BMC Medical Ethics

RESEARCH

Open Access

# Negative performance feedback from algorithms or humans? effect of medical researchers' algorithm aversion on scientific misconduct

Ganli Liao[1*], Feiwen Wang[1], Wenhui Zhu[2] and Qichao Zhang[1*]

## Abstract

Institutions are increasingly employing algorithms to provide performance feedback to individuals by tracking productivity, conducting performance appraisals, and developing improvement plans, compared to traditional human managers. However, this shift has provoked considerable debate over the effectiveness and fairness of algorithmic feedback. This study investigates the effects of negative performance feedback (NPF) on the attitudes, cognition and behavior of medical researchers, comparing NPF from algorithms versus humans. Two scenario-based experimental studies were conducted with a total sample of 660 medical researchers (algorithm group: N1 = 411; human group: N2 = 249). Study 1 analyzes the differences in scientific misconduct, moral disengagement, and algorithmic attitudes between the two sources of NPF. The findings reveal that NPF from algorithms shows higher levels of moral disengagement, scientific misconduct, and negative attitudes towards algorithms compared to NPF from humans. Study 2, grounded in trait activation theory, investigates how NPF from algorithms triggers individual's egoism and algorithm aversion, potentially leading to moral disengagement and scientific misconduct. Results indicate that algorithm aversion triggers individuals' egoism, and their interaction enhances moral disengagement, which in turn leads to increased scientific misconduct among researchers. This relationship is also moderated by algorithmic transparency. The study concludes that while algorithms can streamline performance evaluations, they pose significant risks to scientific misconduct of researchers if not properly designed. These findings extend our understanding of NPF by highlighting the emotional and cognitive challenges algorithms face in decision-making processes, while also underscoring the importance of balancing technological efficiency with moral considerations to promote a healthy research environment. Moreover, managerial implications include integrating human oversight in algorithmic NPF processes and enhancing transparency and fairness to mitigate negative impacts on medical researchers' attitudes and behaviors.

**Keywords** Scientific misconduct, Negative performance feedback, Algorithm aversion, Moral disengagement, Egoism, Algorithmic transparency

*Correspondence:
Ganli Liao
glliao@bistu.edu.cn
Qichao Zhang
zhangqichao@bistu.edu.cn
Full list of author information is available at the end of the article

## Background

In the digital age, algorithms have permeated various fields, emerging as pivotal elements for organizational development. They are instrumental in enhancing service quality, minimizing costs, and optimizing the allocation of resources [129]. Building on these advancements, the application of algorithms now extends to the feedback of individual performance, reflecting their growing role in more personalized aspects of organizational operations. By leveraging big data analytics, organizations delve into biometric and online behavioral data to uncover profound insights into individual performance and emotional dynamics [7, 108]. Then, utilizing sophisticated statistical models and decision-making algorithms, they provide a holistic feedback of individuals' performance [113]. Taking the Diagnosis-Related Group (DRG) performance feedback system, commonly adopted by Chinese medical institutions, as an example, this system has demonstrated its significant effectiveness in refining medical management and enhancing service quality. Further, with the widespread adoption of advanced technologies such as artificial intelligence, big data analysis, deep learning, and neural networks, the precision and efficiency of performance feedback algorithms are experiencing substantial improvements. Compared to using algorithms for performance feedback, traditional human decision-making (such as managers, supervisors, etc.) is susceptible to factors like past experiences and emotions, leading to misleading outcomes or decision errors. Algorithms can process and analyze complex datasets, avoiding fatigue and emotional interference, thus providing a more scientifically and rational performance feedback method. For research institutions, algorithms also have been extensively applied to various aspects of scientific evaluation. For example, big data is used to comprehensively assess scientific outcomes such as the quantity and quality of researchers' academic outputs, academic influence, and innovation capacity. This includes the number of publications and citations, the number of project applications and their novelty, as well as academic collaborations and exchanges. Therefore, the algorithms in performance feedback can identify patterns, trends, and insights that humans may overlook and make decisions based on data. However, while the applications of algorithms in performance feedback has brought revolutionary progress, it also faces unprecedented challenges, especially when using algorithms to provide individuals with negative performance feedback (NPF). For instance, in 2019, a former Microsoft employee shared on social media how he/she was dismissed due to an erroneous evaluation of his/her job performance by Microsoft's performance feedback algorithm. This incident sparked widespread public attention on the transparency and fairness of algorithmic decisions-making. Similarly, in 2020, Google employees revealed that the company used algorithms to assess individual performance and promotion opportunities. There was a growing concern that such algorithms could exacerbate quantitative disparities, as they might overlook qualitative work outcomes like leadership and creativity, focusing excessively on quantifiable metrics.

Such cases are not uncommon, reflecting that NPF from algorithms often tend to be more stringent and uncontrollable. This is because NPF from algorithms are often based on big data samples and machine learning models, which may be seen as quantitative and decontextualized. This feedback may be affected by data quality, sample bias, and algorithmic limitations, leading to inaccuracies and distortions. Additionally, the algorithms lack the capacity to explain NPF, which can easily result in individuals feeling unfairly treated, as the reasons and basis for NPF are difficult for them to accept. Once procedural unfairness arises, individuals may be less tolerant of the NPF from algorithms. It is undeniable that some individuals can learn from NPF and be motivated to strive towards work objectives. However, when the sources of NPF (algorithms or humans) differ, the effects on individuals' cognition, attitudes, emotions, and behaviors can be distinctly different. As a result, some scholars proposed that individuals may tend to accept NPF from humans and reject suggestions from algorithms. This inclination is known as algorithm aversion [36]. Algorithm aversion can lead to a variety of psychological and behavioral reactions, including doubts about the validity of algorithmic feedback and a reluctance to accept their recommendations [21, 82]. For medical researchers, it may even drive them to engage in unethical behaviors like data tampering, plagiarism, or fabricating results to avoid the negative feedback of algorithms. These behaviors jeopardize the reputations and careers of medical researchers and introduce inaccuracies and instability into scientific research. Consequently, the question of whether algorithms or humans should provide NPF has raised a critical issue: how to ensure that the NPF from algorithms in the research field is not only technologically advanced but also sufficiently considers the ethical acceptability and psychological impact on medical researchers.

Previous studies have investigated the key factors influencing algorithm aversion from perspectives such as algorithmic errors, privacy concerns, and misattribution [36, 37, 41, 76, 110]. Scholars believe that individuals form basic cognition and judgments based on their experiences and environment [83, 90]. When individuals observe errors or opacity in algorithms, they not only tend to trust human decisions over algorithms but also may worry about privacy violations. Additionally,

individuals believe in their decision-making capabilities over those of algorithms,thus, when receiving NPF from algorithms, they are likely to attribute the failure to algorithmic faults rather than their own shortcomings [64]. Furthermore, unfamiliarity with algorithms can heighten algorithm aversion [82]. Medical researchers, often lacking specialized knowledge in computational algorithms, are more likely to exhibit negative reactions towards them. Therefore, a thorough exploration of the complex ethical cognition and emotional responses triggered by algorithm aversion among researchers receiving NPF is crucial for uncovering the deep psychological mechanisms and dynamic processes behind human-algorithms interactions in the digital age.

In summary, this study aims to explore the impact of NPF provided by both algorithms and humans on the behavior and psychology of medical researchers. Two scenario-based experimental studies were designed. Study 1 compares the effects of NPF from algorithms versus humans, analyzing whether there are differences in the scientific misconduct, moral disengagement, and algorithmic attitudes. Drawing on trait activation theory, Study 2 investigates how NPF from algorithms triggers feelings of algorithm aversion and how these emotions interact with an individual's egoism, potentially leading to moral disengagement and scientific misconduct. This study proposes that internal traits of individuals are activated by changes in situations, leading to an interaction that produces corresponding cognition and behaviors. This interaction follows the pathway mechanism of "Trait × Situation → Cognition → Behavior," illustrating how dynamic interacts between personal traits and situational factors influence cognition and behaviors. By analyzing the changes in individual psychological motivations when faced with NPF from algorithms and how these changes affect their moral cognition processes, this study aims to provide a deep understanding of the ethical dilemmas and challenges associated with algorithmic decision-making. Additionally, it offers theoretical and managerial suggestions on how to construct more equitable and transparent algorithmic feedback systems to foster scientific integrity.

## Theory and hypotheses

### Feedback from algorithms or humans? fundamentals of medical researchers' NPF

Individual performance feedback are integral to organizations, involving the gathering of job-related data, the assessment of performance, and the provision of guidance for improvement [6, 101]. These activities form the core of the managerial function of information, which necessitates the continuous monitoring of the workplace, including individuals, to generate, process, and disseminate relevant data [88]. As algorithm technology advances, it is increasingly utilized in data analytics to make accurate and comprehensive predictions [54], suggesting that algorithm has the potential to perform the role of managers in these information functions.

The question then arises: who should provide NPF to medical researchers, an algorithm system or a human manager? The integration of algorithms into scientific research presents significant potential for enhancing the efficiency and accuracy of performance feedback. However, this change raises important considerations about the nature and effectiveness of NPF in a research setting. Medical researchers may perceive NPF from algorithms differently than from human, due to the unique characteristics of algorithmic feedback, such as its objectivity, consistency, and lack of emotional nuance. This disparity may lead to differing perceptions of feedback quality, trust in the feedback provider, and ultimately, the researchers' reaction to the NPF. Specifically, NPF from algorithms may lead to a disconnect between the feedback and the researcher's emotional state, potentially leading to a sense of detachment or even resentment. In contrast, NPF from humans can be tailored to the medical researcher's circumstances and emotions, fostering a sense of empathy and understanding. This personalized approach can help them to feel supported and motivated to make the necessary changes in their undervalued performance. Additionally, human managers can provide context and explanation for the NPF, which can help medical researchers to understand the reasons behind the undervalued performance and how to address the issues raised. Therefore, the question of who should provide NPF is not merely a technical consideration but a complex issue that involves ethical, psychological, and organizational factors. This study aims to investigate the impact of NPF provided by algorithms versus humans on researchers' scientific misconduct, moral disengagement, and algorithmic attitudes.

Scientific misconduct within research settings encompasses behaviors that breach recognized ethical norms across various stages of scientific research, such as project application and approval, the conduct of research, and the publication of findings. These behaviors include plagiarism and data fabrication [52, 98, 138]. Studies have shown that the unreasonable evaluation and reward mechanisms lead to the excessive depletion of researchers' psychological resources during involution process [81, 86, 104]. This state of emotional exhaustion can further trigger scientific misconduct [75]. When organizations provide NPF, the pressures from peer competition, academic status, research assessments, and career advancement induce increasing uncertainty about the future for researchers, thereby reducing their sense of

self-efficacy in their research abilities [115]. Particularly when such feedback are linked to promotions and salaries, medical researchers might opt for "shortcuts" to secure their jobs or improve their reputation. Driven by the need to alleviate anxiety and safeguard their career stability, they may resort to unethical practices to achieve their expected performance targets [51]. Drawing on the aforementioned NPF's impact on scientific misconduct, we argue that different NPF providers (algorithms vs. humans) have significant differences on medical researchers' misconduct for three reasons. First, NPF from algorithms is typically based on data and lacks the emotional and empathetic aspects that humans can provide [105]. The lack of emotional engagement in NPF from algorithms can lead medical researchers to feel undervalued or misunderstood, potentially exacerbating their stress and anxiety. This emotional state may push some medical researchers to resort to unethical means to cope with pressure, such as fabricating data or plagiarizing, in order to enhance their research performance. Then, the absence of tailored guidance in NPF from algorithms can lead medical researchers to perceive a lack of targeted support and assistance [38, 44, 94], which can influence their work attitudes and behavior. In the face of pressure and challenges, they may be more inclined to engage in unethical behavior [84]. In contrast, when NPF is provided by humans, they are able to offer emotional support and personalized understanding [40, 124], which may assist medical researchers in better coping with pressure and challenges. The humanized nature of managerial feedback can enhance a sense of belonging and trust among researchers, thereby reducing the risk of scientific misconduct.

Moral disengagement refers to a set of cognitive tendencies in individuals that include redefining their behaviors to appear less harmful and minimizing their responsibility for negative outcomes [9, 94]. Research indicates that moral disengagement can take place through mechanisms such as moral justification, advantageous comparison, diffusion of responsibility, distortion of consequences, and attribution of blame. These mechanisms effectively weaken internal moral condemnation [8, 10]. As a result, moral disengagement can facilitate a range of unethical behaviors. This may include interpersonal deviance [55, 96], workplace deviance [43, 56, 91, 109], workplace incivility [27, 58], and pro-organizational unethical behaviors [11, 25]. Additionally, it can lead to broader organizational misconduct, such as corruption and illegal activities. We propose that the differences between NPF from algorithms and NPF from humans in terms of moral disengagement are attributed to distinct differences in the feedback mechanisms cognitive processing, and emotional impact. Firstly, NPF from algorithms is typically based on pre-established rules, tending to objectively and fairly evaluate an individual's behavior without being influenced by emotional factors. This mode of feedback may focus more on facts and data, with less consideration for individual moral responsibility. Consequently, when individuals receive NPF from algorithms, they may be more inclined to rationalize their behaviors or attribute responsibility to external factors, thus exhibiting a higher propensity for moral disengagement. In contrast, NPF from humans often involves more interpersonal interaction and emotional factors. Humans, in providing feedback, may consider not only work performance but also individual traits, work attitudes, and team dynamics [26, 118]. Consequently, humans' feedback may be more complex and diverse, encompassing not only the feedback of facts but also judgments of the individuals' moral qualities [103]. This kind of feedback may more directly affect their sense of morality and responsibility, making it more difficult for them to alleviate internal guilt through moral disengagement. Furthermore, there may be differences in the cognitive processing of feedback from algorithms and humans. Due to algorithms' lack of human characteristics, individuals may perceive them as an objective feedback tool rather than linking it to their own moral qualities. However, humans, as entities with emotional and social attributes, often evoke a more personalized response from individuals, influencing their self-perception and moral judgment. Thus, we propose that:

> *Hypothesis 1: There are significant differences in medical researchers' scientific misconduct and moral disengagement depending on whether the NPF provided by algorithms or humans.*

### Aversion to algorithms or humans? preferences in NPF makers

Although algorithmic decision-making has brought revolutionary improvements to societal efficiency due to its significant advantages, individual seem reluctant to embrace this emerging technology despite its powerful computational capacity and precise predictions [37]. This hesitation stems from concerns about the transparency and interpretability of algorithms, as well as worries regarding the ethical and fairness implications, and issues related to data security and privacy [32, 61, 121]. Additionally, in managerial practices such as recruitment, promotion, dismissal, and motivation, individuals tend to perceive algorithmic decisions as merely quantitative and decontextualized [80, 121]. They believe that feedback provided by algorithms fail to consider qualitative and environmental factors, which leads to incomprehensive decision-making [71]. Consequently, algorithmic

decisions are often perceived as less fair compared to similar human decisions, potentially reducing emotional commitment to the organization [94]. This phenomenon, known as algorithm aversion, reflects a cautious attitude towards the application of algorithms in decision-making [65, 76, 82]. It is widespread in organizational decision-making and reflects a cognitive bias among individuals. Previous studies indicate that algorithm aversion manifests across multiple dimensions, including cognition, emotion, and behavior. Cognitively, individuals may harbor skepticism towards the decision-making processes of algorithms [24, 61, 82]. Emotionally, they might experience feelings of distrust or resistance towards these algorithmic systems [126]. Behaviorally, they are more likely to disregard or avoid algorithmic recommendations [17, 78], favoring human decisions instead. This multi-faceted response highlights the deep-rooted challenges in integrating algorithmic solutions into the fabric of organizational practices. These challenges significantly hinder the promotion and application of algorithms. Based on these, our study suggests that when organizations provide NPF to medical researchers, they are more inclined to accept NPF from humans rather than from algorithms. Medical researchers' performance is typically measured through a range of indicators, including the quantity of research outputs, management of research projects, and dissemination of knowledge to the public. However, the quantification of these algorithmic indicators may overlook non-quantifiable aspects of medical researchers' achievements, such as the depth of research projects, the innovativeness of studies, the original contributions of researchers, and the broader impact and dissemination of research findings. If the algorithmic negative feedback systems fail to account for the non-quantifiable achievements and specific situational factors of medical researchers, their acceptance of algorithmic assessments may decline. Even though algorithms can often make more accurate decisions than humans, medical researchers might still prefer humans' feedback over algorithmic suggestions when faced with negative feedback that do not meet preset goals or performance standards. Thus, we propose that,

> *Hypothesis 2: Medical researchers tend to avoid NPF from algorithms more than that from humans in terms of cognition, emotion, and behavior.*

### Algorithm aversion: trait activation effect under NPF from algorithm

Thus far, we mainly focus on examining the disparities between NPF from algorithms and that from humans in terms of individual behavior and moral cognition, as well as the degree of aversion individuals exhibit towards these two decision-making mechanisms. However, the question arises as to whether medical researchers, upon receiving NPF from algorithms, would engage in scientific misconduct in an attempt to enhance their personal performance. Therefore, we further employ the trait activation theory to investigate the underlying mechanisms that lead to scientific misconduct in the context of NPF from algorithms.

Trait activation theory suggests that latent personality traits within individuals are activated under appropriate situations, prompting them to demonstrate specific behaviors or intentions that align with these traits. This process is outlined in the pathway "Trait × Situation → Cognition → Behavior" [122, 123]. From the perspective of traits, this study introduces the concept of egoism, which refers to an individual's psychological tendency or internal drive to gain benefits or avoid punishments. [99, 131]. Egoism can significantly shape how individuals respond to different situations, especially when their behaviors are aligned with personal gains or avoidance of negative consequences. Individuals driven by egoism are predisposed to place their personal interests above those of others or societal benefits [120]. In pursuit of safeguarding their core advantages or avoiding penalties, they may deploy a variety of strategies. This psychological trait not only guides individual responses to external situations but also dictates their reactions to NPF from algorithms, influencing the actions they undertake to counteract the detrimental effects. From a situational perspective, NPF from algorithms represent a specific context that signals deficiencies, substandard performance, or unmet expectations to employees. This context triggers egoism, as individuals facing NPF from algorithms may perceive a direct threat to their personal interests. Consequently, they are compelled to act in ways that protect or enhance these interests. This interaction between individual trait and external situation illustrates the sophisticated dynamics involved when individuals respond to algorithmic assessments, especially when such feedback challenge their professional standing and potential career progression.

Therefore, when NPF is provided by algorithms, we argue that individuals with high algorithm aversion will activate their egoism, which will lead to their scientific misconduct. Firstly, NPF from algorithm potentially indicates that individuals' performance has not met the required standards, which can induce their psychological stress. This stress is particularly pronounced in the research area, where research outcomes directly impact researchers' career advancement and academic reputation. When medical researchers receive NPF from algorithms, they may experience intense frustration and anxiety. These emotional responses, in turn, could

prompt considerations of engaging in unethical behaviors as a means to enhance their performance [128, 132]. This tendency is particularly pronounced among those who harbor skepticism and distrust towards algorithms [14, 47]. Such individuals are more likely to perceive algorithms as inaccurate, unfair, or biased. This sense of disbelief may lead them to reject the algorithms' assessment results, promoting a preference for self-serving behaviors driven by personal interests as a response to NPF. Furthermore, previous studies indicate that individual with egoism become more pronounced when faced with stress and challenges [16, 102]. They may manifest as an excessive pursuit of scientific achievements and an overemphasis on personal reputation and benefits [77]. Therefore, when medical researchers encounter the pressure stemming from algorithmic NPF, they might be more inclined to engage in unethical behaviors, such as data manipulation or plagiarism, in an effort to improve their performance outcomes. Especially for individuals with high algorithm aversion, the lack of transparency in how algorithms collect, store, and use private information, as well as unclear perceptions of their functionality and purpose, may enhance their sense of uncertainty [139]. Consequently, these individuals may view algorithmic NPF as an additional challenge and stress. Driven by competition and a desire to maximize personal gains, they may adopt scientific misconduct to maintain their reputation or advantages. Therefore, we propose that:

> *Hypothesis 3: Medical researchers with algorithm aversion who receive NPF from algorithms are egoistically motivated to engage in scientific misconduct.*

With the widespread adoption of algorithms, issues such as social isolation, technological overload at work, and job insecurity have become critical concerns in current algorithmic management practices [62]. These issues reflect significant variations in the acceptance and tolerance of algorithms among individuals. Existing research suggests that external factors can influence an individual's moral cognition [15, 127]. Thus, moral disengagement is more likely to occur under certain contextual conditions [9, 27, 94]. For instance, job insecurity, by impeding individuals with negative emotions and stress-related experiences from engaging fully in their work, can further increase moral disengagement [43, 55, 100]. High levels of moral disengagement can diminish the connection between individuals' moral decision-making and their internal moral standards, resulting in unethical behavior [18, 43, 92], aggressive behaviors [97], etc. Scientific misconduct breaches ethical principles and societal moral norms, thereby constituting unethical behavior [98]. Therefore, we propose that when algorithms provide NPF, the interaction between algorithm aversion and egoism may increase moral disengagement, thereby leading to scientific misconduct. Firstly, due to the heavy daily workload and non-scientific responsibilities occupying a substantial amount of personal time, researchers are left with relatively little time for academic research. Moral disengagement may emerge as a seemingly "easy and reasonable" way to cope with the pressure of evaluation. NPF from algorithms often induce negative emotions among medical researchers, especially when they perceive the algorithms as unreasonable or unfair. This emotional response may lead to increased aversion towards the algorithms, resulting in a tendency to ignore or question the outcomes of algorithmic evaluations. This condition lays the foundation for overlooking ethical guidelines in favor of mitigating personal dissatisfaction and perceived injustice. Therefore, individuals with high algorithm aversion may find it easier to excuse their unethical behaviors through moral disengagement [13, 61], thereby further stimulating egoism. This leads them to seek the most favorable explanations or strategies to deal with algorithms, further alleviating their ethical burden [89, 94, 121], subsequently enhancing individuals' propensity for moral disengagement. Secondly, scientific misconduct, being one of the most prevalent deviant behaviors in the academic field, is widespread within research institutions. Effectively managing and correcting these behaviors poses significant challenges. In fact egoism undermines individuals' ability to address moral dilemmas in the workplace [46]. Highly egoistic individuals tend to attribute problems to external factors, such as environmental harshness and situational pressures. They are more likely to perceive deviant behaviors as acceptable, even if impermissible [5]. Therefore, moral disengagement enables researchers to redefine their behaviors, thereby reducing adherence to ethical norms and making it easier for them to engage in misconduct. Thus, we propose that:

> *Hypothesis 4: The interaction between egoism and algorithm aversion is positively related to moral disengagement. Specifically, the higher a medical researcher's level of algorithm aversion, the stronger the positive relationship between egoism and moral disengagement.*
> *Hypothesis 5: The interaction between egoism and algorithm aversion is positively associated with moral disengagement, which in turn positively related to scientific misconduct.*

### Understanding algorithms: alleviating effect of algorithmic transparency

Because we aim to understand how to increase the value of NPF from algorithms in performance evaluation, we investigate boundary conditions that may alleviate the negative consequences of the algorithms.

According to trait activation theory, reducing scientific misconduct among researchers hinges on decreasing the extent to which algorithm aversion activates egoistic traits, thereby weakening the relationship between "Trait—Cognition—Behavior". Transparency is commonly used to describe the visibility of information and involves a series of deliberate disclosure processes. Through these processes, individuals can gain insights into relevant information, intentions, or behaviors [127]. Specifically, algorithmic transparency refers to the knowability or visibility of how an algorithmic system operates [28, 106]. It includes two core dimensions: the first is accessibility, which involves the public availability of algorithm models,the second is interpretability, which requires that the outcomes of algorithms be explained in a manner understandable to people [48]. Hill et al. [53] suggest that algorithmic transparency can influence behavioral changes by affecting individuals' psychological and cognitive processes. Therefore, this study proposes that algorithmic transparency may play a "buffering" role, helping to reduce the activation of egoism.

On one hand, scholars have examined the impact of algorithms on individual cognition from various perspectives. And it is widely acknowledged that algorithmic transparency is crucial for individuals to feel respected within algorithmic processes [35, 130]. As highlighted by the procedural fairness theory, the accuracy and transparency of decision-making information are crucial because individuals react differently to decision outcomes based on the transparency of the decision-making process [111]. Medical researchers who perceive high algorithmic transparency believe that the mechanisms of the algorithms are sufficiently clear to meet their information transparency needs [134, 142]. When they understand the underlying mechanisms of algorithmic decisions, they are more inclined to trust that these decisions are fair [48, 116]. Consequently, they are less likely to engage in moral disengagement to justify their inappropriate behavior, thereby reducing scientific misconduct.

On the other hand, algorithm-based task allocation, performance management, and reward and punishment incentives are often not fully transparent [19, 20, 107]. The opacity of algorithms is evident in their privacy, proprietary nature, and complexity [68]. Although few question the quality of the algorithms, their opacity often becomes a significant reason for ethical behaviors [61]. When organizations make decisions and oversee their managerial processes through algorithms, medical researchers who perceive low algorithmic transparency may find their work autonomy constrained [125]. They may feel compelled to accept algorithmic decisions without the ability to appeal against NPF from the algorithms, resulting in feelings of helplessness and dissatisfaction [70]. Additionally, individuals with egoism are more likely to focus on personal gains, prioritizing short-term benefits over long-term consequences. Therefore, insufficient algorithmic transparency might further promote moral disengagement, ultimately leading to scientific misconduct. Thus, we propose that,

*Hypothesis 6: Algorithmic transparency moderates the moderating effect of algorithm aversion on the relationship between egoism and moral disengagement. Specifically, higher levels of algorithmic transparency can mitigate the activating effect of algorithm aversion on moral disengagement among medical researchers with egoism.*

*Hypothesis 7: Algorithmic transparency moderates the mediating effect of moral disengagement between algorithm aversion, egoism, and scientific misconduct. Specifically, higher levels of algorithmic transparency diminish the influence of moral disengagement, thus reducing scientific misconduct among medical researchers with egoism and algorithm aversion.*

## Scenario-based experiment and projection technique

### Sample and collection

First, we employed G*Power 3.1 software to estimate the minimum sample size for our experiments [42]. This experiment used the independent samples t-test, assuming a medium effect size of $d=0.5$ and a significance level of $\alpha=0.05$. The result indicates that a minimum of 172 participants are required for this study. Then, we employed scenario-based experiments along with psychological projection techniques selecting participants through random sampling. Our participants included faculty members, graduate students, and Ph.D candidates engaged in medical research from various institutions in China. We designed a series of scientific misconduct scenarios using a situational simulation experiment approach. These materials were presented from a third-person perspective to indirectly ask questions, avoiding direct inquiries to the participants. All participants voluntarily took part in the study with informed consent and received compensation at the conclusion of the experiment. A total of 739 questionnaires were distributed for this study. After excluding 35 questionnaires that failed the attention checks, 704 questionnaires were obtained, resulting in a response rate of 95.26%. Subsequently, after removing 44 invalid questionnaires with more than 10% missing data or excessive selection of the same option, a total of 660 questionnaires were used for statistical analysis (overall response rate of 89.31%). The sample consisted of 52.6% male and 47.4% female participants. Regarding the age distribution of the participants, 27.7% were under

30 years old, 48.0% were 31–40 years old, 14.1% were 41–50 years old, 10.0% were 51–60 years old, and 0.2% were over 60 years old. Regarding marital status, 51.8% of the participants are married, 48.9% are unmarried, 1.5% are divorced, and 0.8% fall into the "other" category. In terms of educational level, 29.2% of the participants obtained a bachelor's degree, 61.5% obtained a master's degree, and 9.2% obtained a doctoral degree. The distribution of participants by years of professional experience is as follows: 16.2% have less than 5 years of experience, 47.7% have 5 to 10 years, 19.8% have 10 to 15 years, 5.3% have 15 to 20 years, 5.6% have 20 to 25 years, and 5.3% have more than 25 years of professional tenure.

Study 1 used a one-way ANOVA to analyze the differences between the algorithm group and the human group. All participants were randomly assigned to read the materials, resulting in 411 participants (N1 = 411) in the algorithm group and 249 participants (N2 = 249) in the human group. These materials were measured based on Newman et al.'s [94] study. Apart from the identity of the decision-maker (i.e., whether the NPF were provided by the algorithm "KH220" or Wang Qi's team), all other content remained unchanged. After reading the scenario materials, participants were asked to fill out questionnaires regarding their attitudes towards the decision-maker, moral disengagement, scientific misconduct and demographic variables. The scenario materials are provided in Supplementary Material 1. Furthermore, Study 2 utilized the 411 questionnaires from the algorithm group to test our hypotheses.

### Measurement

In this study, the mature scale published in TOP journals was translated and back translated in strict accordance with relevant procedures to ensure the consistency of the scale in the Chinese context. Three professors in the field of business administration and several Ph.D candidates were invited to evaluate the Chinese scale with reference to the original scale. After modifying and adjusting to certain items, it is imperative to ensure that the scale exhibits robust content validity. All scales were scored on a Likert's 5-point scale (1 = strongly disagree, 5 = strongly agree).

#### *Algorithm aversion*

This scale comprises three dimensions, including the permissibility, the liking for algorithms, and the utilization intention. The permissibility adopts the scale developed by Bigman and Gray [13], the liking for algorithms employs the scale by Jago [60], and the utilization intention use the scale from Cadario et al. [23]. It consists of a total of 6 items, such as "Is the appraisal decision made by the 'KH220' feedback algorithm appropriate?" and "Should the 'KH220' feedback algorithm be allowed to make these performance feedback decisions?" In this study, the Cronbach's alpha was 0.904, indicating good reliability.

#### *Moral disengagement*

Chen et al.'s scale [25] was used to measure the moral disengagement, which includes 3 items. Sample items like "It is okay for him to use misleading information in order to improve his performance." and "It is okay for him to withhold potentially damaging information in order to improve his performance." In this study, the Cronbach's alpha for this scale was 0.897.

#### *Egoism*

This is measured using the Machiavellianism scale developed by Dahling et al. [31], which comprises 16 items. Examples of items include: "I think it's okay to tell small lies if it keeps me competitive" and "If it helps me succeed, I believe that engaging in some unethical behaviors is acceptable." In this study, the Cronbach's alpha for this scale was 0.933.

#### *Scientific misconduct*

Adapted from a study by Zhang et al. [143], this scale consists of 12 items. Participants read four scenarios sequentially and respond to three questions per scenario to assess their acceptance, consistency and uniformity to test the likelihood of scientific misconduct. Items are "Do you accept the behavior of the researchers described in the context material?", "Do you think the researcher has engaged in such behavior in previous projects?", "Do you think researchers will engage in this behavior in other future projects?". In this study, the Cronbach's alpha for this scale was 0.759.

#### *Algorithmic transparency*

Drawing upon the research of Cadario et al. [23], algorithmic transparency is measured through participants' perceived transparency of decision-making activities conducted by algorithms or humans. Specifically, the item involves asking participants, "To what extent do you understand how the Wang Qi team/algorithm 'KH220' made the above decisions?".

#### *Control variables*

Previous studies have shown that gender, age, educational level, marital status, and years of professional experience may affect researchers' moral disengagement and scientific misconduct. Therefore, these variables are controlled in this study.

The study design was shown in Fig. 1.

## Study 1

The independent samples t-test was conducted using GraphPad Prism 9.5 to examine the differences between human and algorithm groups. Regarding scientific misconduct and moral disengagement, the t-test results in Fig. 2 indicated that the human group showed significantly lower levels of moral disengagement (M = 2.46, SD = 0.99) compared to the algorithm group (M = 3.35, SD = 1.14), with t = 10.206 ($p < 0.001$, Cohen's d = 1.091). Similarly, the human group's scientific misconduct (M = 2.73, SD = 0.54) was significantly lower than the algorithm group's (M = 3.15, SD = 0.66), with t = 8.427 ($p < 0.001$, Cohen's d = 0.618). Then, this study used multivariate analysis of variance (MANOVA) to examine the combined effects of different decision-maker (algorithms vs. humans) on moral disengagement and scientific

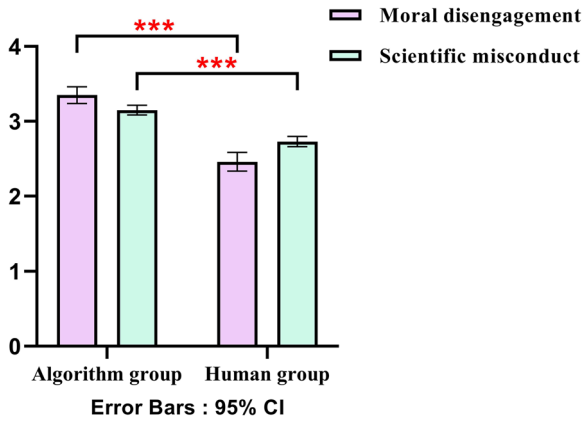


**Fig. 2** t-test results for the moral disengagement and scientific misconduct between the algorithm group and the human group (***$p < 0.001$)

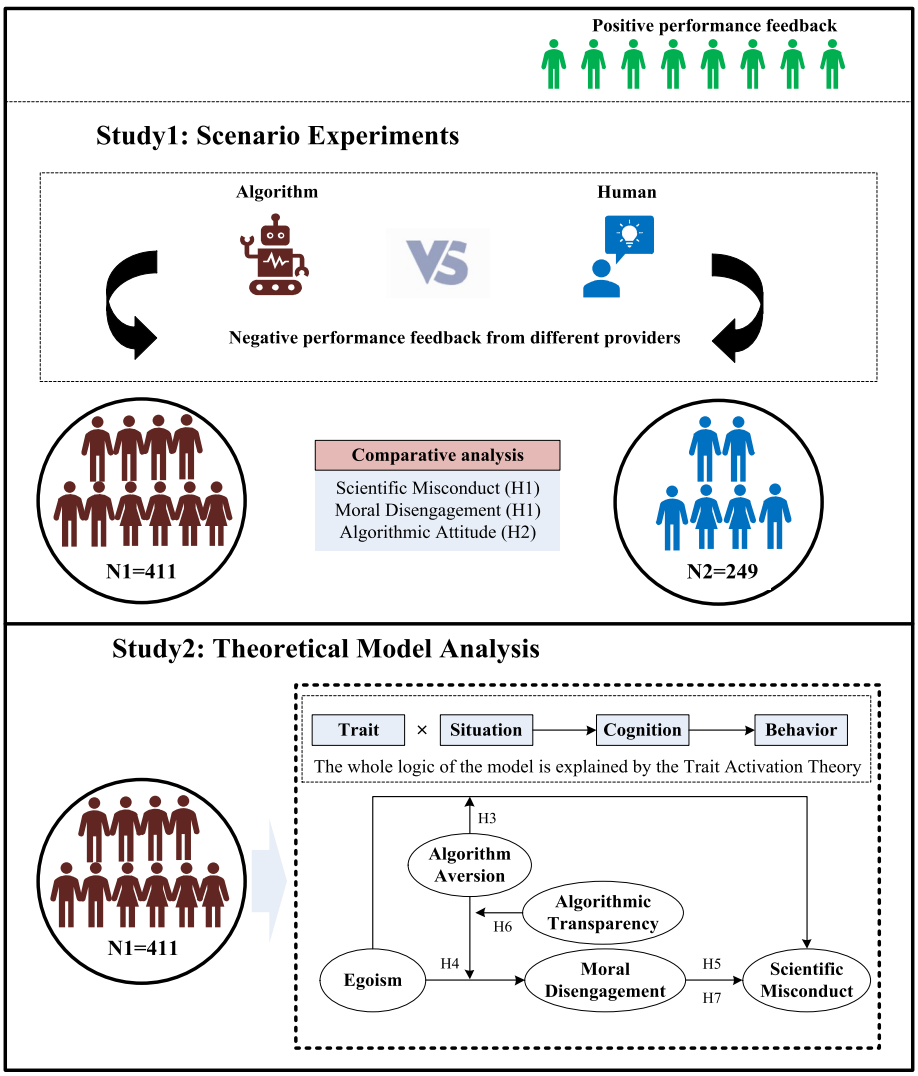


**Fig. 1** Design of experiments

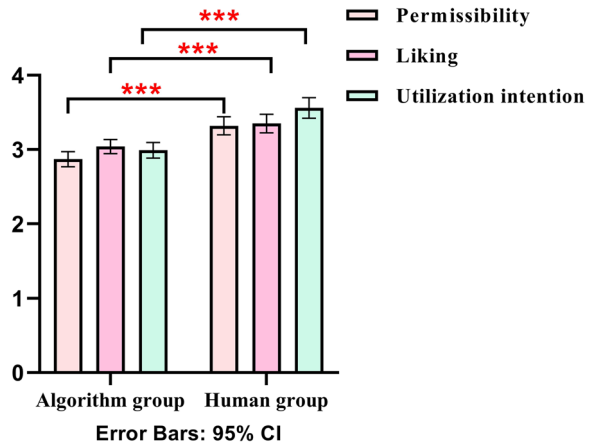


**Fig. 3** t-test results for the algorithm attitude between the algorithm group and the human group (***$p < 0.001$)

misconduct. The results showed that the main effect of the decision-maker was significant (Wilks' $\lambda = 0.820$, $F = 72.277$, $p < 0.001$, $\eta^2 p = 0.180$). Thus, hypothesis 1 was supported.

This study further examined the differences in algorithm attitudes between the algorithm group and the human group, and the results of the t-test were shown in Fig. 3. In this study, the reverse scores of algorithm attitudes were converted to positive scoring. The permissibility of the human group ($M = 3.32$, $SD = 0.97$) was significantly higher than that of the algorithm group ($M = 2.88$, $SD = 1.06$), with $t = 5.354$ ($p < 0.001$, Cohen's $d = 1.028$). The human group's ($M = 3.35$, $SD = 0.99$) level of liking was significantly higher than that of the algorithm group ($M = 3.04$, $SD = 0.97$), with $t = 3.938$ ($p < 0.001$, Cohen's $d = 0.974$). Finally, the utilization intention of the human group ($M = 3.36$, $SD = 1.12$) was significantly higher than that of the algorithm group ($M = 2.99$, $SD = 1.09$), with $t = 4.179$ ($p < 0.001$, Cohen's $d = 1.101$). Furthermore, this study conducted a MANOVA with decision-maker (algorithms vs. humans) as the independent variable and the permissibility, liking for algorithms, and utilization intention as dependent variables. The results showed that the main effect of decision-maker was significant (Wilks' $\lambda = 0.957$, $F = 9.762$, $p < 0.001$, $\eta^2 p = 0.043$). Thus, hypothesis 2 was supported.

## Study 2

### Descriptive analysis

This study conducted descriptive statistics and correlation analysis using SPSS 27.0. The means, standard deviations, and correlation coefficients of all variables are shown in Table 1. The results show that egoism is negatively correlated with scientific misconduct ($r = -0.574$, $p < 0.001$). Moral disengagement and algorithm aversion are positively correlated with scientific misconduct ($r = 0.131$, $p < 0.01$; $r = 0.364$, $p < 0.001$). This provides preliminary evidence for the hypotheses of this study. However, there is no correlation between egoism and moral disengagement, suggesting the possible existence of a moderating or mediating effect between these two variables.

### Confirmatory factor analysis

This study employed Mplus 8.3 to conduct confirmatory factor analysis (CFA). The results, as shown in Table 2, indicate that the four-factor model ($\chi^2/df = 2.848$, $CFI = 0.901$, $TLI = 0.891$, $RMSEA = 0.067$, $SRMR = 0.067$) demonstrates the best fit indices compared to other alternative models (e.g., one factor: $\chi^2/df = 6.492$, $CFI = 0.702$, $TLI = 0.676$, $RMSEA = 0.116$, $SRMR = 0.103$), Thus, the CFA results demonstrate good discriminant validity between all the variables.

### Empirical test

Utilizing the SPSS 27.0 PROCESS software, hierarchical regression analysis was implemented to examine hypotheses 3 to 7, and the results are shown in Table 3. For hypothesis 3 and 4, Model 2 and Model 5 were developed to test whether individuals with algorithm aversion experience egoism when they receive NPF from algorithms, which leads to scientific misconduct and moral disengagement. The results in Model 5 show that after controlling for demographic variables, the interaction between egoism and algorithm aversion is significant for scientific

**Table 1** Descriptive analysis of variables ($N = 411$)

| Variables | M | SD | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| 1.Algorithmic transparency | 3.11 | 1.173 | | | | |
| 2.Moral disengagement | 3.35 | 1.145 | 0.004 | | | |
| 3.Egoism | 3.16 | 0.937 | -0.557*** | 0.071 | | |
| 4.Algorithm aversion | 3.04 | 0.944 | 0.800*** | 0.028 | -0.599*** | |
| 5.Scientific misconduct | 3.15 | 0.662 | 0.311*** | 0.131** | -0.574*** | 0.364*** |

** $p < 0.01$,*** $p < 0.001$

**Table 2** Results of confirmatory factor analysis ($N=411$)

| Models | Factors | $\chi^2/df$ | CFI | TLI | RMSEA | SRMR |
|---|---|---|---|---|---|---|
| Four-factors model | EG, AA, MD, SM | 2.848 | 0.901 | 0.891 | 0.067 | 0.067 |
| Three-factors model 1 | EG + AA, MD, SM | 3.996 | 0.838 | 0.823 | 0.085 | 0.078 |
| Three-factors model 2 | EG, AA, MD + SM | 4.062 | 0.835 | 0.819 | 0.086 | 0.091 |
| Two-factors model 1 | EG + AA, MD + SM | 5.203 | 0.773 | 0.752 | 0.101 | 0.099 |
| Two-factors model 2 | EG, AA + MD + SM | 5.432 | 0.755 | 0.733 | 0.105 | 0.102 |
| One factor model | EG + AA + MD + SM | 6.492 | 0.702 | 0.676 | 0.116 | 0.103 |

EG represents egoism; AA represents algorithm aversion; MD represents moral disengagement; SM represents scientific misconduct

**Table 3** Hierarchical regression results of moderating effect

| Variables | Scientific misconduct | | Moral disengagement | | |
|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| Gender | 0.099* | -0.039 | -0.049 | -0.031 | -0.173 |
| Age | 0.086 | 0.050 | 0.348*** | 0.348*** | 0.389*** |
| Educational level | 0.056 | -0.038 | 0.217*** | 0.227*** | 0.345*** |
| Marriage | -0.018 | -0.016 | 0.013 | 0.010 | -0.089 |
| Year | 0.050 | 0.029 | -0.18* | -0.179* | -0.129 |
| Egoism | | -0.399*** | | 0.088 | 0.087 |
| Algorithm aversion | | 0.003 | | | -0.138 |
| Egoism × Algorithm aversion | | 0.069* | | | 0.428*** |
| Moral disengagement | | | | | |
| F | 2.183 | 0.354 | 8.660*** | 7.803*** | 14.083*** |
| $R^2$ | 0.026 | 27.564*** | 0.097 | 0.104 | 0.219 |

* $p<0.05$, *** $p<0.001$

misconduct ($\beta=0.069$, $p<0.05$). Thus, hypothesis 3 was supported. Model 1 shows that the interaction between egoism and algorithm aversion is significant for scientific misconduct ($\beta=0.428$, $p<0.001$). Thus, hypothesis 4 was supported.

Furthermore, we used a simple slope analysis [3] to examine the moderating effect of algorithm aversion at different levels (Mean ± 1SD). As shown in Fig. 4, compared to low algorithm aversion (Mean—1SD), high algorithm aversion (Mean + 1SD) more strongly enhanced the positive influence of egoism on moral disengagement. Thus, hypothesis 4 was further supported.

Additionally, we used Hayes' Bootstrap method (= 5000 times) to examine the mediating effect of moral disengagement on the relationship between egoism and scientific misconduct under the interaction of algorithm aversion, and the results are shown in Table 4. The results indicate that at three levels of algorithm aversion: low (Mean—1SD), medium (Mean), and high (Mean + 1SD), the mediation effects of moral disengagement were -0.032, 0.009, and 0.049, respectively, with 95% confidence intervals of [-0.061, -0.011], [-0.008, 0.027], and [0.024, 0.081]. This shows that under the activation of algorithm aversion, egoism lead to moral disengagement, and the level of moral disengagement significantly increases with higher algorithm aversion. Moreover, the influence of egoism on scientific misconduct through moral disengagement is also enhanced. Thus, hypothesis 5 was supported.

Next, this study examines H6, which proposes the direct influence of the three-way interaction among egoism, algorithm aversion, and algorithmic transparency on moral disengagement. As shown in Model 9 in Table 5, the interaction has a significant negative correlation with moral disengagement ($\beta=-0.236$, $p<0.01$). This indicates that the higher the algorithmic transparency, the more it can mitigate the activating effect of algorithm aversion on moral disengagement in individuals with egoism. To further test hypothesis 6, following the method of Dawson and Richer [33], we obtained four conditions:

(1) high algorithm aversion (Mean + 1SD) and high algorithmic transparency (Mean + 1SD),
(2) high algorithm aversion (Mean + 1SD) and low algorithmic transparency (Mean—1SD),

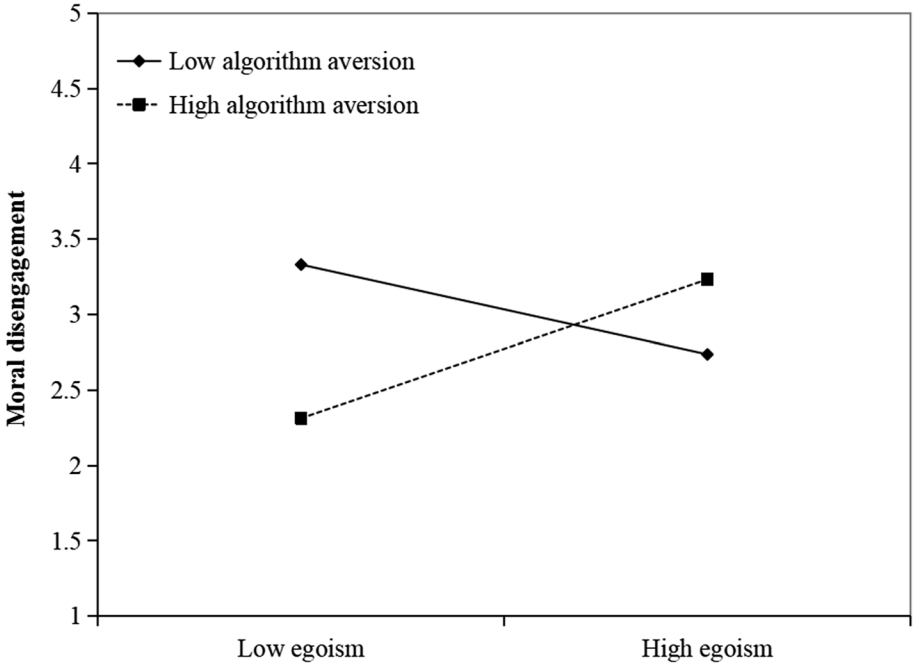


**Fig. 4** The moderating effect of algorithm aversion on the relationship between egoism and moral disengagement

**Table 4** Bootstrap method at different levels of algorithm aversion

| Algorithm aversion | The mediating effects of moral disengagement | Boot SE | Boot LLCI | Boot ULCI |
|---|---|---|---|---|
| -0.944 (low) | -0.032 | 0.013 | -0.061 | -0.011 |
| 0.000 (medium) | 0.009 | 0.009 | -0.008 | 0.027 |
| 0.944 (high) | 0.049 | 0.015 | 0.024 | 0.081 |

**Table 5** Hierarchical regression results of moderating effect

| Variables | Moral disengagement | | | |
|---|---|---|---|---|
| | Model 6 | Model 7 | Model 8 | Model 9 |
| Gender | -0.049 | -0.031*** | -0.093* | -0.106* |
| Age | 0.348*** | 0.343*** | 0.310*** | 0.281*** |
| Educational level | 0.217*** | 0.225 | 0.163*** | 0.153** |
| Marriage | 0.013 | 0.010 | -0.041 | -0.045 |
| Year | -0.180* | -0.173* | -0.150* | -0.139 |
| Egoism | | 0.097 | 0.094 | 0.170** |
| Algorithm aversion | | 0.051 | -0.060** | -0.138 |
| Algorithmic transparency | | -0.037 | -0.066 | -0.101 |
| Egoism×Algorithm aversion | | | 0.245* | 0.220* |
| Egoism×Algorithmic transparency | | | 0.028 | 0.050 |
| Algorithm aversion × Algorithmic transparency | | | -0.156* | -0.096 |
| Egoism×Algorithm aversion × Algorithmic transparency | | | | -0.236** |
| $R^2$ | 8.660*** | 0.105*** | 0.231*** | 0.244*** |
| F | 0.097 | 5.875 | 10.900 | 10.729 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(3) low algorithm aversion (Mean—1SD) and high algorithmic transparency (Mean + 1SD),
(4) low algorithm aversion (Mean—1SD) and low algorithmic transparency (Mean—1SD).

Based on these combinations, we plotted the three-way interaction effect, as shown in Fig. 5. It can be seen that in the scenario of high algorithm aversion and low algorithmic transparency, the positive slope of the fitted curve for the influence of egoism on moral disengagement is the steepest, indicating that in this scenario, the positive influence of egoism on moral disengagement is the most significant. Thus, hypothesis 6 was supported.

We continued to use the Bootstrap method to test the moderated mediation effect, the results are shown in Table 6. The results indicate that under conditions of low algorithm aversion and low algorithmic transparency, as well as low algorithm aversion and high algorithmic transparency, the 95% confidence intervals were [-0.487, 0.017] and [-0.277, 0.638], respectively. Both intervals include 0, indicating that when algorithm aversion is low, the influence of egoism on scientific misconduct through moral disengagement is insignificant. Conversely, under conditions of high algorithm aversion and low algorithmic transparency, and high algorithm aversion and high algorithmic transparency, the 95% confidence intervals were [0.103, 0.961] and [0.160, 0.547], respectively. Both intervals exclude 0, and the mediation effect value for high algorithm aversion and low algorithmic transparency is 0.532, which is higher than the mediation effect value of 0.354 for high algorithm aversion and high algorithmic transparency. Therefore, it can be concluded that the influence of egoism on scientific misconduct through moral disengagement is most significant among researchers with high algorithm aversion and

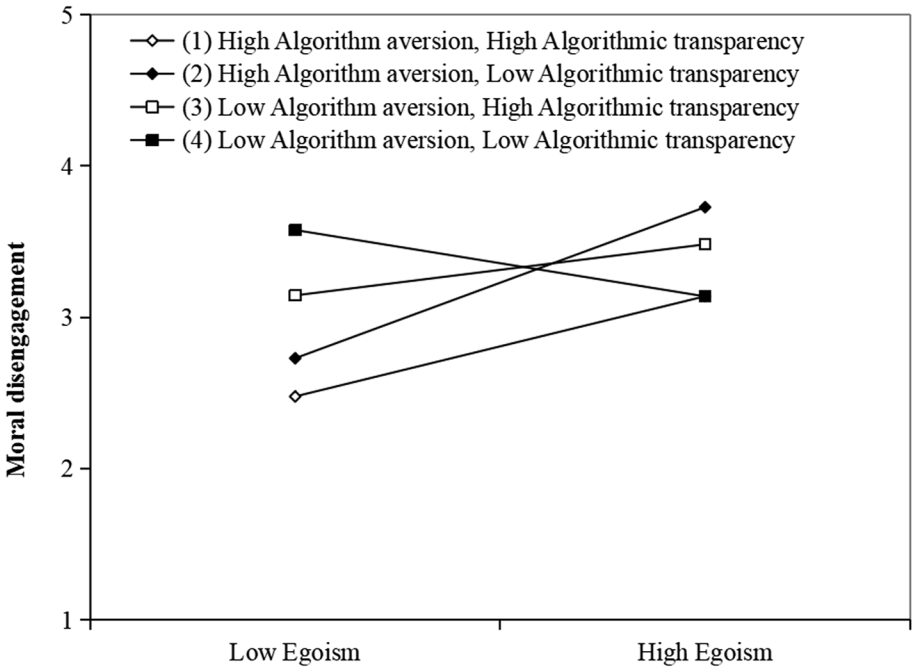


**Fig. 5** Three-way interaction effects of egoism, algorithm aversion, and moral disengagement

**Table 6** Mediating effects and confidence intervals of Bootstrap method at different levels of moderating variables

| Algorithm aversion | Algorithmic transparency | The mediating effects of moral disengagement | Boot SE | Boot LLCI | Boot ULCI |
|---|---|---|---|---|---|
| -0.944 (low) | -1.173 (low) | -0.235 | 0.128 | -0.487 | 0.017 |
| -0.944 (low) | 1.173 (high) | 0.180 | 0.233 | -0.277 | 0.638 |
| 0.944 (high) | -1.173 (low) | 0.532 | 0.218 | 0.103 | 0.961 |
| 0.944 (high) | 1.173 (high) | 0.354 | 0.098 | 0.160 | 0.547 |

low algorithmic transparency. Thus, hypothesis 7 was further supported.

## Discussion and conclusions

Based on a scenario-based experiment in medical fields, we explore how NPF from algorithms or humans influences researchers' moral cognition, algorithmic attitudes, and their tendency toward scientific misconduct. Firstly, empirical analysis indicates that there are pronounced differences between the effects of NPF from algorithms and those from humans on medical researchers. Specifically, in the domains of moral disengagement, algorithm aversion, and scientific misconduct, medical researchers who received NPF from algorithms showed significantly higher levels of these effects compared to the group receiving NPF from humans. Secondly, in the group receiving NPF from algorithms, we find that the interaction between egoism and algorithm aversion has a positive impact on moral disengagement among medical researchers, subsequently leading to an increase in scientific misconduct. As the level of algorithm aversion increases, the positive relationship between egoism and moral disengagement becomes stronger. Moreover, our investigation further demonstrates that moral disengagement serves as a critical mediating variable in the complex interaction between egoism, algorithm aversion, and the scientific misconduct. Specifically, the relationship between an individual's egoism and their engagement in scientific misconduct is mediated by moral disengagement, which is influenced by algorithm aversion. This suggests that the researchers' tendency to engage in unethical conduct is not solely determined by their egoism or algorithm aversion, but is also dependent on their ability for morally disengage from their behaviors. Finally, algorithmic transparency moderates the mediating role of moral disengagement between algorithm aversion and the relationship with egoism and scientific misconduct, indicating that higher levels of algorithmic transparency among medical researchers correspond to a mitigated activation effect of algorithm aversion on egoism.

### Theoretical implications

Our findings reveal that the question of whether algorithms or humans should provide NPF to medical researchers is a complex one that goes beyond technical considerations. The fundamental difference lies in the nature of the NPF and how it is provided to medical researchers. NPF from algorithms, with its inherent objectivity and consistency, lacks the emotional context and nuanced understanding that humans can provide. This can lead to a disconnect between the feedback and the researcher's emotional state, potentially leading to a sense of detachment or even resentment. In contrast, NPF from humans can be tailored to the medical researcher's circumstances and emotions, fostering a sense of empathy and understanding. This personalized approach can help medical researchers to feel supported and motivated to make the necessary changes in their undervalued performance. Additionally, humans can provide an explanation for the NPF, which can help medical researchers to understand the reasons behind the undervalued performance and how to address the issues raised. Therefore, the study makes the following three theoretical contributions.

Firstly, this study extends the comparison of decision-making providers—algorithms and humans—into the area of NPF in medical research. By exploring the differences in NPF providers concerning scientific misconduct, ethical cognition, and decision-maker preferences, further highlighting the consistency in individuals' perceptions against algorithms. Previous studies have found that individuals exhibit a consistently stable rejection and aversion towards algorithmic management in the workplace [21, 79, 82]. Our findings are generally consistent with these results: both the general public and managers demonstrate apprehension and opposition towards algorithmic decision-making [1, 34, 50, 95]. This sentiment manifests itself in distrust and dislike of algorithms, revealing the limitations of algorithms in the decision-making process as well as the inherent bias of humans towards it. However, while these studies provide deep insights into algorithmic management, they primarily focus on the algorithms themselves, paying less attention to the differences between algorithms and humans in the NPF decision-making process. Furthermore, existing studies predominantly focus on specific groups such as gig workers [30, 39, 111, 133, 135], with relatively less attention paid to medical researchers. Therefore, this study further enriches our understanding of the differences between algorithms and humans by expanding the research field. Further, most of the current literature on performance is on algorithmic positive feedback [19, 72], leader positive feedback [22, 57, 66, 114], leader's NPF [59, 93, 119, 137, 144]. However, in the process of NPF for medical researchers, scientific integrity, ethical cognition, and decision-maker preferences may be affected in more complex ways. By comparison with positive performance feedback, NPF is more likely to trigger emotional volatility and cognitive biases in individuals [12]. That is, this study focuses on NPF, enhancing our understanding of the differences between algorithms and humans in negative performance evaluations. These findings not only highlight the challenges algorithms may encounter when applied in medical research but also provide new

perspectives and considerations for research management and decision-making.

Secondly, drawing on trait activation theory, this study analyzes the underlying mechanisms of scientific misconduct when NPF is provided by algorithms. This exploration not only extends the application of trait activation theory to the field of algorithmic management but also proposes a novel pathway through which trait activation triggers scientific misconduct. Previous research on algorithm management has primarily been grounded in theories such as control theory [2], socio-technical system theory [140], surveillance theory [4], and self-determination theory [136], focusing on the impact of algorithms on individual behavior and the underlying psychological mechanisms. For instance, while self-determination theory emphasizes that individuals actively exert effort and engage in proactive behaviors to fulfill their psychological needs in response to algorithm management, it does not thoroughly explore how stressful environments, as specific scenarios, and then interact with individual traits to influence behavior [117]. This study employs trait activation theory, considering NPF from algorithms and algorithm aversion as significant contextual cues. These cues activate latent traits within individuals, which then interact with egoism to foster scientific misconduct. The results indicate that algorithm aversion plays a critical role in triggering medical researchers' egoism. When individuals exhibit a strong aversion to algorithms, their intrinsic egoism is more likely to be activated in response to NPF, thereby predisposing them to engage in inappropriate scientific behavior. This finding offers a significant contrast to the traditional situational strength theory [67, 87], which emphasizes the interaction between individual traits and specific contexts [63]. This interaction influences the behavioral decisions of medical researchers. Consequently, a detailed exploration of algorithm management practices through the lens of trait activation theory not only enhances our comprehensive understanding of researchers' traits and behaviors but provides a crucial theoretical basis for formulating effective research management and regulatory strategies.

Thirdly, the study unveils the internal mechanisms behind scientific misconduct among medical researchers in the scenario of NPF from algorithms. On one hand, from the perspective of moral disengagement, it offers a new viewpoint on the influence mechanisms of scientific misconduct under algorithmic NPF, proposing an "Trait × Situation → Cognition → Behavior" pathway. This pathway underscores the interaction between egoism and specific scenarios, and how this interaction influences an individual's moral cognition, then shaping their behavior. Ethical issues arising from algorithms have been primarily focused on in previous research [49, 69, 85, 89], conversely insufficient attention has been paid to the ethical behavior of individuals. Additionally, these studies predominantly emphasize theoretical analysis, with relatively less application of empirical testing. This study illustrates the intricate moral mechanisms that drive scientific misconduct among researchers in the algorithmic management, addressing the gaps prevalent in existing literature. It provides a robust theoretical framework that supports the development of effective preventive and ethical intervention strategies. On the other hand, to decrease the negative impacts of NPF from algorithms, this study enriches the research on boundary conditions of this influence mechanism from the perspective of algorithmic transparency. Current scholarly discourse emphasizes the essential role of algorithmic transparency in protecting individual rights and ensuring social fairness [35, 73, 116]. It is crucial for strengthening algorithmic trust and ensuring the sustained advancement of technology [45, 112]. Our study supports similar conclusions: algorithmic transparency significantly decreases algorithm aversion and its negative consequences [32, 141]. Among different levels of algorithm aversion, individuals with a higher algorithmic transparency show lower levels of moral disengagement compared to those with lower algorithmic transparency. In summary, the pathways and boundary conditions influencing scientific misconduct have been explained in this study from the perspectives of moral disengagement and algorithmic transparency. Not only does this study provide a conceptual framework for alleviating unethical practices in scientific research, but it also establishes a theoretical foundation and offers insights for the further management of algorithm-driven decision processes.

#### Practical implications

The practical implications of this study are as follows:

Firstly, managers should explore internal and external reasons for NPF from algorithms. It is essential to implement personalized tutoring and training programs tailored to researchers' unique needs, thereby providing skills enhancement and psychological support to help researchers overcome frustration and achieve mutual benefit with algorithms. Moreover, our study indicates that anthropomorphizing algorithms will bring more benefits as well as positive experience [23]. Consequently, it is advised that humanizing algorithm management be adopted to enhance the trust and acceptance of algorithms, thereby improving researchers' emotional response to NPF.

Then, managers are advised to strengthen researchers' sense of ethical responsibility, prompting research ethics education such as institutional advocacy, exemplary leadership, cultural immersion and so on. Organizations

should ensure that all researchers are aware of the definitions and consequences of scientific misconduct, as well as how to conduct their research activities while observing ethical principles. In the process of recruitment, institutions should prioritize ethical values as a crucial screening criterion through questionnaire surveys and in-depth interviews. Additionally, for the purpose of proactively preventing and addressing scientific misconduct, it is also advisable to establish probationary periods and conduct regular assessments.

Finally, organizations should further promote algorithmic transparency and engagement. On the one hand, it is necessary to clarify the processes and evidence of algorithms as decision-making tools, then making their inherently complex and opaque processes more comprehensible and interpretable. On the other hand, it is imperative that the operating rules and logical foundations of the algorithm be clearly elucidated to users [29, 74]. Attitudinal effects of mere exposure reveal a psychological phenomenon: repeated exposure to specific external information can enhance individuals' preference for it. Accordingly, increasing familiarity with algorithms and understanding their operations is expected to boost their acceptance. This could enhance the popularity of algorithm management across society. To achieve this goal, the key is that organizations should foster greater openness and transparency in the design, testing, and adjustment stages of the algorithm. Especially, this entails inviting researchers to engage and providing them an opportunity to offer specialized advice for the improvement and optimization of the algorithms. It not only enhances the researches' understanding of the algorithmic decision-making process and reduces the misunderstandings and contradictions due to opacity, but also aligns the algorithms' applications with reality's requests, as well as fostering a virtuous cycle of continuous improvement and optimization.

### Limitations and future directions

Despite the valuable insights gained from this study, several limitations should be acknowledged, which provide opportunities for future research.

Firstly, this study employed the scenario-based experiment that may have certain uncertainties and biases. On the one hand, relying only on situational materials to design experimental conditions may make it difficult to ensure the natural state in real environment, leading to limited external validity of results. On the other hand, participants may not provide truthful answers due to social desirability effect when it comes to sensitive topics such as scientific misconduct. Therefore, future study should enhance the authenticity of scenario-based experiments by advancing and evaluating more intricate and diverse scenarios. Additionally, employing a mixed method design is needed, including quantitative situational experiments and qualitative in-depth interviews, to yield more comprehensive and precise data. Secondly, this research merely focuses on moral disengagement as a mediating mechanism, neglecting other potential factors such as self-efficacy, cognitive dissonance, and stress perception. Future studies are encouraged to extend to other mediating variables in order to fully understand the relationship between algorithm aversion and scientific misconduct. Thirdly, this study primarily examines the negative impacts of algorithm aversion on researchers, but it does not explore potential positive effects such as algorithmic improvement. Future studies could aim to develop a more comprehensive and integrated framework to explore both sides of algorithm feedback. This approach will provide insights into the potential benefits of algorithm aversion for researchers and investigate additional situational variables such as algorithmic appreciation.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12910-024-01121-0.

Supplementary Material 1.

**Acknowledgements**
The authors would like to express their sincere thanks to all the participants.

**Clinical trial number**
Not applicable.

**Authors' contributions**
Ganli Liao and Feiwen Wang wrote the main manuscript text, Qichao Zhang prepared the figures and tables, Ganli Liao offered the fundings, and Wenhui Zhu provided methodology. All authors reviewed the manuscript.

**Funding**
This research was funded by the National Social Science Fund of China (No. 24CGL122).

**Data availability**
The datasets utilized and analyzed during the current study are not publicly available but can be obtained from the first author upon reasonable request.

## Declarations

**Ethics approval and consent to participate**
The study design and ethical considerations were approved by the Ethics Committee of the School of Economics and Management, Beijing Information Science and Technology University. All participants provided informed written consent, ensuring data privacy and adherence to ethical guidelines.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Business School, Beijing Information Science and Technology University, Beijing, China. [2]Zhongguancun Smart City Co., Ltd, Beijing, China.

Received: 23 August 2024 Accepted: 17 October 2024
Published online: 23 October 2024

**References**

1. Acikgoz Y, Davison KH, Compagnone M, Laske M. Justice perceptions of artificial intelligence in selection. Int J Sel Assess. 2020;28(4):399–416. https://doi.org/10.1111/ijsa.12306.
2. Ågerfalk PJ. Artificial intelligence as digital agency. Eur J Inf Syst. 2020;29(1):1–8. https://doi.org/10.1080/0960085X.2020.1721947.
3. Aiken LS, West SG. Multiple regression: Testing and interpreting interactions-institute for social and economic research. Sage: Newbury Park; 1991. p. 167–8.
4. Almeida D, Shmarko K, Lomas E. The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. AI and Ethics. 2022;2(3):377–87. https://doi.org/10.1007/s43681-021-00077-w.
5. Ashforth BE, Anand V. The normalization of corruption in organizations. Res Organ Behav. 2003;25:1–52. https://doi.org/10.1016/S0191-3085(03)25001-2.
6. Balcazar F, Hopkins BL, Suarez Y. A critical, objective review of performance feedback. J Organ Behav Manag. 1985;7(3–4):65–89. https://doi.org/10.1300/J075v07n03_05.
7. Ball KS, Margulis ST. Electronic monitoring and surveillance in call centres: a framework for investigation. N Technol Work Employ. 2011;26(2):113–26. https://doi.org/10.1111/j.1468-005X.2011.00263.x.
8. Bandura A. Social foundations of thought and action. Englewood Cliffs, NJ. 1986;1986(23–28):2. https://doi.org/10.5465/amr.1987.4306538.
9. Bandura A, Barbaranelli C, Caprara GV, Pastorelli C. Mechanisms of moral disengagement in the exercise of moral agency. J Pers Soc Psychol. 1996;71(2):364–74. https://doi.org/10.1037/0022-3514.71.2.364.
10. Bandura A. A commentary on moral disengagement: the rhetoric and the reality. Am J Psychol. 2018;131(2):246–51. https://doi.org/10.5406/amerjpsyc.131.2.0246.
11. Basaad S, Bajaba S, Basahal A. Uncovering the dark side of leadership: How exploitative leaders fuel unethical pro-organizational behavior through moral disengagement. Cogent Bus Manage. 2023;10(2):2233775. https://doi.org/10.1080/23311975.2023.2233775.
12. Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD. Bad is stronger than good. Rev Gen Psychol. 2001;5(4):323–70. https://doi.org/10.1037/1089-2680.5.4.323.
13. Bigman YE, Gray K. People are averse to machines making moral decisions. Cognition. 2018;181:21–34. https://doi.org/10.1016/j.cognition.2018.08.003.
14. Bigman YE, Wilson D, Arnestad MN, Waytz A, Gray K. Algorithmic discrimination causes less moral outrage than human discrimination. J Exp Psychol Gen. 2023;152(1):4–27. https://doi.org/10.1037/xge0001250.
15. Blasi A. Bridging moral cognition and moral action: A critical review of the literature. Psychol Bull. 1980;88(1):1–45. https://doi.org/10.1037/0033-2909.88.1.1.
16. Bowles S, Gintis H. Reciprocity, self-interest, and the welfare state. Nordic J Pol Econ. 2000;26(1):33–53 http://www.nopecjournal.org/NOPEC_2000_a02.pdf.
17. Bozdag E. Bias in algorithmic filtering and personalization. Ethics Inf Technol. 2013;15:209–27. https://doi.org/10.1007/s10676-013-9321-6.
18. Barsky A. Investigating the effects of moral disengagement and participation on unethical work behavior. J Bus Ethics. 2011;104:59–75. https://doi.org/10.1007/s10551-011-0889-7.
19. Bucher EL, Schou PK, Waldkirch M. Pacifying the algorithm–Anticipatory compliance in the face of algorithmic management in the gig economy. Organization. 2021;28(1):44–67. https://doi.org/10.1177/1350508420961531.
20. Buhmann A, Paßmann J, Fieseler C. Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse. J Bus Ethics. 2020;163(2):265–80. https://doi.org/10.1007/s10551-019-04226-4.
21. Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. J Behav Decis Mak. 2020;33(2):220–39. https://doi.org/10.1002/bdm.2155.
22. Byron K, Khazanchi S. Rewards and creative performance: a meta-analytic test of theoretically derived hypotheses. Psychol Bull. 2012;138(4):809–30. https://doi.org/10.1037/a0027652.
23. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nat Hum Behav. 2021;5(12):1636–42. https://doi.org/10.1038/s41562-021-01146-0.
24. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. J Mark Res. 2019;56(5):809–25. https://doi.org/10.1177/0022243719851788.
25. Chen M, Chen CC, Sheldon OJ. Relaxing moral reasoning to win: How organizational identification relates to unethical pro-organizational behavior. J Appl Psychol. 2016;101(8):1082–96. https://doi.org/10.1037/apl0000111.
26. Cianci AM, Klein HJ, Seijts GH. The effect of negative feedback on tension and subsequent performance: The main and interactive effects of goal content and conscientiousness. J Appl Psychol. 2010;95(4):618–30. https://doi.org/10.1037/a0019130.
27. Claybourn M. Relationships between moral disengagement, work characteristics and workplace harassment. J Bus Ethics. 2011;100(2):283–301. https://doi.org/10.1007/s10551-010-0680-1.
28. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff. 2014;33(7):1139–47. https://doi.org/10.1377/hlthaff.2014.0048.
29. Confalonieri R, Coba L, Wagner B, Besold TR. A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Min Knowl Discov. 2021;11(1):e1391. https://doi.org/10.1002/widm.139.
30. Curchod C, Patriotta G, Cohen L, Neysen N. Working for an algorithm: Power asymmetries and agency in online work settings. Adm Sci Q. 2020;65(3):644–76. https://doi.org/10.1177/0001839219867024.
31. Dahling JJ, Whitaker BG, Levy PE. The development and validation of a new Machiavellianism scale. J Manag. 2009;35(2):219–57. https://doi.org/10.1177/0149206308318617.
32. Dargnies, M. P., Hakimov, R., & Kübler, D. (2024). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. Management Science, ahead of print (ahead of print). https://doi.org/10.1287/mnsc.2022.02774.
33. Dawson JF, Richter AW. Probing three-way interactions in moderated multiple regression: development and application of a slope difference test. J Appl Psychol. 2006;91(4):917–26. https://doi.org/10.1037/0021-9010.91.4.917.
34. Diab DL, Pui SY, Yankelevich M, Highhouse S. Lay perceptions of selection decision aids in US and non-US samples. Int J Sel Assess. 2011;19(2):209–16. https://doi.org/10.1111/j.1468-2389.2011.00548.x.
35. Diakopoulos N, Koliska M. Algorithmic transparency in the news media. Digit Journal. 2017;5(7):809–28. https://doi.org/10.1080/21670811.2016.1208053.
36. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen. 2015;144(1):114–26. https://doi.org/10.1037/xge0000033.
37. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Manage Sci. 2018;64(3):1155–70. https://doi.org/10.1287/mnsc.2016.2643.
38. Dietvorst BJ, Bharti S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci. 2020;31(10):1302–14. https://doi.org/10.1177/0956797620948841.
39. Duggan J, Sherman U, Carbery R, McDonnell A. Algorithmic management and app-work in the gig economy: A research agenda for employment relations and HRM. Hum Resour Manag J. 2020;30(1):114–32. https://doi.org/10.1111/1748-8583.12258.
40. Erickson D, Holderness DK Jr, Olsen KJ, Thornock TA. Feedback with feeling? How emotional language in feedback affects individual

performance. Acc Organ Soc. 2022;99:101329. https://doi.org/10.1016/j.aos.2021.101329.
41. Fast NJ, Jago AS. Privacy matters...or does It? Algorithms, rationalization, and the erosion of concern for privacy. Curr Opin Psychol. 2020;31:44–8. https://doi.org/10.1016/j.copsyc.2019.07.011.
42 Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009;41(4):1149–60. https://doi.org/10.3758/BRM.41.4.1149.
43. Fida R, Paciello M, Tramontano C, Fontaine RG, Barbaranelli C, Farnese ML. An integrative approach to understanding counterproductive work behavior: The roles of stressors, negative emotions, and moral disengagement. J Bus Ethics. 2015;130:131–44. https://doi.org/10.1007/s10551-014-2209-5.
44. Gillespie, N., Lockey, S., & Curtis, C. (2021). Trust in Artificial Intelligence: A Five Country Study. The University of Queensland and KPMG Australia. https://doi.org/10.14264/e34bfa3.
45. Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. Acad Manag Ann. 2020;14(2):627–60. https://doi.org/10.5465/annals.2018.0057.
46. Graham KA, Resick CJ, Margolis JA, Shao P, Hargis MB, Kiker JD. Egoistic norms, organizational identification, and the perceived ethicality of unethical pro-organizational behavior: A moral maturation perspective. Human Relations. 2020;73(9):1249–77. https://doi.org/10.1177/0018726719862851.
47. Gratch J, Fast NJ. The power to harm: AI assistants pave the way to unethical behavior. Curr Opin Psychol. 2022;47:101382. https://doi.org/10.1016/j.copsyc.2022.101382.
48. Grimmelikhuijsen S. Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. Public Adm Rev. 2023;83(2):241–62. https://doi.org/10.1111/puar.13483.
49. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205–11. https://doi.org/10.1136/medethics-2019-105586.
50. Haesevoets T, De Cremer D, Dierckx K, Van Hiel A. Human-machine collaboration in managerial decision making. Comput Hum Behav. 2021;119:106730. https://doi.org/10.1016/j.chb.2021.106730.
51. Hambrick DC, Finkelstein S, Mooney AC. Executive job demands: New insights for explaining strategic decisions and leader behaviors. Acad Manag Rev. 2005;30(3):472–91. https://doi.org/10.5465/amr.2005.17293355.
52. Hesselmann F, Graf V, Schmidt M, Reinhart M. The visibility of scientific misconduct: A review of the literature on retracted journal articles. Curr Sociol. 2017;65(6):814–45. https://doi.org/10.1177/0011392116663807.
53. Hill AD, Johnson SG, Greco LM, O'Boyle EH, Walter SL. Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. J Manage. 2021;47(1):105–43. https://doi.org/10.1177/0149206320960533.
54. Huang MH, Rust RT. Artificial intelligence in service. J Serv Res. 2018;21(2):155–72. https://doi.org/10.1177/1094670517752459.
55 Huang GH, Wellman N, Ashford SJ, Lee C, Wang L. Deviance and exit: The organizational costs of job insecurity and moral disengagement. J Appl Psychol. 2017;102(1):26–42. https://doi.org/10.1037/apl0000158.
56. Hystad SW, Mearns KJ, Eid J. Moral disengagement as a mechanism between perceptions of organisational injustice and deviant work behaviours. Safety Sci. 2014;68:138–45. https://doi.org/10.1016/j.ssci.2014.03.012.
57. Ilgen DR, Fisher CD, Taylor MS. Consequences of individual feedback on behavior in organizations. J Appl Psychol. 1979;64(4):349–71. https://doi.org/10.1037/0021-9010.64.4.349.
58. Ilies R, Guo CY, Lim S, Yam KC, Li X. Happy but uncivil? Examining when and why positive affect leads to incivility. J Bus Ethics. 2020;165:595–614. https://doi.org/10.1007/s10551-018-04097-1.
59. Ilies R, Judge TA. Goal regulation across time: the effects of feedback and affect. J Appl Psychol. 2005;90(3):453–67. https://doi.org/10.1037/0021-9010.90.3.453.
60. Jago AS. Algorithms and authenticity. Acad Manage Discov. 2019;5(1):38–56. https://doi.org/10.5465/amd.2017.0002.
61. Jauernig J, Uhl M, Walkowitz G. People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency. Philos Technol. 2022;35(1):2. https://doi.org/10.1007/s13347-021-00495-y.
62. Johnson A, Dey S, Nguyen H, Groth M, Joyce S, Tan L, Harvey SB. A review and agenda for examining how technology-driven changes at work will impact workplace mental health and employee well-being. Australian J Manag. 2020;45(3):402–24. https://doi.org/10.1177/0312896220922292.
63. Judge TA, Zapata CP. The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. Acad Manag J. 2015;58(4):1149–79. https://doi.org/10.5465/amj.2010.0837.
64. Jung M, Seiter M. Towards a better understanding on mitigating algorithm aversion in forecasting: An experimental study. J Manag Control. 2021;32(4):495–516. https://doi.org/10.1007/s00187-021-00326-3.
65. Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion. Eur Conf Information Systems, 1–16. https://aisel.aisnet.org/ecis2020_rp/168.
66. Kahai SS, Huang R, Jestice RJ. Interaction effect of leadership and communication media on feedback positivity in virtual teams. Group Org Manag. 2012;37(6):716–51. https://doi.org/10.1177/1059601112462061.
67. Keeler KR, Kong W, Dalal RS, Cortina JM. Situational strength interactions: Are variance patterns consistent with the theory? J Appl Psychol. 2019;104(12):1487–513. https://doi.org/10.1037/apl0000416.
68. Kellogg KC, Valentine MA, Christin A. Algorithms at work: The new contested terrain of control. Acad Manag Ann. 2020;14(1):366–410. https://doi.org/10.5465/annals.2018.0174.
69. Kraemer F, Van Overveld K, Peterson M. Is there an ethics of algorithms? Ethics Inf Technol. 2011;13:251–60. https://doi.org/10.1007/s10676-010-9233-7.
70. Kuhn KM, Maleki A. Micro-entrepreneurs, dependent contractors, and instaserfs: Understanding online labor platform workforces. Acad Manag Perspect. 2017;31(3):183–200. https://doi.org/10.5465/amp.2015.0111.
71. Langer M, König CJ. Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. Hum Resour Manag Rev. 2023;33(1):100881. https://doi.org/10.1016/j.hrmr.2021.100881.
72. Lata LN, Burdon J, Reddel T. New tech, old exploitation: Gig economy, algorithmic control and migrant labour. Sociol Compass. 2023;17(1):e13028. https://doi.org/10.1111/soc4.13028.
73. Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–26. https://doi.org/10.1145/3359284.
74. Leichtmann B, Humer C, Hinterreiter A, Streit M, Mara M. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. Comput Hum Behav. 2023;139:107539. https://doi.org/10.1016/j.chb.2022.107539.
75. LePine JA, Podsakoff NP, LePine MA. A meta-analytic test of the challenge stressor–hindrance stressor framework: An explanation for inconsistent relationships among stressors and performance. Acad Manag J. 2005;48(5):764–75. https://doi.org/10.5465/amj.2005.18803921.
76. Liu NTY, Kirshner SN, Lim ET. Is algorithm aversion WEIRD? A cross-country comparison of individual-differences and algorithm aversion. J Retail Consum Serv. 2023;72:103259. https://doi.org/10.1016/j.jretconser.2023.103259.
77. Locke EA, Woiceshyn J. Why businessmen should be honest: The argument from rational egoism. J Organ Behav. 1995;16(5):405–14. https://doi.org/10.1002/job.4030160503.
78. Logg JM, Minson JA, Moore DA. Algorithm appreciation: People prefer algorithmic to human judgment. Organ Behav Hum Decis Process. 2019;151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.
79. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res. 2019;46(4):629–50. https://doi.org/10.1093/jcr/ucz013.
80. Maasland C, Weißmüller KS. Blame the machine? Insights from an experiment on algorithm aversion and blame avoidance in computer-aided human resource management. Front Psychol. 2022;13:779028. https://doi.org/10.3389/fpsyg.2022.779028.
81. Mahmoudi M, Ameli S, Moss S. The urgent need for modification of scientific ranking indexes to facilitate scientific progress and diminish

academic bullying. BioImpacts: BI. 2020;10(1):5–7. https://doi.org/10.15171/bi.2019.30.
82. Mahmud H, Islam AN, Ahmed SI, Smolander K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technol Forecast Soc Chang. 2022;175:121390. https://doi.org/10.1016/j.techfore.2021.121390.
83. Mai R, Hoffmann S, Lasarov W, Buhs A. Ethical products= less strong: How explicit and implicit reliance on the lay theory affects consumption behaviors. J Bus Ethics. 2019;158:659–77. https://doi.org/10.1007/s10551-017-3669-1.
84. Mai, K. M., Welsh, D. T., Wang, F., Bush, J., & Jiang, K. (2022). Supporting creativity or creative unethicality? Empowering leadership and the role of performance pressure. J Business Ethics, 1–21. https://doi.org/10.1007/s10551-021-04784-6.
85. Martin K. Ethical implications and accountability of algorithms. J Bus Ethics. 2019;160(4):835–50. https://doi.org/10.1007/s10551-018-3921-3.
86. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol. 2001;52(1):397–422. https://doi.org/10.1146/annurev.psych.52.1.397.
87. Meyer RD, Dalal RS, Hermida R. A review and synthesis of situational strength in the organizational sciences. J Manag. 2010;36(1):121–40. https://doi.org/10.1177/0149206309349309.
88. Mintzberg H. The design school: reconsidering the basic premises of strategic management. Strateg Manag J. 1990;11(3):171–95. https://doi.org/10.1002/smj.4250110302.
89. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. Big Data Soc. 2016;3(2):2053951716679679. https://doi.org/10.1177/2053951716679679.
90. Molden DC, Dweck CS. Finding "meaning" in psychology: a lay theories approach to self-regulation, social perception, and social development. Am Psychol. 2006;61(3):192–203. https://doi.org/10.1037/0003-066X.61.3.192.
91. Moore C, Detert JR, Klebe Treviño L, Baker VL, Mayer DM. Why employees do bad things: Moral disengagement and unethical organizational behavior. Pers Psychol. 2012;65(1):1–48. https://doi.org/10.1111/j.1744-6570.2011.01237.x.
92. Moore C. Moral disengagement in processes of organizational corruption. J Bus Ethics. 2008;80(1):129–39. https://doi.org/10.1007/s10551-007-9447-8.
93. Motro D, Comer DR, Lenaghan JA. Examining the effects of negative performance feedback: the roles of sadness, feedback self-efficacy, and grit. J Bus Psychol. 2021;36(3):367–82. https://doi.org/10.1007/s10869-020-09689-1.
94. Newman DT, Fast NJ, Harmon DJ. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. Organ Behav Hum Decis Process. 2020;160:149–67. https://doi.org/10.1016/j.obhdp.2020.03.008.
95. Nørskov S, Damholdt MF, Ulhøi JP, Jensen MB, Mathiasen MK, Ess CM, Seibt J. Employers' and applicants' fairness perceptions in job interviews: using a teleoperated robot as a fair proxy. Technol Forecast Soc Chang. 2022;179:121641. https://doi.org/10.1016/j.techfore.2022.12164.
96. Ogunfowora B, Stackhouse M, Maerz A, Varty C, Hwang C, Choi J. The impact of team moral disengagement composition on team performance: The roles of team cooperation, team interpersonal deviance, and collective extraversion. J Bus Psychol. 2021;36:479–94. https://doi.org/10.1007/s10869-020-09688-2.
97. Paciello M, Fida R, Tramontano C, Lupinetti C, Caprara GV. Stability and change of moral disengagement and its impact on aggression and violence in late adolescence. Child Dev. 2008;79(5):1288–309. https://doi.org/10.1111/j.1467-8624.2008.01189.x.
98. Paruzel-Czachura M, Baran L, Spendel Z. Publish or be ethical? Publishing pressure and scientific misconduct in research. Research Ethics. 2021;17(3):375–97. https://doi.org/10.1177/1747016120980562.
99. Paulhus DL, John OP. Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. J Pers. 1998;66(6):1025–60. https://doi.org/10.1111/1467-6494.00041.
100. Probst TM, Petitta L, Barbaranelli C, Austin C. Safety-related moral disengagement in response to job insecurity: Counterintuitive effects of perceived organizational and supervisor support. J Bus Ethics. 2020;162(2):343–58. https://doi.org/10.1007/s10551-018-4002-3.
101. Prue DM, Fairbank JA. Performance feedback in organizational behavior management: A review. J Organ Behav Manag. 1981;3(1):1–16. https://doi.org/10.1300/J075v03n01_01.
102. Qin G, Zhang L. How compulsory citizenship behavior depletes individual resources—a moderated mediation model. Curr Psychol. 2024;43(2):969–83. https://doi.org/10.1007/s12144-023-04386-7.
103. Quade MJ, Greenbaum RL, Mawritz MB. "If only my coworker was more ethical": When ethical and performance comparisons lead to negative emotions, social undermining, and ostracism. J Bus Ethics. 2019;159(2):567–86. https://doi.org/10.1007/s10551-018-3841-2.
104. Quan W, Shu F, Yang M, Larivière V. Publish and flourish: investigating publication requirements for PhD students in China. Scientometrics. 2023;128(12):6675–93. https://doi.org/10.1007/s11192-023-04854-8.
105. Raamkumar AS, Yang Y. Empathetic conversational systems: A review of current advances, gaps, and opportunities. IEEE Trans Affect Comput. 2022;14(4):2722–39. https://doi.org/10.1109/TAFFC.2022.3226693.
106. Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1–13). https://doi.org/10.1145/3173574.3173677.
107. Rahman HA. The invisible cage: Workers' reactivity to opaque algorithmic evaluations. Adm Sci Q. 2021;66(4):945–88. https://doi.org/10.1177/00018392211010118.
108. Ravid DM, Tomczak DL, White JC, Behrend TS. EPM 20/20: A review, framework, and research agenda for electronic performance monitoring. J Manag. 2020;46(1):100–26. https://doi.org/10.1177/0149206319869435.
109. Raza A, Ishaq MI, Jamali DR, Zia H, Haj-Salem N. Testing workplace hazing, moral disengagement and deviant behaviors in hospitality industry. Int J Contemp Hosp Manag. 2024;36(3):743–68. https://doi.org/10.1108/IJCHM-06-2022-0715.
110. Reich T, Kaju A, Maglio SJ. How to overcome algorithm aversion: Learning from mistakes. J Consum Psychol. 2023;33(2):285–302. https://doi.org/10.1002/jcpy.1313.
111. Rosenblat A, Stark L. Algorithmic labor and information asymmetries: A case study of Uber's drivers. Int J Commun. 2016;10:3758–84. https://doi.org/10.2139/ssrn.2686227.
112. Schiff DS, Schiff KJ, Pierson P. Assessing public value failure in government adoption of artificial intelligence. Public Administration. 2022;100(3):653–73. https://doi.org/10.1111/padm.12742.
113. Schildt H. Big data and organizational design–the brave new world of algorithmic management and computer augmented transparency. Innovation. 2017;19(1):23–30. https://doi.org/10.1080/14479338.2016.1252043.
114. Schroeder J, Fishbach A. How to motivate yourself and others? Intended and unintended consequences. Res Organ Behav. 2015;35:123–41. https://doi.org/10.1016/j.riob.2015.09.001.
115. Schweitzer ME, Ordóñez L, Douma B. Goal setting as a motivator of unethical behavior. Acad Manag J. 2004;47(3):422–32. https://doi.org/10.5465/20159591.
116. Shin D, Park YJ. Role of fairness, accountability, and transparency in algorithmic affordance. Comput Hum Behav. 2019;98:277–84. https://doi.org/10.1016/J.CHB.2019.04.019.
117. Shuang ZHAO, Jun MA. Algorithmic management and employee creativity: create by the potential of algorithm or stay in the digital cage. J Systems Manage. 2024;33(3):782–800. https://doi.org/10.3969/j.issn1005-2542.2024.03.016.
118. Simon LS, Rosen CC, Gajendran RS, Ozgen S, Corwin ES. Pain or gain? Understanding how trait empathy impacts leader effectiveness following the provision of negative feedback. J Appl Psychol. 2022;107(2):279–97. https://doi.org/10.1037/apl0000882.
119. Su W, Zhang Y. Supervisor negative feedback, subordinate prevention focus and performance: testing a mediation model. Curr Psychol. 2023;42(28):24613–22. https://doi.org/10.1007/s12144-022-03494-0.
120. Swami V, Chamorro-Premuzic TOMAS, Snelgar R, Furnham A. Egoistic, altruistic, and biospheric environmental concerns: A path analytic investigation of their determinants. Scand J Psychol. 2010;51(2):139–45. https://doi.org/10.1111/j.1467-9450.2009.00760.x.

121. Tambe P, Cappelli P, Yakubovich V. Artificial intelligence in human resources management: Challenges and a path forward. Calif Manage Rev. 2019;61(4):15–42. https://doi.org/10.1177/0008125619867910.
122. Tett RP, Burnett DD. A personality trait-based interactionist model of job performance. J Appl Psychol. 2003;88(3):500–17. https://doi.org/10.1037/0021-9010.88.3.500.
123. Tett RP, Guterman HA. Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. J Res Pers. 2000;34(4):397–423. https://doi.org/10.1006/jrpe.2000.2292.
124. Tong S, Jia N, Luo X, Fang Z. The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. Strateg Manag J. 2021;42(9):1600–31. https://doi.org/10.1002/smj.3322.
125. Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, Floridi L. The ethics of algorithms: key problems and solutions. Ethics Governance Pol Artif Intell. 2021;144:97–123. https://doi.org/10.1007/978-3-030-81907-1_8.
126. Turel O, Kalhan S. Prejudiced against the Machine? Implicit Associations and the Transience of Algorithm Aversion. MIS Quarterly. 2023;47(4):1396. https://doi.org/10.25300/MISQ/2022/17961.
127. Turilli M, Floridi L. The ethics of information transparency. Ethics Inf Technol. 2009;11:105–12. https://doi.org/10.1007/s10676-009-9187-9.
128. Tzini K, Jain K. Unethical behavior under relative performance evaluation: Evidence and remedy. Hum Resour Manage. 2018;57(6):1399–413. https://doi.org/10.1002/hrm.21913.
129. Van der Wees PJ, Nijhuis-van der Sanden MW, van Ginneken E, Ayanian JZ, Schneider EC, Westert GP. Governing healthcare through performance measurement in Massachusetts and the Netherlands. Health Policy. 2014;116(1):18–26. https://doi.org/10.1016/j.healthpol.2013.09.009.
130. Wang Q, Huang Y, Jasin S, Singh PV. Algorithmic transparency with strategic users. Manage Sci. 2023;69(4):2297–317. https://doi.org/10.1287/mnsc.2022.4475.
131. Weigel RH, Hessing DJ, Elffers H. Egoism: Concept, measurement and implications for deviance. Psychology, Crime and Law. 1999;5(4):349–78. https://doi.org/10.1080/10683169908401777.
132. Welsh DT, Ordóñez LD. The dark side of consecutive high performance goals: Linking goal setting, depletion, and unethical behavior. Organ Behav Hum Decis Process. 2014;123(2):79–89. https://doi.org/10.1016/j.obhdp.2013.07.006.
133. Wiener M, Cram WA, Benlian A. Algorithmic control and gig workers: a legitimacy perspective of Uber drivers. Eur J Inf Syst. 2023;32(3):485–507. https://doi.org/10.1080/0960085X.2021.1977729.
134. Wilson HJ, Daugherty PR. Collaborative intelligence: Humans and AI are joining forces. Harv Bus Rev. 2018;96(4):114–23.
135. Wood AJ, Graham M, Lehdonvirta V, Hjorth I. Good gig, bad gig: autonomy and algorithmic control in the global gig economy. Work Employ Soc. 2019;33(1):56–75. https://doi.org/10.1177/0950017018785616.
136. Xia Q, Chiu TK, Lee M, Sanusi IT, Dai Y, Chai CS. A self-determination theory (SDT) design approach for inclusive and diverse artificial intelligence (AI) education. Comput Educ. 2022;189:104582. https://doi.org/10.1016/j.compedu.2022.104582.
137. Xing L, Sun JM, Jepsen D. Feeling shame in the workplace: examining negative feedback as an antecedent and performance and well-being as consequences. J Organ Behav. 2021;42(9):1244–60. https://doi.org/10.1002/job.2553.
138. Yu L, Miao M, Liu W, Zhang B, Zhang P. Scientific misconduct and associated factors: a survey of researchers in three Chinese tertiary hospitals. Account Res. 2021;28(2):95–114. https://doi.org/10.1080/08989621.2020.1809386.
139. Yu TW, Chen TJ. Online travel insurance purchase intention: A transaction cost perspective. J Travel Tour Mark. 2018;35(9):1175–86. https://doi.org/10.1080/10548408.2018.1486781.
140. Yu X, Xu S, Ashton M. Antecedents and outcomes of artificial intelligence adoption and application in the workplace: the socio-technical system theory perspective. Inf Technol People. 2023;36(1):454–74. https://doi.org/10.1108/ITP-04-2021-0254.
141. Zerilli J, Bhatt U, Weller A. How transparency modulates trust in artificial intelligence. Patterns. 2022;3(4):100455. https://doi.org/10.1016/j.patter.2022.100455.
142. Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol. 2019;32:661–83. https://doi.org/10.1007/s13347-018-0330-6.
143. Zhang N, Guo M, Jin C, Xu Z. Effect of medical researchers' creative performance on scientific misconduct: a moral psychology perspective. BMC Med Ethics. 2022;23(1):137. https://doi.org/10.1186/s12910-022-00876-8.
144. Zhu C, Zhang F, Ling CD, Xu Y. Supervisor feedback, relational energy, and employee voice: The moderating role of leader–member exchange quality. Int J Human Res Manage. 2023;34(17):3308–35. https://doi.org/10.1080/09585192.2022.2119093.


## Publisher's Note