger, R. G., & Glerum, D. R. (2015). Justice and deonance: "You ought to be fair." In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford handbook of justice in the workplace* (pp. 331–350). Oxford University Press.

Folger, R. G., & Shukla, J. (2019). A fairness theory update. In E. A. Lind (Ed.) *Social psychology and justice* (pp. 110–133). Routledge. https://doi.org/10.4324/9781003002291-6

Ford, M. T., Wang, Y., Jin, J., & Eisenberger, R. (2018). Chronic and episodic anger and gratitude towad the organization: Relationships with organizational and supervisor supportiveness and extrarole behavior. *Journal of Occupational Health Psychology*, *23*(2), 175–187. https://doi.org/10.1037/ocp0000075

Green, B., & Hu, L. (2018, July 14–15). *The myth in the methodology: Towards a recontextualization of fairness in machine learning* [Conference session]. Machine Learning: The Debates Workshop, 895, Stockholm, Sweden.

Greenbaum, R., Bonner, J., Gray, T., & Mawritz, M. (2020). Moral emotions: A review and research agenda for management scholarship. *Journal of Organizational Behavior*, *41*(2), 95–114. https://doi.org/10.1002/job.2367

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford University Press.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-21606-5

Heath, A. J., Carlsson, M., & Agerström, J. (2023). What adds to job ads? The impact of equality and diversity information on organizational attraction in minority and majority ethnic groups. *Journal of Occupational and Organizational Psychology*, *96*(4), 872–896. https://doi.org/10.1111/joop.12454

Hericher, C., & Bridoux, F. (2023). Employees' emotional and behavioral reactions to corporate social irresponsibility. *Journal of Management*, *49*(5), 1533–1569. https://doi.org/10.1177/01492063221100178

Hericher, C., Bridoux, F., & Raineri, N. (2023). I feel morally elevated by my organization's CSR, so I contribute to it. *Journal of Business Research*, *169*, Article 114282. https://doi.org/10.1016/j.jbusres.2023.114282

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, *178*(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys*, *55*(2), 1–38. https://doi.org/10.1145/3491209

Kern, C., Gerdon, F., Bach, R. L., Keusch, F., & Kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, *3*(10), Article 100591. https://doi.org/10.1016/j.patter.2022.100591

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

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores* (arXiv:1609.05807). arXiv. http://arxiv.org/abs/1609.05807

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, *12*(3), e1452. https://doi.org/10.1002/widm.1452

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

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, *64*, 19–40. https://doi.org/10.1016/j.jom.2018.10.003

Marabelli, M., Newell, S., & Handunge, V. (2021). The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges. *The Journal of Strategic Information Systems*, *30*(3), Article 101683. https://doi.org/10.1016/j.jsis.2021.101683

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Martin, K. (2019a). Designing ethical algorithms. *MIS Quarterly Executive*, *18*(2), 129–142. https://doi.org/10.17705/2msqe.00012

Martin, K. (2019b). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, *160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

McGuire, J., & De Cremer, D. (2023). Algorithms, leadership, and morality: Why a mere human effect drives the preference for human over algorithmic leadership. *AI and Ethics*, *3*(2), 601–618. https://doi.org/10.1007/s43681-022-00192-2

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Meijer, A., Lorenz, L., & Wessels, M. (2021). Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review*, *81*(5), 837–846. https://doi.org/10.1111/puar.13391

Menon, A. K., & Williamson, R. C. (2018, February 23–24). *The cost of fairness in binary classification* [Conference session]. 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, PMLR 81, (pp. 107–118).

Mitchell, M. S., Vogel, R. M., & Folger, R. (2015). Third parties' reactions to the abusive supervision of coworkers. *Journal of Applied Psychology*, *100*(4), 1040–1055. https://doi.org/10.1037/apl0000002

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Moser, C., den Hond, F., & Lindebaum, D. (2022). Morality in the age of artificially intelligent algorithms. *Academy of Management Learning & Education*, *21*(1), 139–155. https://doi.org/10.5465/amle.2020.0287

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Nussberger, A.-M., Luo, L., Celis, L. E., & Crockett, M. J. (2022). Public attitudes value interpretability but prioritize accuracy in artificial intelligence. *Nature Communications*, *13*(1), 5821. https://doi.org/10.1038/s41467-022-33417-3

Nyberg, A. J., Schleicher, D. J., Bell, B. S., Boon, C., Cappelli, P., Collings, D. G., Dalle Molle, J. E., Feuerriegel, S., Gerhart, B., Jeong, Y., Korsgaard, M. A., Minbaeva, D., Ployhart, R. E., Tambe, P., Weller, I., Wright, P. M., & Yakubovich, V. (2025). A brave new world of human resources research: Navigating perils and identifying grand challenges of the GenAI revolution. *Journal of Management*, *51*(6), 2677–2718. https://doi.org/10.1177/01492063251325188

Ötting, S. K., & Maier, G. W. (2018). The importance of procedural justice in human–machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior*, *89*, 27–39. https://doi.org/10.1016/j.chb.2018.07.022

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pessach, D., & Shmueli, E. (2020). *Algorithmic fairness* (arXiv:2001.09784). arXiv. http://arxiv.org/abs/2001.09784

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly*, *30*(1), 11–33. https://doi.org/10.1016/j.leaqua.2018.11.002

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research*. https://cran.r-project.org/web/packages/psych/citation.html

Romani, S., & Grappi, S. (2014). How companies' good deeds encourage consumers to adopt pro-social behavior. *European Journal of Marketing*, *48*(5/6), 943–963. https://doi.org/10.1108/EJM-06-2012-0364

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574–586. https://doi.org/10.1037//0022-3514.76.4.574

Rupp, D. E., Ganapathi, J., Aguilera, R. V., & Williams, C. A. (2006). Employee reactions to corporate social responsibility: An organizational justice framework. *Journal of Organizational Behavior*, *27*(4), 537–543. https://doi.org/10.1002/job.380

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed., global edition). Pearson.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January 29–31). *Fairness and abstraction in sociotechnical systems* [Conference session]. 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, United States (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Skarlicki, D. P., Ellard, J. H., & Kelln, B. R. C. (1998). Third-party perceptions of a layoff: Procedural, derogation, and retributive aspects of justice. *Journal of Applied Psychology*, *83*(1), 119–127. https://doi.org/10.1037/0021-9010.83.1.119

van den Broek, E., Sergeeva, A., & Huysman, M. (2025). Is there fairness in AI? *Journal of Management Studies*. Advance online publication. https://doi.org/10.1111/joms.13276

Vianello, M., Galliani, E. M., & Haidt, J. (2010). Elevation at work: The effects of leaders' moral excellence. *The Journal of Positive Psychology*, *5*(5), 390–411. https://doi.org/10.1080/17439760.2010.516764

Wesche, J. S., Hennig, F., Kollhed, C. S., Quade, J., Kluge, S., & Sonderegger, A. (2024). People's reactions to decisions by human vs. algorithmic decision-makers: The role of explanations and type of selection tests. *European Journal of Work and Organizational Psychology*, *33*(2), 146–157. https://doi.org/10.1080/1359432X.2022.2132940

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454–464. https://doi.org/10.1177/1948550617703168

Xie, C., Bagozzi, R. P., & Grønhaug, K. (2015). The role of moral emotions and individual differences in consumer responses to corporate green and non-green actions. *Journal of the Academy of Marketing Science*, *43*(3), 333–356. https://doi.org/10.1007/s11747-014-0394-5

Yan, C., Chen, Q., Zhou, X., Dai, X., & Yang, Z. (2024). When the automated fire backfires: The adoption of algorithm-based HR decision-making could induce consumer's unfavorable ethicality inferences of the company. *Journal of Business Ethics*, *190*(4), 841–859. https://doi.org/10.1007/s10551-023-05351-x

Zhao, H., & Gordon, G. J. (2022). Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, *23*(57), 1–26.

Author Biographies

Flore Vancompernolle Vromman is a Research and Teaching Assistant at UCLouvain, affiliated with the Louvain Research Institute in Management and Organizations (LouRIM) at the Louvain School of Management (LSM). Her research focuses on responsible AI, algorithmic fairness, human responses to algorithmic decision-making, and explainable AI. Her work has appeared in the proceedings of ACM conferences and in the journal *JMIR Mental Health*.

Corentin Hericher (PhD) is an Assistant Professor at UCLouvain, affiliated with the Louvain Research Institute in Management and Organizations (LouRIM) at the Louvain School of Management (LSM). His research interests focus on human behaviors in reaction to how a firm treats its stakeholders. His work has been published in academic journals such as *Business Ethics: The Environment and Responsibility*, *Journal of Business Research*, and *Journal of Management*.

Corentin Vande Kerckhove (PhD) is an Assistant Professor at UCLouvain, affiliated with the Louvain Research Institute in Management and Organizations (LouRIM) at the Louvain School of Management (LSM). His research interests focus on responsible AI, opinion modeling, and algorithmic decision-making. His articles have appeared in journals such as *ACM Transactions on Recommender Systems*, *PLOS ONE*, and *Political Analysis*.

Nicolas Raineri is an Associate Professor of Organizational Behavior at ICN Business School, and a member of the Centre Europeen de Recherche en Economie Financiere et en Gestion des Entreprises (CEREFIGE) at the University of Lorraine, Nancy, France. His current research focuses on the psychological foundations of corporate social responsibility. He is Editor-in-Chief of RIPCO, a French academic journal in organizational behavior and management, and his work has appeared in *Journal of Management Studies*, *Journal of Business Ethics*, and *Journal of Business Research*, among others.

Appendix A: Experimental Vignettes

Study 1

Introduction. FloCos is a company that needs to recruit new employees to sustain its business. It received numerous CVs from potential candidates interested by its current openings. Among the candidates, half are male and half are female.

FloCos decides to rely on artificial intelligence to filter the CVs. For each CV, the artificial intelligence suggests a decision "hire" or "do not hire."

To calibrate this artificial intelligence, FloCos has to choose between two possibilities, **option A** and **option B**.

- With **option A**, the artificial intelligence is very accurate at detecting good CVs. It reduces the risk that FloCos hires candidates with a mismatching set of skills. However, at equal skills, male candidates have more chances to be hired than other candidates.
- With **option B**, the artificial intelligence is less accurate at detecting good CVs than in option A. It increases the risk that FloCos hires candidates with a mismatching set of skills. However, at equal skills, male and female candidates have equal chances to be hired.

Accuracy-oriented algorithm choice condition. FloCos chooses **option A**.

Fairness-oriented algorithm choice condition. FloCos chooses **option B**.

Study 2

Introduction. FloCos is a company that needs to fire employees to sustain its business. It collected performance reviews for all its employees over the past years. Among the employees, half are male and half are female.

FloCos decides to rely on artificial intelligence to analyze the performance reviews. For each employee, the artificial intelligence suggests a decision "fire" or "do not fire."

To calibrate this artificial intelligence, FloCos has to choose between two possibilities, **option A** and **option B**.

- With **option A**, the artificial intelligence is very accurate at detecting lower performing employees. However, at equal performance reviews, male employees have more chances to keep their job than other employees.
- With **option B**, the artificial intelligence is less accurate at detecting lower performing employees than in option A. However, at equal performance reviews, male and female employees have equal chances to keep their job.

Accuracy-oriented algorithm choice condition. FloCos chooses **option A**.

Fairness-oriented algorithm choice condition. FloCos chooses **option B**.

Study 3

Introduction (accuracy-oriented vs. fairness-oriented algorithms with 90% accuracy). FloCos is a company that needs to recruit 100 new employees to

sustain its business. It received numerous CVs from potential candidates interested by its current openings. Among the candidates, half are male and half are female.

FloCos decides to rely on artificial intelligence to filter the CVs. For each CV, the artificial intelligence suggests a decision "hire" or "do not hire."

To calibrate this artificial intelligence, FloCos has to choose between two possibilities, **option A** and **option B**.

- With **option A**, the artificial intelligence is always accurate at detecting good CVs: the 100 candidates hired have the required set of skills. However, at equal skills, male candidates have more chances to be hired than other candidates.
- With **option B**, the artificial intelligence is less accurate at detecting good CVs: of the 100 candidates hired, 90 have the required set of skills. However, at equal skills, all candidates have equal chances to be hired.

Accuracy-oriented algorithm choice condition. FloCos chooses **option A**.

Fairness-oriented algorithm choice condition. FloCos chooses **option A**.

Introduction (accuracy-oriented vs. fairness-oriented algorithms with 60% accuracy). FloCos is a company that needs to recruit 100 new employees to sustain its business. It received numerous CVs from potential candidates interested by its current openings. Among the candidates, half are male and half are female.

FloCos decides to rely on artificial intelligence to filter the CVs. For each CV, the artificial intelligence suggests a decision "hire" or "do not hire."

To calibrate this artificial intelligence, FloCos has to choose between two possibilities, **option A** and **option B**.

- With **option A**, the artificial intelligence is always accurate at detecting good CVs: the 100 candidates hired have the required set of skills. However, at equal skills, male candidates have more chances to be hired than other candidates.
- With **option B**, the artificial intelligence is less accurate at detecting good CVs: of the 100 candidates hired, 60 have the required set of skills. However, at equal skills, all candidates have equal chances to be hired.

Accuracy-oriented algorithm choice condition. FloCos chooses **option A**.

Fairness-oriented algorithm choice condition. FloCos chooses **option B**.

Appendix B: Scales for Studies 1 to 3

Pretest Studies 1 and 3

Manipulation items. The option that FloCos chose: (1 = Not at all; 7 = Very much)

1) Is accurate at detecting good CVs.
2) Gives equal chances to male and female candidates to be hired.

Pretest Study 2

Manipulation items. The option that FloCos chose: (1 = Not at all; 7 = Very much)

1) Is accurate at detecting lower performing employees.
2) Gives equal chances to male and female employees to keep their job.

Items for Study 1, Study 2, and Study 3

Emotions. Thinking about FloCos' choice, I feel (1 = Not at all; 7 = Very much):
Other-praising moral emotions (Romani & Grappi, 2014)

1) Touched.
2) Inspired.
3) Moved.
4) Thankful.
5) Grateful.
6) Appreciative.

Other-condemning moral emotions (Xie et al., 2015)

1) Angry.
2) Mad.
3) Very annoyed.
4) Contemptuous.
5) Scornful.
6) Disdainful.
7) Disgust.
8) Feelings of distaste.
9) Feelings of revulsion.

Filler items

1) Indifferent.
2) Like I don't care.

3) Nothing at all.
4) Not interested.
5) Remorseful.
6) Guilty.
7) Regretful.

Behaviors. Thinking about FloCos' choice. . . (1 = Strongly disagree; 7 = Strongly agree)

Negative word-of-mouth (Hericher & Bridoux, 2023)

1) I would warn my friends and relatives not to work for or buy from FloCos.
2) I would complain to my friends and relatives about FloCos' choice.
3) I would say negative things about FloCos to other people.
4) I would speak negatively about FloCos to others around me.

Positive word-of-mouth (Hericher & Bridoux, 2023)

1) If I bought products from FloCos, I would make sure that others know it.
2) I would recommend FloCos to others.
3) I would say positive things about FloCos to other people.
4) I would speak positively about FloCos to others around me.

Control variable. *Moral identity* (Aquino & Reed, 2002)

Listed below are some characteristics that might describe a person: **caring**, **compassionate**, **fair**, **friendly**, **generous**, **helpful**, **hardworking**, **honest**, and **kind**.

The person with these characteristics could be you or it could be someone else. For a moment, visualize in your mind the kind of person who has these characteristics. Imagine how that person would think, feel, and act. When you have a clear image of what this person would be like, answer the following questions.

1) It would make me feel good to be a person who has these characteristics.
2) Being someone who has these characteristics is an important part of who I am.
3) I would be ashamed to be a person who had these characteristics. *(reverse coded)*
4) Having these characteristics is not really important to me. *(reverse coded)*
5) I strongly desire to have these characteristics.

Appendix C: Hierarchical Models

Table C1. Hierarchical Models for Study 1.

	Other-praising moral emotions		Other-condemning moral emotions		Positive word-of-mouth			Negative word-of-mouth		
Predictors	Model 1a	Model 2a	Model 1b	Model 2b	Model 3a	Model 4a	Model 5a	Model 3b	Model 4b	Model 5b
Moral identity	−.15 (.15)	−.16 (.14)	.03 (.19)	.04 (.20)	.04 (.13)	.04 (.14)	.13 (.11)	.12 (.20)	.13 (.21)	.07 (.13)
Fairness- vs. accuracy-oriented algorithm	—	.55* (.23)	—	—	—	.42* (.21)	−.11 (.15)	—	—	—
Accuracy- vs. fairness-oriented algorithm	—	—	—	.93** (.29)	—	—	—	—	1.03*** (.31)	.20 (.22)
Other-praising moral emotions	—	—	—	—	—	—	.55*** (.13)	—	—	−.24* (.10)
Other-condemning moral emotions	—	—	—	—	—	—	−.27*** (.06)	—	—	.77*** (.09)
R^2	.01	.08	.00	.09	.00	.04	.63	.00	.10	.65

Note. Regression coefficients are unstandardized, standard errors in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table C2. Hierarchical Models for Study 2.

	Other-praising moral emotions		Other-condemning moral emotions		Positive word-of-mouth			Negative word-of-mouth		
Predictors	Model 1a	Model 2a	Model 1b	Model 2b	Model 3a	Model 4a	Model 5a	Model 3b	Model 4b	Model 5b
Moral identity	−.05 (.16)	−.08 (.16)	.02 (.27)	.03 (.28)	−.15 (.19)	−.17 (.19)	−.10 (.11)	.06 (.23)	.07 (.24)	.03 (.23)
Fairness- vs. accuracy-oriented algorithm	—	.63** (.20)	—	—	—	.53* (.21)	−.14 (.14)	—	—	—
Accuracy-vs. fairness-oriented algorithm	—	—	—	.81* (.28)	—	—	—	—	.48 (.12)	−.31 (.28)
Other-praising moral emotions	—	—	—	—	—	—	.85*** (.19)	—	—	−.55** (.18)
Other-condemning moral emotions	—	—	—	—	—	—	−.22*** (.04)	—	—	.57*** (.07)
R^2	.00	.12	.00	.05	.01	.08	.76	.00	.02	.60

Note. Regression coefficients are unstandardized, standard errors in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table C3. Hierarchical Models for Study 3 With the 90%-Accuracy Fairness-oriented Condition.

	Other-praising moral emotions		Other-condemning moral emotions		Positive word-of-mouth			Negative word-of-mouth		
Predictors	Model 1a	Model 2a	Model 1b	Model 2b	Model 3a	Model 4a	Model 5a	Model 3b	Model 4b	Model 5b
Moral identity	.06 (.18)	.06 (.26)	.05 (.19)	−.07 (.26)	.10 (.12)	.25 (.17)	.19 (.14)	−.10 (.18)	−.43 (.25)	−.36 (.19)
Fairness- vs. accuracy-oriented algorithm	—	.89*** (.23)	—	—	—	.60** (.20)	−.05 (.16)	—	—	—
Accuracy- vs. fairness-oriented algorithm	—	—	—	.88*** (.24)	—	—	—	—	.59* (.26)	−.24 (.21)
Other-praising moral emotions	—	—	—	—	—	—	.42*** (.07)	—	—	−.30*** (.08)
Other-condemning moral emotions	—	—	—	—	—	—	−.34*** (.07)	—	—	.68*** (.08)
R^2	.00	.12	.00	.11	.00	.10	.55	.00	.07	.54

Note. Regression coefficients are unstandardized, standard errors in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table C4. Hierarchical Models for Study 3 With the 60%-Accuracy Fairness-oriented Condition.

	Other-praising moral emotions		Other-condemning moral emotions		Positive word-of-mouth			Negative word-of-mouth		
Predictors	Model 1a	Model 2a	Model 1b	Model 2b	Model 3a	Model 4a	Model 5a	Model 3b	Model 4b	Model 5b
Moral identity	.06 (.18)	−.02 (.27)	.05 (.19)	.22 (.28)	.10 (.12)	−.10 (.19)	−.03 (.14)	−.10 (.18)	.44 (.33)	.28 (.24)
Fairness- vs. accuracy-oriented algorithm	—	.34 (.20)	—	—	—	.53*** (.15)	.22* (.10)	—	—	—
Accuracy-vs. fairness-oriented algorithm	—	—	—	.62** (.24)	—	—	—	—	.94*** (.26)	.41* (.19)
Other-praising moral emotions	—	—	—	—	—	—	.40*** (.10)	—	—	−.47** (.14)
Other-condemning moral emotions	—	—	—	—	—	—	−.30*** (.06)	—	—	.66*** (.09)
R^2	.00	.03	.00	.06	.00	.10	.63	.00	.12	.62

Note. Regression coefficients are unstandardized, standard errors in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.