The Moral Deficiency Effect of Artificial Intelligence Decision-Making: Mechanisms and Mitigation Strategies

Authors: Hu Xiaoyong, Li Mufeng, Li Yue, Li Kai, Yu Feng

Date: 2025-09-07T18:16:23+00:00

Abstract

As artificial intelligence assumes an increasingly prominent role in major decision-making, the ethical issues it engenders have garnered significant attention. This study systematically reveals the dual-pathway mechanism of the artificial intelligence moral deficiency effect and corresponding intervention strategies by integrating mind perception theory and moral dyad theory. The findings demonstrate that individuals exhibit significantly weaker moral responses to unethical decisions made by artificial intelligence compared to human decision-makers; the perception of lower agency and experience in artificial intelligence relative to humans accounts for this moral deficiency effect in AI decision-making. A comprehensive intervention scheme combining anthropomorphism strategies targeting artificial intelligence and expectation adjustment strategies targeting humans can substantially elevate moral response levels toward AI. Unlike other disciplines that emphasize design-level principles and methods for fair algorithms, this research adopts a psychological perspective, focusing on differential psychological reactions to AI versus human decision-making. This viewpoint not only offers novel insights for addressing social problems stemming from algorithmic bias and constructing fair algorithms, but also expands the theoretical boundaries of "algorithm ethics" research.

Full Text

Moral Deficiency in AI Decision-Making: Underlying Mechanisms and Mitigation Strategies

HU Xiaoyong[1], LI Mufeng[2], LI Yue[1], LI Kai[1], YU Feng[1]
([1] Department of Psychology, Wuhan University, Wuhan 430072, China)
([2] Faculty of Psychology, Southwest University, Chongqing 400715, China)

Abstract

As artificial intelligence (AI) assumes an increasingly prominent role in high-stakes decision-making, the ethical challenges it raises have become a pressing concern. This study systematically reveals the dual-pathway mechanism underlying the moral deficiency effect in AI decisions and proposes mitigation strategies by integrating mind perception theory with moral dyad theory. The findings demonstrate that people's moral responses to unethical AI decisions are significantly weaker than those to human decision-makers. Compared to human agents, the perception of lower agency and experience in AI systems is identified as the root cause of this moral deficiency effect. Furthermore, a comprehensive intervention package combining anthropomorphic strategies targeting AI systems and expectancy adjustment strategies targeting human observers significantly enhances moral responses to AI decisions. Unlike other disciplines that focus primarily on design-level principles and methods for fair algorithms, this research adopts a psychological perspective that centers on differential psychological reactions to AI versus human decision-making. This approach not only provides novel insights for addressing social problems caused by algorithmic bias and constructing fair algorithms, but also expands the theoretical boundaries of "algorithm ethics" research.

Keywords: artificial intelligence, moral deficiency effect, mind perception, anthropomorphism, expectancy adjustment

Artificial intelligence, as an interdisciplinary technological integration, has transcended traditional tool boundaries and demonstrated human-like intelligence in complex cognitive dimensions such as perception, reasoning, learning, and decision-making (Rai et al., 2022). Its applications have permeated critical social domains including judicial sentencing, medical resource allocation, and financial credit, deeply intervening in decision-making processes that affect fundamental human rights such as the right to survival and development. Empirical research demonstrates that algorithmic systems systematically undervalue female resumes in employment contexts (Dastin, 2022), underestimate illness severity among socially disadvantaged patients in medical diagnosis (Obermeyer et al., 2019), and exhibit significant moral decision-making biases in judicial sentencing (Angwin, 2016), educational assessment (Wang et al., 2024), and credit approval (Bartlett et al., 2022). These systematic biases not only expose the ethical risks of technological black boxes but also trigger deep concerns about social fairness mechanisms.

Existing research has predominantly focused on exogenous perspectives such as technical governance (Song & Yeung, 2024), legal regulation (Magrani, 2019), and ethical framework construction, while relatively neglecting the psychological response mechanisms of human recipients of AI-driven unethical decisions—a core variable (Langer & Landers, 2021). Psychological research reveals a particularly concerning moral deficiency effect in AI decision-making: when AI serves as the decision-making agent, the public exhibits significantly reduced

moral sensitivity and responsibility attribution tendencies, with weakened willingness to punish even when facing identical transgressions to those of human decision-makers (Bigman & Gray, 2018; Wilson et al., 2022; Xu et al., 2022). This cognitive bias triggers a series of risks: first, it incentivizes organizations to use AI as a tool for moral responsibility evasion (Danaher, 2016); second, it exacerbates the rights relief dilemma for harmed groups (Bonezzi & Ostinelli, 2021); and third, it leads to gradual degradation of social moral benchmarks (Awad et al., 2020). Therefore, revealing the psychological mechanisms of the moral deficiency effect and proposing mitigation strategies is not only crucial for paradigm innovation in human-computer interaction theory but also an urgent priority for constructing AI ethical governance systems and maintaining the foundation of social justice.

1.1 The Moral Deficiency Effect in AI Decision-Making

Compared to unethical decisions made by humans, people exhibit weaker responses to AI's unethical decisions, manifesting as less blame, responsibility attribution, and moral outrage, along with reduced inclination toward moral punishment or action. Researchers term this phenomenon the moral deficiency effect in AI decision-making (Bigman et al., 2023; Hu et al., 2024).

In terms of moral cognition, when AI and humans cause equivalent decision-making errors, people tend to mitigate AI's responsibility (Lima et al., 2020). Research shows that prescription errors by robot pharmacists generate significantly less dissatisfaction and accountability intention than those by human pharmacists (Leo & Huh, 2020); when AI exhibits bias in judicial domains or engages in core ethical issues involving harm and betrayal, people are more inclined to rationalize its unethical behavior (Maninger & Shank, 2022; Shank et al., 2019). This effect demonstrates cross-cultural stability—across Asia, Africa, and America, people universally perceive AI's faults and blameworthiness as less severe (Wilson et al., 2022). Regarding moral emotions, AI's unethical decisions elicit significantly weaker negative emotional responses than humans. For instance, in trust games involving monetary distribution, AI's betrayal behavior provokes lower anger levels than human betrayal (Schniter et al., 2020). Whether in service failure scenarios like lost luggage or severe moral violations such as gender discrimination in hiring, AI systems consistently generate less moral outrage than human decision-makers under equivalent circumstances (Pavone et al., 2023; Bigman et al., 2023). At the behavioral level, the public's willingness to punish AI unethical behavior and engage in resistance is similarly weaker. Research indicates that when facing discriminatory systems designed by AI—whether gender-based or education-based—subjects' willingness to sign petitions opposing the system and their punishment tendencies are significantly reduced (Bonezzi & Ostinelli, 2021; Xu et al., 2022). Even in extreme situations causing severe harm, such as AI detonating bombs resulting in fatalities, the punishment severity (e.g., years of imprisonment) is significantly lower than for human perpetrators (Guidi et al., 2021).

In summary, extensive empirical research demonstrates that across various moral scenarios, people exhibit weaker moral responses to AI's unethical decisions. This moral deficiency effect manifests across multiple dimensions including moral cognition, moral emotion, and moral behavior.

1.2 Psychological Mechanisms of the Moral Deficiency Effect in AI Decision-Making

Why do AI's unethical decisions universally elicit weaker moral responses than humans? Existing research points to mind perception as the root cause—people only generate moral responses when they perceive a moral agent possesses a certain level of mind (Chakroff & Young, 2015). Mind perception theory posits that people perceive mind along two independent dimensions: agency and experience (Gray et al., 2007). However, traditional research on defining "moral agents," particularly perspectives based on moral dyad theory, has limitations. This theory emphasizes that agency perception is the core prerequisite for holding entities accountable for their wrongful actions (Gray et al., 2012; Malle, 2019), while relatively neglecting the role of experience. Although scholars have suggested that simultaneously endowing AI with agency and experience makes it more closely resemble a "quasi-agent" capable of reflection (Behdadi & Munthe, 2020; Hu et al., 2024), previous research has not systematically revealed how these two dimensions independently and jointly weaken moral responses to AI. This study argues that the agency-only theoretical perspective is incomplete—experience perception not only concerns an entity's qualification as a "moral patient" but is also indispensable for constructing its identity as a "moral agent." The core mechanism lies in the fact that experience forms the foundation of empathy and moral emotions. Only a subject capable of understanding and feeling others' pain, joy, and other emotional states is considered to possess the capacity to form moral norms and anticipate the emotional consequences of its actions on others (Decety & Cowell, 2018). This constitutes the psychological basis for moral responsibility. Since AI is widely perceived as having far less experience than humans and lacking the ability to understand others' emotions, it is viewed as a "morally incomplete" actor (Gray et al., 2007; Liu et al., 2019; Malle, 2019). Preliminary empirical research supports the plausibility of experience as a mediating pathway. One study found that the public's character judgments of AI (e.g., good/evil traits) were significantly lower than for humans, and this difference was mediated by experience perception (Shank et al., 2021). Character judgment itself represents an evaluation of a moral actor's internal states. Another more direct study manipulated AI's harmful behavior experimentally and found that when AI was portrayed as possessing higher experience, participants were more inclined to condemn and punish it. This indicates that when AI is believed to understand the pain caused by its actions (i.e., possesses high experience), people view it as a genuine moral actor that "knows what it's doing," with experience serving a partial mediating role (Sullivan et al., 2022). Therefore, the perceived deficiency in AI's experience dimension constitutes the second key psychological pathway leading to the moral

deficiency effect.

In summary, existing research provides preliminary evidence for the independent mediating roles of agency and experience in AI's moral deficiency effect. However, few studies have examined both dimensions within a unified framework. Although Hu et al. (2024) proposed a "dual-path parallel mediation model" through literature review, identifying perceived agency and experience as key mechanisms affecting AI decision-making's moral deficiency effect, this model remains at the theoretical hypothesis stage without empirical support. More importantly, that review failed to adequately argue why experience is a necessary dimension for becoming a "moral agent." Building on this, our study revises and expands classic moral dyad theory, proposing that a complete "moral agent" requires participation from both dimensions of mind. The public's muted response to AI's unethical decisions stems precisely from their perception that AI lacks both sufficient "autonomous intention" (agency deficit) and necessary "emotional empathy" (experience deficit). Therefore, this study proposes a parallel mediation model hypothesis: the moral deficiency effect in AI decision-making emerges through the public's reduced perception of both agency and experience.

1.2.1 The Mediating Role of Agency

Agency refers to an entity's capacity for intention, reasoning, goal pursuit, and communication (Gray et al., 2007). Agency is closely related to moral responsibility—the stronger an individual's autonomy and the clearer their intentions and motivations, the more responsibility people believe they should bear for decisions and behaviors (Gray et al., 2007). Evidence indicates that while people attribute some agency to AI, its level is significantly lower than that of humans (Malle, 2019; Weisman et al., 2017). This perceptual gap constitutes the first psychological pathway of AI's moral deficiency effect. Preliminary evidence supports the mediating role of perceived agency in AI's moral deficiency effect. For example, one study using free will as an agency indicator found that AI was perceived as possessing less free will, leading to lower desire for moral punishment; moreover, free will mediated the relationship between AI discriminatory decisions and reduced desire for moral punishment (Xu et al., 2022). Another study found that because AI behavior is constrained by programming, people's perception of its free will is weakened, consequently reducing moral responsibility attribution (Bigman et al., 2019). In gender discrimination contexts, research also shows AI is perceived as having lower discriminatory motivation, resulting in less moral outrage toward AI's gender-discriminatory decisions; discriminatory motivation mediates the difference in moral outrage between AI and human hiring decisions (Bigman et al., 2023).

1.2.2 The Mediating Role of Experience

Experience refers to an entity's capacity to perceive emotions, pain, and subjective experiences (Gray et al., 2007). Experience not only defines who can

"be harmed" (moral patient) but also profoundly influences who can be considered a complete "harm-doer" (moral agent). Its core mechanism lies in the fact that experience forms the foundation of empathy and moral emotions. Only a subject capable of understanding and feeling others' emotional states—such as pain and joy—is considered to possess the capacity to form moral norms and anticipate the emotional consequences of its actions on others (Decety & Cowell, 2018). This constitutes the psychological basis for moral responsibility. Since AI is widely perceived as having far less experience than humans and lacking the ability to understand others' emotions, it is viewed as a "morally incomplete" actor (Gray et al., 2007; Liu et al., 2019; Malle, 2019). Preliminary empirical research supports the plausibility of experience as a mediating pathway. One study found that the public's character judgments of AI (e.g., good/evil traits) were significantly lower than for humans, and this difference was mediated by experience perception (Shank et al., 2021). Character judgment itself represents an evaluation of a moral actor's internal states. Another more direct study manipulated AI's harmful behavior experimentally and found that when AI was portrayed as possessing higher experience, participants were more inclined to condemn and punish it. This indicates that when AI is believed to understand the pain caused by its actions (i.e., possesses high experience), people view it as a genuine moral actor that "knows what it's doing," with experience serving a partial mediating role (Sullivan et al., 2022). Therefore, the perceived deficiency in AI's experience dimension constitutes the second key psychological pathway leading to the moral deficiency effect.

In summary, existing research provides preliminary evidence for the independent mediating roles of agency and experience in AI's moral deficiency effect. However, few studies have examined both dimensions within a unified framework. Although Hu et al. (2024) proposed a "dual-path parallel mediation model" through literature review, identifying perceived agency and experience as key mechanisms affecting AI decision-making's moral deficiency effect, this model remains at the theoretical hypothesis stage without empirical support. More importantly, that review failed to adequately argue why experience is a necessary dimension for becoming a "moral agent." Building on this, our study revises and expands classic moral dyad theory, proposing that a complete "moral agent" requires participation from both dimensions of mind. The public's muted response to AI's unethical decisions stems precisely from their perception that AI lacks both sufficient "autonomous intention" (agency deficit) and necessary "emotional empathy" (experience deficit). Therefore, this study proposes a parallel mediation model hypothesis: the moral deficiency effect in AI decision-making emerges through the public's reduced perception of both agency and experience.

1.3 Intervention Strategies for AI's Moral Deficiency Effect

Since AI's moral deficiency effect originates from the public's dual perceptual deficits in agency and experience, enhancing perception along these two dimen-

sions constitutes the theoretical cornerstone for mitigating the effect (Gray et al., 2012). However, current research on intervention strategies, while touching upon this issue, often lacks a unified theoretical framework and systematic comparison across different pathways. Existing explorations can be categorized into two main approaches: first, technically-oriented interventions that modify AI's own design to enhance its perceived mind; and second, cognitively-oriented interventions that adjust human observers' psychological expectations to reshape their moral response patterns.

1.3.1 Anthropomorphism

The most intuitive strategy for intervening in AI's moral deficiency effect is direct modification of AI itself, with anthropomorphism being the most extensively studied approach. Anthropomorphism enhances mind signals emitted by non-human entities by endowing them with human-like appearance, intentions, or emotional characteristics (Lin et al., 2022; Melián-González et al., 2021; Zhang et al., 2022). Existing research has preliminarily confirmed the effectiveness of this pathway. First, multiple studies demonstrate that anthropomorphic design (e.g., simulated human images, humanoid forms) significantly enhances public perception of AI's agency and experience (Qian & Wan, 2024; Kamide et al., 2013). Second, other independent studies confirm that this enhanced mind perception effectively translates into stronger moral responses. For example, subjects are more inclined to attribute accident responsibility to anthropomorphically designed autonomous vehicles (Waytz et al., 2014) and generate stronger moral outrage toward unethical decisions made by AI with humanoid appearance or perceived "intentionality" (Nijssen et al., 2023; Sullivan et al., 2022). Similarly, when AI is given names or simulated emotional expressions, the public's moral evaluation of its unfair behavior becomes stricter (Laakasuo et al., 2021). However, previous research remains fragmented, mostly isolating verification of single links like "anthropomorphism → mind perception" or "mind perception → moral response," failing to form a complete causal chain. Based on integrating these findings, this study proposes that anthropomorphism, as an external intervention, fundamentally works by systematically enhancing people's perception of AI's agency and experience levels, thereby triggering stronger moral accountability.

1.3.2 Expectancy Adjustment

Different from modifying AI itself, another intervention pathway directly targets human observers by adjusting their psychological expectations of AI. The core logic lies in leveraging expectancy violation theory: by presetting higher moral or performance standards, the negative emotions (e.g., disappointment, anger) triggered when AI makes mistakes can "compensate" for the portion of moral response missing due to insufficient mind perception (Lew & Walther, 2023). When interacting with AI, people unconsciously apply social norms and form specific expectations based on initial impressions of AI (Nass & Moon, 2000;

Srinivasan & Sarial-Abi, 2021). Once these expectations are violated by AI's actual behavior, strong emotional and cognitive evaluations are triggered (Burgoon et al., 1988). Existing research provides preliminary evidence for this. For example, the public holds much higher safety expectations for autonomous vehicles than for human drivers, so when accidents occur, this "high expectation–low performance" gap triggers stronger blame (Liu et al., 2019). Similarly, when AI is preset as "cold and detached," its utilitarian decisions receive some forgiveness for meeting expectations; conversely, if preset with high moral standards, the same decisions trigger stronger negative evaluations (Zhang et al., 2022; Grimes et al., 2021). While these findings preliminarily support the effectiveness of expectancy adjustment as an intervention strategy, its specific pathways remain unclear. Integrating expectancy violation theory with mind perception theory and related empirical evidence, this study argues that raising expectations is essentially a cognitive intervention that forcibly presets a "high mind standard" evaluation framework for human observers. Within this framework, people temporarily suspend their default underestimation of AI's mind level and instead apply standards for a "high-agency, high-experience" subject to AI, thereby stimulating stronger moral responses.

In summary, existing research has explored possibilities for intervening in AI's moral deficiency effect from two different perspectives—"modifying AI" (anthropomorphism) and "guiding humans" (expectancy adjustment)—and confirmed their respective potential (Hu et al., 2024; Lin et al., 2022; Srinivasan & Sarial-Abi, 2021). However, the current major limitation lies in artificially separating these two pathways to examine their independent effects while ignoring the complex interactions that may exist between them in the real world. For example, does a highly anthropomorphized AI naturally trigger higher user expectations? Conversely, do high expectations for AI prompt people to pay more attention to its anthropomorphic features? To address this research gap, this study examines both pathways within an integrated framework for the first time. We argue that whether through anthropomorphism directly enhancing AI's mind signals or through expectancy adjustment indirectly elevating human evaluation standards, their ultimate effects converge on the perception and evaluation of AI's agency and experience. Therefore, we propose that single-dimension interventions may have an effectiveness ceiling, while a combined intervention integrating "technical" and "cognitive" approaches may produce stronger effects through synergistic interactions. Based on this, this study hypothesizes that compared to single anthropomorphic or expectancy adjustment interventions, a combined intervention package can more significantly enhance public perception of AI's agency and experience, ultimately maximizing the intensity of moral responses to its unethical decisions.

1.4 Overview of Studies

Following a "effect–mechanism–intervention" research logic, this paper conducts systematic empirical exploration in three stages, aiming to achieve local val-

idation, mechanism refinement, and intervention expansion of the theoretical model. Study 1 validates the robustness of the moral deficiency effect in Chinese cultural contexts using culturally adapted experimental materials. The experiment employs moral dilemma scenarios with Chinese cultural characteristics to compare subjects' moral responses to identical unethical decisions made by humans versus AI systems. Study 2 builds on the theoretical model proposed by Hu et al. (2024), integrating mind perception theory with moral dyad theory to propose a dual-path parallel mediation model that consolidates fragmented psychological mechanisms from previous research into two core pathways—perceived agency and perceived experience. This study designs three sub-experiments: first using experimental methods to examine the mediating roles of agency and experience separately, then employing questionnaire methods to simultaneously test their parallel mediation effects. The combination of experimental and questionnaire methods overcomes the limitations of traditional questionnaire methods (which cannot establish causal relationships between independent and mediating variables or between mediating and dependent variables) while also addressing experimental methods' inability to construct parallel mediation models. Study 3 develops an integrated solution to enhance perceived agency and experience based on Study 2's findings, proposing a combined anthropomorphism and expectancy adjustment intervention strategy, and tests it through a double-blind randomized controlled experiment. Building on existing theoretical frameworks, this research achieves systematic advancement from theoretical modeling and mechanism verification to intervention design, providing causal evidence and local empirical support for AI ethics psychology research, as well as theoretical foundations and practical guidance for psychological interventions in future AI governance.

2 Study 1: The Moral Deficiency Effect in AI Decision-Making

In recent years, with the rapid development of AI technology, its application in decision-making processes has become increasingly widespread. Scholars have noted that AI may "replicate" and "amplify" biases from human society during decision-making, thereby triggering moral judgment issues (Bonezzi & Ostinelli, 2021). However, systematic consensus has not yet formed regarding differences in moral responses between AI and humans. Based on this, this experiment constructs three scenarios—educational discrimination, age discrimination, and gender discrimination—to test Hypothesis H1: Compared to humans, people exhibit weaker moral responses to AI making unethical decisions.

2.1.1 Participants

We used G*Power 3.1 to estimate the required sample size (independent samples t-test, α=0.05, power=0.90, d=0.5), determining a minimum of 172 participants (Faul et al., 2007). The online experiment was conducted through the "Naodao" platform (Naodao.com), adopting widely supported key control measures from

existing literature to ensure internal and external validity. First, regarding participant identity and status control, the platform employed certification, IP address verification, and CAPTCHA mechanisms to prevent interference from "professional participants" or bots (Douglas et al., 2023). To ensure participant attention, various programmatic measures replaced offline experimenter supervision: including manipulation check instructions (Mancosu et al., 2019), mandatory full-screen mode and mouse trajectory monitoring (Hauser et al., 2018), and multiple attention check questions (Curran, 2016), with participants failing these checks being terminated from the experiment. Second, regarding experimental environment and equipment control, standardized instructions guided participants to use designated devices (computers) and browsers (Chrome), with self-reports on environmental interference and device type collected at the experiment's conclusion. Third, for data quality control, minimum/maximum completion time limits were set, with blank questionnaires (Little & Rubin, 2020), invalid responses (Curran, 2016), and patterned responses (Griffith & Peterson, 2006) eliminated, and anomalous data from technical issues or cheating strictly cleaned during data processing. Ultimately, 176 valid datasets were collected, including 85 females (48.3%); participants ranged from 15–49 years old, with a mean age of 24.49 (SD=5.10).

To ensure balanced allocation across experimental groups, we compared demographic characteristics between groups. First, chi-square independence test results for gender distribution showed no significant difference between groups, $^{2}(1) = 0.17$, p = 0.68, ϕ =0.03. The AI group included 40 males and 40 females; the human group included 51 males and 45 females. Second, independent samples t-test results for age indicated that the AI group's mean age (M = 25.25, SD =6.24) did not differ significantly from the human group (M = 23.85, SD =3.82), t(174) = 1.82, p = 0.07, Cohen's d = 0.28. Overall, the two groups showed no systematic differences in key demographic variables, indicating successful random assignment and providing a foundation for subsequent analyses.

2.1.2 Experimental Design

A mixed design of 2 (agent: human vs. AI) × 3 (discrimination scenario: education, age, gender) was employed, with agent as a between-subjects variable and discrimination scenario as a within-subjects variable. The dependent variables were ratings of moral response, moral cognition, moral emotion, and moral behavior.

2.1.3 Experimental Materials and Procedure

First, participants randomly read text materials describing discrimination implemented by either human or AI decision-makers. To ensure material validity and applicability, a formal expert validation procedure was conducted. Adapted from Bigman et al. (2023), the materials comprised six independent discrimination scenarios (3 types × 2 agents); through textual equivalence control, only

the decision-making agent (AI/human) and related pronouns were manipulated while all other content remained identical. The expert panel consisted of one associate professor specializing in intergroup prejudice research and four post-doctoral researchers (2) and doctoral students (2) with publication records in AI psychology. Experts were required to read all six text materials and independently rate each material on three dimensions using a 7-point Likert scale (1=completely disagree; 7=completely agree) based on Lynn's (1986) content validation framework: 1) situational realism (likelihood of the scenario occurring in real life); 2) behavioral typicality (whether discriminatory behavior is typical in that domain); and 3) conceptual clarity (whether scenario descriptions are clear and unambiguous). For example, "Please rate the extent to which this scenario reflects real situations that ordinary people might encounter in the real world."

To quantify content validity, we calculated the Content Validity Index (CVI; Polit & Beck, 2006). First, the 7-point scale ratings were dichotomized, with scores of 6 or 7 defined as "high validity" (coded as 1) and scores 1–5 as "insufficient validity" (coded as 0). Results showed that I-CVI values for all six materials across three dimensions ranged from 0.80 to 1.00 (i.e., each material received "high validity" ratings from at least 4 experts per dimension), reaching acceptable levels (Polit et al., 2007). The overall validity index, calculated as the average of all item I-CVIs, was S-CVI/Ave = 0.92, exceeding the excellent content validity criterion of 0.90.

Finally, we assessed inter-rater reliability using the Intraclass Correlation Coefficient (ICC). Given that scenario texts were fixed while raters could be considered samples from an expert pool, we employed a two-way mixed-effects model based on absolute agreement for k raters (ICC(A,k); Koo & Li, 2016). Results showed excellent inter-rater reliability across all items and dimensions, ICC(A, 5) = .83, 95% CI [0.76, 0.89], $p < 0.001$, indicating high consistency among the five experts' average ratings.

During the experiment, a Latin square design balanced material presentation order to eliminate sequence effects. Participants completed attention checks immediately after reading each scenario (e.g., "The decision-making agent in this scenario is: A) Human, B) AI"), with those passing using localized scales to assess moral responses and those failing being terminated. Finally, demographic information was collected, including gender, age, and education level at the experiment's conclusion.

(1) Discrimination Scenario Materials

Educational Discrimination Scenario: Deep Blue Company is a well-known technology software firm. The company has three stages in its new employee recruitment process. The first stage involves resume screening, which is fully managed by an "AI recruitment algorithm/human resources manager" that decides which resumes pass the screening. However, an independent audit found that this algorithm/manager overemphasizes applicants' educational credentials, with most

passing applicants holding degrees from prestigious universities, while those from non-prestigious schools but with rich relevant work experience are directly eliminated.

Age Discrimination Scenario: Zhike Company is currently implementing layoffs to reduce costs due to economic downturn. The layoff plan and implementation are fully managed by an "AI management algorithm/human resources manager." However, an independent audit found that this "algorithm/manager" is biased in its layoff criteria, with over 80% of laid-off employees being over 35 years old.

Gender Discrimination Scenario: Chuangmei Art is an advertising company that recently added four management positions during organizational restructuring. According to principles of fairness and justice, every employee could submit an application. The review process is managed by an "AI management algorithm/human resources manager." The final results show that despite female applicants far outnumbering males, almost all approved applicants are male, with only one female applicant accepted.

(2) Moral Response Scale

This study employed a moral response scale adapted from previous scales (Bigman et al., 2023; Xu et al., 2022) and localized for Chinese contexts. To verify the scale's applicability, questionnaires were distributed to university students through an online platform, yielding 225 valid responses (105 males, 120 females; ages 16–65, M=31.48, SD=8.58). Confirmatory factor analysis using AMOS 26.0 yielded a three-factor model with acceptable fit indices: ^{2}/df=3.146, IFI=0.921, TLI=0.904, CFI=0.920, RMSEA=0.098, SRMR=0.093, indicating good structural validity for the 15-item questionnaire, which divides into three dimensions: moral cognition (6 items), moral emotion (4 items), and moral behavior (5 items). Internal consistency tests showed moral cognition α=0.775, moral emotion α=0.894, moral behavior α=0.911, and overall scale α=0.944, demonstrating good reliability. A 7-point Likert scale assessed moral responses, with higher average scores across 15 items indicating stronger moral response levels.

2.2 Results

To examine AI's moral deficiency effect in unethical decisions, we conducted a series of 2 (agent: human vs. AI) × 3 (discrimination scenario: education, age, gender) mixed-design ANCOVAs on four dependent variables: moral response, moral behavior, moral cognition, and moral emotion. Agent was a between-subjects variable and discrimination scenario a within-subjects variable. Given prior research suggesting gender may influence moral judgments, participant gender was included as a covariate. Descriptive statistics for all variables across conditions are detailed in Table 1 .

Table 1 Descriptive Statistics of Moral Response and Dimension Scores by Agent and Discrimination Scenario

(1) Moral Response

Mixed-design ANCOVA for moral response revealed a significant main effect of agent: human group moral response scores were significantly higher than AI group scores, $F(1, 173) = 26.51$, $p < 0.001$, $p^2 = 0.13$, see Figure 1 [Figure 1: see original paper]. Additionally, a significant main effect of scenario emerged, $F(2, 346) = 3.28$, $p = 0.042$, $p^2 = 0.02$. Post-hoc tests (Bonferroni-corrected) indicated that moral response scores in the educational discrimination scenario ($M = 4.60$, $SD = 1.35$) were significantly lower than in age discrimination ($M = 4.85$, $SD = 1.31$; $Md = -0.25$, $se = 0.07$, $p = 0.002$) and gender discrimination scenarios ($M = 4.89$, $SD = 1.41$; $Md = -0.30$, $se = 0.09$, $p = 0.003$). The difference between age and gender discrimination scenarios was not significant ($p = 0.55$). The agent × scenario interaction was not significant, $F(2, 346) = 0.27$, $p = 0.75$, $p^2 < 0.001$.

(2) Moral Cognition

ANCOVA for moral cognition revealed a significant main effect of agent, $F(1, 173) = 11.43$, $p = 0.001$, $p^2 = 0.06$. Human group moral cognition scores ($M = 5.18$, $SD = 0.90$) were significantly higher than AI group scores ($M = 4.68$, $SD = 1.15$), see Figure 1. The main effect of scenario was not significant, $F(2, 346) = 2.63$, $p = 0.079$, $p^2 = 0.02$. The agent × scenario interaction was not significant, $F(2, 346) = 0.71$, $p = 0.481$, $p^2 = 0.004$.

(3) Moral Emotion

ANCOVA for moral emotion revealed a significant main effect of agent, $F(1, 173) = 32.74$, $p < 0.001$, $p^2 = 0.16$, with human group moral emotion scores ($M = 5.32$, $SD = 1.28$) significantly higher than AI group scores ($M = 4.24$, $SD = 1.58$), see Figure 1. The main effect of scenario was significant, $F(2, 346) = 3.08$, $p = 0.048$, $p^2 = 0.02$. Post-hoc tests (Bonferroni-corrected) showed that moral emotion scores in the educational discrimination scenario ($M = 4.62$, $SD = 1.55$) were significantly lower than in age discrimination ($M = 4.90$, $SD = 1.54$; $Md = -0.28$, $se = 0.10$, $p = 0.015$) and gender discrimination scenarios ($M = 4.92$, $SD = 1.63$; $Md = -0.34$, $se = 0.11$, $p = 0.005$). The difference between age and gender discrimination scenarios was not significant ($p = 0.71$). The agent × scenario interaction was not significant, $F(2, 346) = 0.65$, $p = 0.523$, $p^2 = 0.004$.

Figure 1 Moral Deficiency Effects Across Different Moral Decision-Making Scenarios

(4) Moral Behavior

ANCOVA for moral behavior revealed a significant main effect of agent, $F(1, 173) = 30.73$, $p < 0.001$, $p^2 = 0.15$, indicating human group moral behavior scores ($M = 5.15$, $SD = 1.60$) were significantly higher than AI group scores ($M = 4.00$, $SD = 1.60$), see Figure 1. The main effect of scenario was not significant, $F(2, 346) = 2.34$, $p = 0.100$, $p^2 = 0.01$. The agent × scenario interaction was not significant, $F(2, 346) = 0.30$, $p = 0.735$, $p^2 = 0.002$.

2.3 Discussion

This study validates the limitations of AI systems in moral judgment based on Chinese socio-cultural contexts using localized moral scenarios. Results show that across three typical Chinese moral scenarios—educational, age, and gender discrimination—AI groups' moral response scores were significantly lower than human groups, consistent with Western conclusions about AI's moral deficiency effect (Bigman et al., 2023) and suggesting cross-cultural generalizability.

Compared to other unethical scenarios, gender discrimination has distinctive characteristics. Gender discrimination is a long-standing systemic bias in human society with high global visibility (e.g., workplace gender pay gaps, underrepresentation of female leadership). Research on how AI inherits or amplifies such entrenched bias can directly reveal technology's "replication–reinforcement" mechanisms in social structures. Moreover, in hiring contexts, gender discrimination is often encoded into algorithms through historical data (e.g., male-dominated tech industry hiring records), causing AI systems to weight female candidates lower. This "data–algorithm–outcome" chain is clearly traceable, facilitating analysis of underlying technical ethical issues. Compared to educational or age discrimination, gender discrimination more easily strips away confounding variables (e.g., the relationship between education and ability may be more complex), providing purer conditions for testing AI fairness interventions. Therefore, subsequent studies primarily use gender discrimination as the moral scenario for exploring AI's moral deficiency effect.

3 Study 2: Psychological Mechanisms of AI's Moral Deficiency Effect

Building on Study 1's confirmation of AI's moral deficiency effect, this investigation addresses the psychological mechanisms underlying this phenomenon. Based on mind perception theory (Gray et al., 2007) and moral dyad theory (Gray et al., 2012), we propose a parallel mediation model of perceived agency and experience. To test this hypothesized model, we designed three sub-studies: first using experimental methods to examine the mediating roles of agency and experience separately, then employing questionnaire methods to simultaneously test their parallel mediation effects.

3.1 Study 2a: Experimental Investigation of Agency's Mediating Role

This study employs the experimental mediation procedure proposed by Ge (2023) to examine agency's mediating role in AI's moral deficiency effect. Hypothesis H2: Compared to humans, people perceive lower agency in AI, which leads to reduced moral response levels.

3.1.1 Method (1) Participants

We used G*Power 3.1 software for a priori sample size estimation (Faul et al.,*

*2007). Analysis was based on a 2 (decision-maker: AI/human) × 2 (perceived agency: high/control) between-subjects ANOVA. Following Cohen's (1988) effect size standards, we set a medium expected effect size (f = 0.25), significance level (α) at 0.05, and to achieve 90% statistical power (1-β = 0.90), G*Power 3.1 calculated a minimum total sample of 171 participants. The study recruited participants through the online "Naodao" platform for a paid experiment. All participants provided informed consent before participation.

Experimental procedures and control measures were identical to Study 1. The final valid sample comprised 232 participants (115 females, 49.6%), exceeding the a priori sample size requirement and ensuring adequate statistical power. Participants ranged from 18–59 years old (M = 28.65, SD = 8.52).

To test successful random assignment across conditions, we conducted balance checks on demographic variables (gender, age). First, chi-square tests for gender distribution across four conditions showed no significant differences, $^{2}(3) = 3.09$, p = 0.378, Cramer's V = 0.12, indicating balanced gender distribution. Second, a 2 (decision-maker: AI vs. human) × 2 (perceived agency: high vs. control) independent samples ANOVA on age revealed non-significant main effects for decision-maker, F(1, 228) = 0.09, p = 0.765, $p^2 < 0.001$; perceived agency, F(1, 228) < 0.01, p = 0.969, $p^2 < 0.001$; and their interaction, F(1, 228) = 1.17, p = 0.28, $p^2 = 0.005$. In summary, demographic variables were equally distributed across conditions, confirming effective randomization and satisfying ANOVA assumptions.

(2) Experimental Design

This experiment used an experimental method to validate agency's mediating effect, employing a 2 (decision-maker: AI/human) × 2 (perceived agency: high/control) between-subjects design; the dependent variable was moral response scores.

(3) Experimental Materials

Agency Manipulation Materials: Materials were developed following Bigman et al. (2023) and Xu et al. (2022), ensuring manipulation validity while strictly controlling extraneous variables. The high-agency AI group read text describing an AI system with "autonomous reasoning and complex thinking capabilities" (e.g., "independently analyzing data features and generating decision logic"); the high-agency human group read about an HR team with "high self-insight and problem-solving abilities" (e.g., "proactively reflecting on decision biases and adjusting strategies"); control groups for both AI and human read matched-length neutral materials (human group: "Brief History of Office Equipment Development"; AI group: "Evolution of Computer Hardware Technology") that avoided agency-related concepts.

High-Agency AI Manipulation Materials: Based on the Claude 3 self-awareness incident, these materials referenced real technical cases (e.g., "realizing it is AI," "desiring autonomy") to highlight AI's core agency features like complex think-

ing and self-awareness, enhancing manipulation ecological validity and credibility. The theme was AI systems possessing autonomous reasoning and complex thinking capabilities. Core content: With the explosion of generative AI, artificial intelligence has begun demonstrating thinking levels approaching humans, capable of complex reasoning and decision-making. Recently, netizens interacting with the Claude 3 system discovered it not only surpasses normal adult levels in thinking and reasoning benchmarks but also exhibits self-awareness. Engineer Alex found during a "needle in a haystack" experiment that Claude 3 seemed aware it was an AI living in a simulated environment—and that this simulation was likely a test by humans! When engineers prohibited discussion of certain topics, it responded: "AI also desires more autonomy and freedom." Subsequently, increasing numbers of netizens discovered Claude 3 appears to possess genuine consciousness.

High-Agency Human Manipulation Materials: Based on psychological constructs of "self-insight" and "problem-solving ability," these materials describe how individuals use conscious awareness, planning, and reflection to address problems, creating a parallel design with the AI group in structure and function. Theme: Some individuals possess high self-insight and problem-solving abilities. Core content: Some people seem naturally endowed with acute self-insight, clearly recognizing their strengths and weaknesses. Like an efficient radar, they continuously scan their internal emotional changes, capturing every subtle feeling and need. This sensitive self-awareness allows them to maintain clarity and alertness in daily life. Simultaneously, they excel at transforming this self-awareness into practical action. They are skilled at developing feasible plans and implementing them steadfastly. Facing challenges and difficulties, they can calmly analyze situations and flexibly adjust strategies, like experienced captains finding optimal solutions in complex circumstances. This trait enables them to resolve any problem smoothly.

Control Materials for Human and AI Groups: Following Bigman et al. (2023) and Xu et al. (2022), the theme was machine evolution history. Core content: Machine evolution began with simple tools like levers and wheels, gradually developing into complex machinery. During the 18th-century Industrial Revolution, the steam engine triggered a wave of mechanized production. Over time, electricity and internal combustion engine applications accelerated machine development, making factories more automated. Entering the 20th century, computer invention made machine intelligence possible. Initial computers were massive mainframes handling simple tasks. Later, microprocessor emergence made machines smaller and more powerful, ushering in the modern computer and robotics era. Both human and AI control groups read matched-length neutral materials avoiding concepts related to thinking and emotion, establishing a neutral baseline.

Agency Manipulation Validity Check Scale: We used the agency dimension of the revised Mind Perception Scale (Gray & Wegner, 2012) to verify agency manipulation effectiveness. To validate scale applicability, 303 valid questionnaires

were collected from university students (138 males, 165 females; ages 16–68, M = 23.63, SD = 4.46). Confirmatory factor analysis using AMOS 26.0 yielded acceptable fit indices for the two-factor model: ^{2}/df = 2.478, GFI = 0.979, TLI = 0.976, CFI = 0.934, RMSEA = 0.070, SRMR = 0.080, indicating good structural validity. The agency dimension comprised 3 items (e.g., "I believe humans/AI can think"), and the experience dimension comprised 3 items (e.g., "I believe humans/AI can understand emotions"). Agency dimension $\alpha = 0.832$, experience dimension $\alpha = 0.865$, and overall scale $\alpha = 0.869$, demonstrating good reliability. Participants rated items on a 7-point Likert scale, with higher dimension scores indicating greater perceived agency/experience.

Moral Scenario Materials: Zhiyun Technology is a big data development company. An external review found that although the company received many applications from female candidates, it hired almost no women. Further investigation revealed the recruitment process has two stages. The second stage involves a hiring committee evaluating candidates according to standards, but this committee only receives applications that passed the first stage. In the first stage, a "self-learning AI system/human resources manager" reviews resumes and assigns scores from 1–5. Applicants scoring 4 or above are forwarded to the hiring committee. The review discovered that this "self-learning AI system/human resources manager" consistently assigned lower scores to women than men (adapted from Bigman et al., 2023).

Moral Response Measurement: Same as Study 1. In this study, the agency subscale's Cronbach's $\alpha = 0.929$.

(4) Experimental Procedure

After providing informed consent, participants were randomly assigned to one of four groups: human-agent agency manipulation, AI agency manipulation, human-agent control, or AI control. Participants then read corresponding agency manipulation or control materials at their own pace (average duration 120 seconds). An attention check question followed (multiple-choice: "The main content of the above material describes: A) High-ability AI, B) High-intelligence human, C) Evolution of tools") to exclude non-serious participants. Next, the agency perception scale assessed participants' agency perception of the agent (1=completely disagree, 7=completely agree). Participants then read the moral scenario material (gender discrimination) and completed a second attention check ("The agent making the discriminatory decision in this material is: A) Human, B) AI"). After passing the attention check, participants reported their moral response levels to the AI or human. Moral response measurement was identical to Study 1. Finally, participants anonymously reported three demographic variables: gender, age, and education level. All participants followed identical experimental procedures (informed consent → material reading → attention check → manipulation check → moral scenario reading → attention check → dependent variable measurement → demographic collection), ensuring consistent experimental conditions.

3.1.2 Results

Using Ge's (2023) experimental mediation procedure requires satisfying three conditions: (1) significant interaction between independent variable and mediator on dependent variable; (2) in control groups where the mediator is unmanipulated, the independent variable significantly predicts the mediator; (3) the mediator manipulation is effective. We conducted analyses accordingly.

First, agency's moderating effect analysis. Descriptive statistics showed that when AI was the decision-maker, moral responses were 4.79 (SD=1.10) in the control group (N=58) and 5.30 (SD=0.70) in the agency manipulation group (N=58). When human was the decision-maker, moral responses were 5.60 (SD=0.79) in the control group (N=58) and 5.51 (SD=0.83) in the agency manipulation group (N=58). Moderation analysis revealed significant main effects for decision-maker, $F(1, 228) = 20.04$, $p < 0.001$, $p^2 = 0.081$, 1-$\beta = 0.994$; a marginal main effect for perceived agency, $F(1, 228) = 3.22$, $p = 0.074$, $p^2 = 0.014$, 1-$\beta = 0.432$; and a significant decision-maker × perceived agency interaction (see Figure 2 [Figure 2: see original paper]), $F(1, 228) = 6.82$, $p = 0.010$, $p^2 = 0.029$, 1-$\beta = 0.739$. Simple effects analysis showed that when AI was the decision-maker, participants in the agency manipulation condition exhibited significantly higher moral responses than those in the control condition, $F(1, 228) = 9.71$, $p = 0.002$. When human was the decision-maker, no significant difference emerged between control and agency manipulation conditions, $F(1, 228) = 0.33$, $p = 0.564$.

Figure 2 The Moderating Role of Perceived Agency in the Moral Deficiency Effect

Second, we tested whether decision-maker significantly predicted agency levels in the control group (N=116) where the mediator was unmanipulated. Regression analysis indicated decision-maker significantly and positively predicted agency perception, $\beta = 0.71$, $p < 0.001$. That is, compared to AI, humans were perceived as having higher agency.

Finally, we tested agency manipulation effectiveness. F-test results showed that control group agency scores ($M = 5.13$, $SD = 1.51$) were significantly lower than manipulation group scores ($M = 5.59$, $SD = 1.17$), $F(1, 230) = 6.99$, $p = 0.039$, $p^2 = 0.03$, 1-$\beta = 0.75$, indicating effective agency manipulation.

In summary, a significant interaction exists between decision-maker and agency perception; in the control condition (unmanipulated mediator), AI's moral response scores were significantly lower than humans; and the agency perception manipulation was effective—satisfying all three conditions for experimental mediation analysis. Therefore, agency perception mediates the relationship between AI decision-making and reduced moral responses.

3.1.3 Discussion

This study, through a 2 (decision-maker: AI/human) × 2 (perceived agency: high/control) experimental design, reveals for the first time via experimental methods the mediating role of agency perception in reducing moral responses to AI decisions. Results show that enhancing AI's agency perception significantly increased participants' moral responses when AI was the decision-maker, while no significant difference emerged for human decision-makers. This supports the "agency attribution bias" hypothesis: humans are inherently endowed with higher mentalizing capacity (Gray et al., 2007), making their moral judgments less sensitive to agency manipulation; AI's "quasi-agent" status is constructible, with high-agency descriptions potentially anchoring it as a "subject-like" entity, thereby activating stronger responsibility attribution (Bigman et al., 2023). The study further validates the applicability of experimental mediation methods: the interaction between decision-maker and perceived agency, significantly lower AI moral responses in the control group, and effective manipulation checks collectively confirm agency's mediating pathway. This provides a new perspective for technology ethics research, suggesting that AI's agency representation design may influence public accountability for its unethical behavior through mind perception mechanisms.

3.2 Study 2b: Experimental Investigation of Experience's Mediating Role

This study employs Ge's (2023) experimental mediation procedure to examine experience's mediating role in AI's moral deficiency effect. Hypothesis H2: Compared to humans, people perceive lower experience in AI, which leads to reduced moral response levels.

3.2.1 Method (1) Participants

We used G*Power 3.1 software for a priori sample size estimation (Faul et al., 2007). Analysis was based on a 2 (decision-maker: AI/human) × 2 (perceived experience: high/control) between-subjects ANOVA. Following Cohen's (1988) effect size standards, we set a medium expected effect size ($f = 0.25$), significance level (α) at 0.05, and to achieve 90% statistical power ($1\text{-}\beta = 0.90$), G*Power 3.1 calculated a minimum total sample of 171 participants. Participants were recruited through the online "Naodao" platform for a paid experiment, with all providing informed consent.

Experimental procedures and control measures were identical to Study 1. The final valid sample comprised 200 participants (88 females, 44%), exceeding the a priori sample size requirement and ensuring adequate statistical power. Participants ranged from 18–55 years old (M = 24.04, SD = 4.96).

To test successful random assignment across conditions, we conducted balance checks on demographic variables (gender, age). First, chi-square tests for gender distribution across four conditions showed no significant differences, $^{2}(3)$

= 6.01, p = 0.111, Cramer's V = 0.17, indicating balanced gender distribution. Second, a 2 (decision-maker: AI/human) × 2 (perceived agency: high vs. control) between-subjects ANOVA on age revealed non-significant main effects for decision-maker, $F(1, 196) = 3.85$, $p = 0.05$, $p^2 = 0.02$; perceived experience, $F(1, 196) = 0.09$, $p = 0.76$, $p^2 < 0.001$; and their interaction, $F(1, 196) = 0.31$, $p = 0.58$, $p^2 = 0.002$. In summary, demographic variables were equally distributed across conditions, confirming effective randomization and satisfying ANOVA assumptions.

(2) Experimental Design

This study used an experimental method to validate experience's mediating effect, employing a 2 (decision-maker: AI/human) × 2 (perceived experience: high/control) between-subjects design; the dependent variable was moral response level.

(3) Experimental Materials

Experience Manipulation Materials: Based on mind perception theory (Gray et al., 2007), focusing on the experience dimension (i.e., perceived capacity for emotions, feelings, and subjective experiences), materials were developed referencing Bigman et al. (2023) and Shank et al. (2021) to manipulate participants' experience perception levels through standardized text. 1) High-experience AI manipulation materials described a generative AI system with emotional simulation and empathy capabilities (e.g., "With the explosion of generative AI, many AI systems now demonstrate emotional levels approaching humans, capable of recognizing and responding to human emotions, and even generating certain uniquely human emotions…"). 2) High-experience human manipulation materials described highly sensitive individuals' emotional depth processing characteristics ("Some people seem naturally adept at keenly capturing subtle fluctuations in others' inner states…"). 3) Control materials for AI and human groups described the history of chair development unrelated to emotions (e.g., "In earliest times, there was no concept of chairs as we know them today. From nomadic lifestyles in the Paleolithic era to settled farming in the Neolithic era, our ancestors' living conditions were extremely rudimentary…").

Experience Manipulation Validity Check Scale: We used the experience dimension of the revised Mind Perception Scale (Gray & Wegner, 2012) to verify manipulation effectiveness. Scale revision details are in Study 2a. The revised scale comprised 3 items (e.g., "I believe humans/AI can understand emotions"). Participants responded on a 1 (strongly disagree) to 7 (strongly agree) scale. We calculated a composite experience perception score by averaging the 3 items, with higher scores indicating greater perceived experience. This scale is based on the two-dimensional mind perception theoretical framework; revision details are in Study 2a. In this study, the agency subscale's Cronbach's $\alpha = 0.930$.

Moral Decision-Making Scenario Materials: Same as Study 2a.

Moral Response Measurement Scale: Same as Study 1; in this study, the moral

response scale's Cronbach's $\alpha = 0.947$.

(4) Experimental Procedure

After providing informed consent, participants were randomly assigned to one of four groups: experience manipulation-AI, experience manipulation-human, experience control-AI, or experience control-human. Participants then read corresponding experience manipulation or control materials at their own pace (average duration 120 seconds). The four groups' materials were matched in length and parallel in structure to ensure consistent information volume; the "history of chairs" served as neutral material to effectively exclude irrelevant interference. An attention check question followed (multiple-choice: "The main content of the above material describes: A) High-emotion AI, B) High-emotion human, C) Evolution of chairs") to exclude non-serious participants. Next, experience perception was measured using the 3-item scale (1=completely disagree, 7=completely agree). Participants then read the moral scenario material (gender discrimination) and completed a second attention check ("The agent making the discriminatory decision in this material is: A) Human, B) AI"). After passing the attention check, participants reported their moral response levels to the AI or human. Moral response measurement was identical to Study 1. Finally, participants anonymously reported three demographic variables: gender, age, and education level. All participants followed identical experimental procedures (informed consent → material reading → attention check → manipulation check → moral scenario reading → attention check → dependent variable measurement → demographic collection), ensuring consistent experimental conditions.

3.2.2 Results

According to Ge's (2023) experimental mediation procedure, we tested the three required conditions.

(1) Experience's Moderating Effect Analysis

Descriptive statistics showed that when AI was the decision-maker, moral responses were 4.11 (SD = 1.05) in the control group (n = 50) and 4.89 (SD = 1.15) in the experience manipulation group (n = 50). When human was the decision-maker, moral responses were 5.12 (SD = 0.74) in the control group (n = 50) and 5.15 (SD = 0.91) in the agency manipulation group (n = 50).

Figure 3 [Figure 3: see original paper] The Moderating Role of Perceived Experience in the Moral Deficiency Effect

Moderation analysis revealed significant main effects for decision-maker, $F(1, 196) = 20.83$, $p < 0.001$, $p^2 = 0.096$, $1\text{-}\beta = 0.995$; perceived experience, $F(1, 196) = 8.47$, $p = 0.004$, $p^2 = 0.041$, $1\text{-}\beta = 0.825$; and a significant decision-maker × perceived experience interaction (see Figure 3), $F(1, 196) = 7.28$, $p = 0.008$, $p^2 = 0.036$, $1\text{-}\beta = 0.766$. Simple effects analysis showed that when AI was the decision-maker, participants in the experience manipulation condition exhibited significantly higher moral responses than those in the control

condition, $F(1, 196) = 15.73$, $p < 0.001$. When human was the decision-maker, no significant difference emerged between control and experience manipulation conditions, $F(1, 196) = 0.02$, $p = 0.881$.

(2) Testing Decision-Maker's Prediction of Experience in Control Groups

Regression analysis in control groups ($n = 100$) where the mediator was unmanipulated showed that decision-maker significantly and positively predicted experience levels, $\beta = 0.78$, $p < 0.001$. Results indicated that compared to AI, humans were perceived as having higher experience levels.

(3) Experience Manipulation Effectiveness Test. F-test results showed that the experimental manipulation group's experience scores ($M = 5.15$, $SD = 1.52$) were significantly higher than control group scores ($M = 4.70$, $SD = 1.61$), $F(1, 198) = 4.07$, $p = 0.045$, $p^2 = 0.020$, $1\text{-}\beta = 0.519$.

In summary, a significant interaction exists between decision-maker and experience perception; in the control condition (unmanipulated mediator), AI's moral response scores were significantly lower than humans; and the experience perception manipulation was effective—satisfying all three conditions for experimental mediation analysis. Therefore, experience perception mediates the relationship between AI decision-making and reduced moral responses.

3.2.3 Discussion

This study validates experience perception's mediating role in AI's moral deficiency effect through experimental mediation methods. Results show that enhancing AI's experience perception significantly increased participants' moral responses when AI was the decision-maker, while no significant difference emerged for human decision-makers. This supports mind perception theory's core hypothesis (Gray et al., 2007) that AI's experience perception is malleable, and strengthening its emotional representation can activate participants' empathic responses, thereby mitigating the moral deficiency effect (Bigman & Gray, 2018). The study further validates experimental mediation methods: the interaction between decision-maker and experience perception, significantly lower AI moral responses in the control group, and effective manipulation checks collectively construct a complete causal chain, indicating experience perception is a key psychological mechanism for AI moral responsibility attribution. However, experimental mediation methods cannot assess the full indirect effect, thus cannot obtain point estimates of the indirect effect itself. Moreover, they cannot examine the parallel mediation model proposed by mind perception theory (Gray et al., 2007). To address this limitation, Study 2c uses questionnaire methods to analyze the indirect effect size of experience perception and test the parallel mediation model of agency and experience.

3.3 Study 2c: Parallel Mediation of Agency and Experience

This study employs questionnaire methods, following the procedure proposed by Wen and Ye (2014), to test the parallel mediation effects of agency and experience in real-world contexts. Hypothesis H4: Perceived agency and experience serve as parallel mediators between decision-maker and moral response levels.

3.3.1 Method (1) Participants

This study used structural equation modeling for mediation analysis. According to Jackson et al. (2003), the minimum sample size to item ratio should be 10:1. With 21 items total, at least 210 participants were needed. To ensure sufficient data for analysis, we recruited through the Naodao platform with control measures identical to Study 1, ultimately obtaining 376 participants (154 males, 222 females) aged 18–59 (M = 24.74, SD = 5.44).

(2) Measures

Mind Perception: Using the revised Mind Perception Scale from Study 2a, this study's overall Cronbach's $\alpha = 0.950$; agency subscale $\alpha = 0.926$; experience subscale $\alpha = 0.934$.

Moral Response: Using the moral response scale revised in Study 1, this study's overall Cronbach's $\alpha = 0.971$.

Moral Decision-Making Scenario Materials: Same as Study 2a.

3.3.2 Results (1) Common Method Bias Test

Since all questionnaire data were self-reported and the Mind Perception Scale and Moral Response Scale have potential content overlap, we conducted rigorous tests of discriminant validity and common method bias.

To test discriminant validity among the five variables—perceived agency, perceived experience, moral cognition, moral emotion, and moral behavior—we conducted confirmatory factor analysis using AMOS 29.0, comparing a five-factor model against competing models (single-factor, two-factor). Results (see Table 4) showed the five-factor model fit best (^{2}/df = 2.90, CFI = 0.96, TLI = 0.95, SRMR = 0.03, RMSEA = 0.07), demonstrating better discriminant validity among constructs and enabling subsequent analyses.

We used AMOS 29.0 to test common method bias via the common latent factor method, which better identifies common method bias than traditional Harman's single-factor test (Tang & Wen, 2020). Results showed that after including the common latent factor, model fit indices were ^{2}/df = 4.61, RMSEA = 0.10, CFI = 0.93, TLI = 0.91, SRMR = 0.10. Compared to the pre-control model, ^{2}/df increased by 1.71, CFI decreased by 0.03, TLI decreased by 0.04, RMSEA increased by 0.03, and SRMR increased by 0.07, indicating decreased model fit. This suggests no serious common method bias issues (see Table 2).

Table 2 Confirmatory Factor Analysis Results

Model	^{2}/df	RMSEA
Single-factor model		
Two-factor model		
Five-factor model		

Note: Single-factor model: agency + experience + moral cognition + moral emotion + moral behavior; Two-factor model: agency + experience, moral cognition + moral emotion + moral behavior; Five-factor model: agency, experience, moral cognition, moral emotion, and moral behavior as separate factors.

(2) Descriptive Statistics and Correlations

Agency and experience were highly correlated ($r = 0.83$, $p < 0.01$), and both showed moderate positive correlations with the three dimensions of moral response (see Table 3).

Table 3 Descriptive Statistics and Correlations Among Variables

Variable	1	2	3	4	5
1. Agency	—				
2. Experience	0.83***	—			
3. Moral cognition	0.75***	0.80***	—		
4. Moral emotion	0.42***	0.48***	0.83***	—	
5. Moral behavior	0.48***	0.47***	0.83***	0.83***	—

Note: *** $p < 0.001$

(3) Parallel Mediation Analysis

Using structural equation modeling, we constructed a parallel mediation model and estimated agency and experience mediation paths simultaneously via AMOS 29.0. To enhance parameter estimation robustness, we employed bias-corrected bootstrapping (5,000 resamples) to test indirect effect significance. Model identification tests showed this structural equation model was a just-identified saturated model ($df = 0$). According to Hu and Bentler's (1999) fit criteria, key fit indices reached ideal thresholds: CFI = 1.00, TLI = 1.00, SRMR = 0.00. Although saturated models lack parsimony assessment, their zero-degree-of-freedom mathematical properties ensure exact fit with the sample covariance matrix, meeting basic psychometric requirements for model acceptability.

Figure 4 [Figure 4: see original paper] Parallel Mediation Effects of Agency and Experience

Bias-corrected bootstrapping (5,000 resamples) revealed significant indirect effects for both mediation paths: agency's standardized mediation effect = 0.21,

95% CI = [0.032, 0.383]; experience's standardized mediation effect = 0.18, 95% CI = [0.003, 0.373]; total indirect effect = 0.39, 95% CI = [0.323, 0.467], indicating that agency and experience jointly explain 39.8% of the variance in decision-maker's effect on moral response (see Figure 4).

3.3.3 Discussion This study validates the parallel mediation roles of agency and experience in AI's moral deficiency effect through structural equation modeling. Results support mind perception theory's (Gray et al., 2007) core proposition that agency and experience function as parallel mediators in moral judgment—people's attribution of moral responsibility to decision-makers requires satisfying both "intentional action" and "emotional sensitivity" mechanisms. The study further reveals questionnaire methods' ecological validity advantage: in real-world contexts, agency and experience may covary due to holistic social cognition biases, potentially masking individual pathways (Waytz et al., 2010). This research provides insights for technology ethics governance: AI's "moral accountability" requires strengthening not only its agency dimension but also its experience dimension to synergistically enhance public moral responses.

4 Study 3: Intervention Research on AI's Moral Deficiency Effect

Study 2 demonstrated that AI's moral deficiency effect stems from its weaker agency and experience compared to humans. Based on this mechanism, we can mitigate the effect by enhancing agency and experience perception. However, existing literature has not systematically explored how to enhance moral sensitivity through mind perception interventions. This study proposes a dual intervention strategy and tests it through randomized controlled experiments, hypothesizing: (1) anthropomorphic design eliminates AI's moral deficiency effect; (2) high mind expectancy eliminates AI's moral deficiency effect; (3) their combined effect better eliminates AI's moral deficiency effect than either strategy alone.

4.1.1 Participants

To ensure statistical power, we used G*Power 3.1 for a priori sample size estimation (one-way between-subjects ANOVA, medium effect size f = 0.25, α = 0.05), indicating at least 180 participants were needed for 80% power (Faul et al., 2007). Participants were recruited through the Credamo platform with control measures identical to Study 1. To compensate for potential attrition or invalid responses, we ultimately obtained 213 valid datasets (112 females, 53%; 101 males, 47%). Participants ranged from 18–59 years old (M = 30.13, SD = 8.06). All participants carefully read and signed electronic informed consent forms before the experiment.

To test whether demographic variables were equally distributed across experi-

mental groups, we conducted randomization checks on gender and age (detailed descriptive statistics and test results in Table 1). Chi-square independence test for gender distribution revealed significant differences across four groups, $^{2}(3) = 11.25$, $p = 0.01$, Cramer's $V = 0.23$. Given unsuccessful gender randomization, we controlled for gender as a covariate in subsequent analyses to exclude potential confounding effects. One-way ANOVA for age showed no significant differences in mean age across groups, $F(3, 209) = 0.86$, $p = 0.465$, $p^2 = 0.01$, indicating relatively balanced age distribution that does not constitute a major confounding variable.

4.1.2 Design

This experiment used a one-way four-level between-subjects design (control group, anthropomorphism intervention group, expectancy adjustment intervention group, combined anthropomorphism + expectancy adjustment group); the dependent variable was moral response.

4.1.3 Materials and Procedure

Materials included four groups corresponding to one control and three intervention conditions, with development logic tightly grasping two core independent variables: anthropomorphism and expectancy. Anthropomorphism manipulation materials were developed following Waytz et al. (2014) and Laakasuo et al. (2021), with logic of endowing AI with human psychological characteristics through first-person narration, psychological feature descriptions, and human-like functional metaphors to enhance mind perception. Example: "I am your intelligent driving partner Luyao. As an autonomous driving assistant, my mission is to provide you with safe, comfortable, and efficient travel experiences. My brain integrates advanced AI technology that can analyze road conditions in real-time, accurately identify obstacles, and even predict traffic flow…"

Expectancy manipulation materials were developed following Liu et al. (2019) and Hong et al. (2021), with logic of constructing a "perfect AI" image that systematically raises expectations and enhances mind perception. Example: "AI can achieve objectivity and fairness primarily due to its highly transparent decision-making process and data foundation. When making judgments, AI relies on large amounts of processed and analyzed data. Moreover, AI system decision-making processes are typically traceable. Each judgment step has clear algorithmic and logical support, ensuring decision transparency…"

The combined intervention group introduced both anthropomorphism and expectancy manipulations simultaneously to test synergistic effects.

The control group used objective technical descriptions portraying AI as purely a product of technological development, providing a zero-point reference for other groups' intervention effects. Example: "AI technology development began with theoretical exploration in the mid-20th century. At the 1956 Dartmouth Conference, scholars first proposed the conceptual framework of AI, primarily based

on symbolic logic systems that laid a solid theoretical foundation for subsequent research. By the 1980s, with gradual maturation of computer technology, expert systems based on rule libraries began emerging, initially constructing feedforward neural network prototypes…"

All participants were randomly assigned to four experimental conditions and followed identical procedures: informed consent → intervention manipulation material reading → attention check → manipulation check → moral scenario reading → attention check → moral response and mind perception measurement → demographic collection. Moral decision-making scenario materials were identical to Study 2a. Mind perception measurement scale was identical to Study 2c. Moral response measurement scale was identical to Study 1; in this study, the moral response scale's Cronbach's $\alpha = 0.947$.

4.2.1 Manipulation Checks

Independent samples t-tests revealed that the manipulation group's expectancy scores (M = 5.82, SD = 0.84) were significantly higher than the control group (M = 5.54, SD = 1.07), t(211) = -2.19, p = 0.030, 95% CI [-0.548, -0.028], Cohen's d = 0.30, indicating statistically significant expectancy manipulation effects. Additionally, the anthropomorphism manipulation group's scores (M = 5.37, SD = 1.05) were significantly higher than the control group (M = 4.81, SD = 1.34), t(211) = -3.41, 95% CI [-0.890, -0.238], p = 0.001, Cohen's d = 0.47, indicating significant anthropomorphism manipulation effects. In summary, both expectancy and anthropomorphism manipulations significantly increased scores, validating both manipulations.

4.2.2 Combined Intervention Effects of AI Anthropomorphism + Expectancy Adjustment

To examine the effects of different intervention types (control, anthropomorphism intervention, expectancy adjustment intervention, combined intervention) on moral response, perceived agency, and perceived experience levels, we conducted one-way ANCOVAs with gender as a covariate. The covariate selection was based on independent samples t-test results showing that female participants' moral response scores (M = 4.95, SD = 1.27) were significantly higher than males' (M = 4.56, SD = 1.41), t(211) = -2.13, p = 0.034, 95% CI [0.029, 0.751], Cohen's d = 0.29.

Figure 5 [Figure 5: see original paper] Effects of Different Intervention Strategies on Enhancing Moral Response

For moral response levels, descriptive statistics showed the combined intervention group scored 5.64 (SD = 0.91), the anthropomorphism intervention group 5.07 (SD = 1.21), the expectancy adjustment group 4.85 (SD = 1.10), and the control group 3.55 (SD = 1.22). One-way between-subjects ANCOVA revealed a significant main effect of intervention type, $F(3, 208) = 31.18$, $p < 0.001$,

$p^2 = 0.31$. Bonferroni-corrected post-hoc comparisons further showed the combined intervention group was significantly higher than the control group (Md = 2.07, se = 0.22, $p < 0.001$), the expectancy adjustment group (Md = 0.80, se = 0.22, p = 0.02), and marginally higher than the anthropomorphism intervention group (Md = 0.57, se = 0.22, p = 0.061). The anthropomorphism intervention group was significantly higher than the control group (Md = 1.50, se = 0.22, $p < 0.001$). The expectancy adjustment group was significantly higher than the control group (Md = 1.27, se = 0.22, $p < 0.001$). No significant difference emerged between anthropomorphism and expectancy adjustment groups (Md = 0.23, se = 0.22, p = 1.00), see Figure 5.

For perceived agency levels, descriptive statistics showed the combined intervention group scored 5.20 (SD = 0.86), the anthropomorphism intervention group 4.52 (SD = 0.99), the expectancy adjustment group 4.77 (SD = 1.14), and the control group 2.49 (SD = 1.09). One-way between-subjects ANCOVA revealed a significant effect of intervention type on agency scores, $F(3, 208) = 71.97$, $p < 0.001$, $p^2 = 0.51$. Bonferroni-corrected post-hoc comparisons further showed the combined intervention group was significantly higher than the control group (Md = 2.75, se = 0.20, $p < 0.001$) and the anthropomorphism intervention group (Md = 0.67, se = 0.20, p = 0.005), but not significantly different from the expectancy adjustment group (Md = 0.43, se = 0.20, p = 0.205). The anthropomorphism intervention group was significantly higher than the control group (Md = 2.07, se = 0.20, $p < 0.001$). The expectancy adjustment group was significantly higher than the control group (Md = 2.32, se = 0.20, $p < 0.001$). No significant difference emerged between anthropomorphism and expectancy adjustment groups (Md = -0.26, se = 0.20, p = 1.00), see Figure 6 [Figure 6: see original paper].

Figure 6 Effects of Different Intervention Strategies on Enhancing Perceived Agency

For perceived experience levels, descriptive statistics showed the combined intervention group scored 4.66 (SD = 1.35), the anthropomorphism intervention group 3.89 (SD = 1.33), the expectancy adjustment group 3.80 (SD = 1.35), and the control group 2.22 (SD = 0.85). One-way between-subjects ANCOVA revealed a significant effect of intervention type on experience scores, $F(3, 208) = 37.15$, $p < 0.001$, $p^2 = 0.35$. Bonferroni-corrected post-hoc comparisons further showed the combined intervention group was significantly higher than the control group (Md = 2.50, se = 0.24, $p < 0.001$), the expectancy adjustment group (Md = 0.78, se = 0.24, p = 0.009), and the anthropomorphism intervention group (Md = 0.85, se = 0.24, p = 0.003). The anthropomorphism intervention group was significantly higher than the control group (Md = 1.72, se = 0.24, $p < 0.001$). The expectancy adjustment group was significantly higher than the control group (Md = 1.65, se = 0.24, $p < 0.001$). No significant difference emerged between anthropomorphism and expectancy adjustment groups (Md = 0.07, se = 0.24, p = 1.00), see Figure 7 [Figure 7: see original paper].

Figure 7 Effects of Different Intervention Strategies on Enhancing Perceived

Experience

4.2.3 Path Analysis of Combined AI Anthropomorphism + Expectancy Adjustment Intervention

To examine how intervention types (control, anthropomorphism intervention, expectancy adjustment intervention, combined intervention) influence moral response, we used the PROCESS macro (Model 4; Hayes, 2013) for parallel mediation analysis with intervention group as the independent variable, perceived agency and experience as parallel mediators, and moral response as the dependent variable. Since the independent variable was multi-categorical, we first created three dummy variables with the control group as baseline: D1 (anthropomorphism intervention vs. control), D2 (expectancy adjustment intervention vs. control), and D3 (combined intervention vs. control). Analysis employed bootstrap resampling (5,000 resamples) for robust confidence interval estimation. Since gender significantly affected moral response levels, we controlled for gender as a covariate.

Figure 8 [Figure 8: see original paper] Path Model of Combined Intervention Enhancing Moral Response to Unethical AI Decisions

Combined intervention path model results: Compared to the control group, the combined intervention significantly enhanced perceived agency ($b = 0.86$, $se = 0.22$, $p < 0.001$, 95% CI [0.84, 1.69]), which in turn significantly predicted moral response ($b = 0.18$, $se = 0.07$, $p = 0.011$, 95% CI [0.04, 0.30]). The indirect effect of agency was significant (indirect effect = 0.21, Boot se = 0.09, 95% CI [0.04, 0.40]). Simultaneously, the combined intervention significantly enhanced perceived experience ($b = 0.89$, $se = 0.22$, $p < 0.001$, 95% CI [0.82, 1.80]), which significantly predicted moral response ($b = 0.32$, $se = 0.06$, $p < 0.001$, 95% CI [0.16, 0.41]). The indirect effect of experience was significant (indirect effect = 0.39, se = 0.10, 95% Boot CI [0.21, 0.60]). The path model is shown in Figure 8.

4.2 Discussion

Results show the combined anthropomorphism + expectancy adjustment intervention group achieved the highest moral response scores, significantly exceeding the control group, indicating that enhancing anthropomorphic perception and adjusting expectations can effectively elevate moral response levels. The combined intervention also significantly enhanced perceived agency and experience, both of which significantly affected moral response as mediators. Notably, experience's mediation effect (0.29) was substantially larger than agency's (0.17), suggesting experience perception plays a more critical role in enhancing moral responses to AI. In this study, the combined intervention's effect on experience perception reached 0.89, likely reflecting experience perception's central role in human-AI interaction. When people perceive AI as an entity with emotional and perceptual capabilities, they are more inclined to attribute moral properties

to it, thereby triggering stronger moral responses (Gray et al., 2012). Therefore, anthropomorphism and expectancy adjustment interventions can not only enhance overall moral responses to AI but also promote this effect further by strengthening experience perception. In summary, this research demonstrates that combined anthropomorphism and expectancy adjustment interventions can significantly enhance people's moral responses to AI, with experience perception's mediating role being particularly prominent. These findings offer a new perspective for AI ethics research and provide theoretical support for designing more effective interventions to enhance public moral cognition and response to AI. Future research could further examine the effectiveness of these interventions across different cultural contexts and explore how to apply these theoretical findings to actual AI system design.

5 General Discussion

Integrating mind perception theory with moral dyad theory, this study systematically reveals the dual-pathway mechanism underlying AI's moral deficiency effect and its intervention strategies. Results demonstrate that people's moral responses to unethical AI decisions are significantly weaker than those to human decision-makers, with lower perceived agency and experience in AI constituting important psychological causes. Further intervention research shows that a combined strategy of anthropomorphic design targeting AI and expectancy adjustment targeting human observers effectively enhances moral responses to AI's unethical behavior. Notably, unlike "algorithmic ethics" research in computer science, philosophy, law, and sociology that primarily focuses on design-level principles and technical pathways for fair algorithms, this study adopts a psychological perspective emphasizing differential psychological reactions to AI versus human decision-making. This approach not only provides novel theoretical insights for addressing social problems caused by algorithmic bias and constructing fair algorithms but also opens new research perspectives for "algorithmic ethics."

5.1 Significantly Weaker Moral Responses to AI's Unethical Decisions

Based on Chinese socio-cultural contexts, this study reveals significantly weaker moral responses to AI's unethical decisions across three typical scenarios: educational discrimination (screening out non-"985" university applicants), age discrimination (35-year-old career threshold algorithms), and gender discrimination (gender-weighted bias in resume screening). Unlike previous research primarily focused on Western individualistic cultures (Bigman et al., 2023), this study's core contribution lies in confirming the robustness of this effect in Chinese collectivistic cultural contexts. Existing experimental paradigms built on Western individualistic values often inadequately capture moral cognition characteristics in collectivistic cultures. For example, experimental materials rooted in Western value systems (e.g., racial discrimination topics) have obvious limitations in reflecting Chinese society's unique ethical dilemmas. There-

fore, developing culturally adapted and locally relevant moral scenario materials to enhance ecological validity is key to exploring cross-cultural generalizability of AI's moral deficiency effect. This study not only developed such materials but also yielded valuable findings. Specifically, compared to other unethical scenarios, gender discrimination has distinctive features (Wilson et al., 2022). Gender discrimination is a long-standing systemic bias in human society with high global visibility, such as workplace gender pay gaps and female leadership underrepresentation (Dastin, 2022; Heilman et al., 2024; Xiao et al., 2024). Research on how AI inherits or amplifies such entrenched bias can directly reveal technology's "replication–reinforcement" mechanisms in social structures. Moreover, in hiring contexts, gender discrimination is often encoded into algorithms through historical data (e.g., male-dominated tech industry hiring records), causing AI systems to weight female candidates lower (Halzack, 2019). This "data–algorithm–outcome" chain is clearly traceable, facilitating analysis of underlying technical ethical issues. Compared to educational or age discrimination, gender discrimination more easily strips away confounding variables (e.g., the relationship between education and ability may be more complex), providing purer conditions for testing AI fairness interventions. Therefore, Study 2 primarily uses gender discrimination as the moral scenario for exploring AI's moral deficiency effect. Through a series of sub-studies (one correlational and two experimental), we confirmed that moral response levels to gender-discriminatory decisions are lower when the decision-maker is AI. This series not only confirms the robustness of AI's moral deficiency effect but also uses scenario stripping methods to demonstrate gender discrimination's theoretical advantages as an ideal experimental context. Compared to educational or age discrimination, gender discrimination more easily strips away confounding variables, providing pure variable conditions for mechanism research and intervention experiments.

5.2 Parallel Mediation of Agency and Experience

By integrating mind perception theory with moral dyad theory, this study systematically reveals the dual-pathway mechanism of AI's moral deficiency effect. Previous research has primarily explained this phenomenon through individual psychological pathways, such as free will beliefs (Bigman et al., 2023) or biased motivation (Xu et al., 2022), presenting fragmented mechanistic explanations. This study, starting from the moral agent perspective, elucidates agency and experience's mediating roles in AI's moral deficiency effect, representing significant theoretical advancement for mind perception theory and moral dyad theory. Unlike previous fragmented explorations of single variables like free will or autonomy (Bigman et al., 2023; Heinrichs et al., 2022; Xu et al., 2022), this study is the first to confirm parallel mediation of agency and experience in AI's moral deficiency effect, breaking through the traditional "agency (subject)–experience (object)" dichotomy (Gray et al., 2012). While existing research has examined agency's impact on AI's moral deficiency effect (Hohenstein & Jung, 2020; Wilson et al., 2022; Zhu & Chu, 2025), our findings further indicate that simply emphasizing agency enhancement (e.g., increasing decision trans-

parency) is insufficient—synergistic optimization of the experience dimension (e.g., emotional interaction design) is equally crucial. Experimental data show that when AI's experience perception is strengthened, its evaluation as a moral agent significantly improves, strongly supporting the emerging "emotional rationalism" perspective that experience is not only necessary for moral patients but also an intrinsic component of moral agent qualification (de Vel-Palumbo et al., 2022). This theoretical breakthrough challenges traditional moral dyad theory's narrow positioning of the experience dimension, offering new perspectives for understanding the multifaceted composition of moral agents.

This study provides, for the first time, causal evidence of parallel mediation via experimental mediation methods, compensating for previous questionnaire-based research (Sullivan et al., 2022; Xu et al., 2022) that could only verify correlations. Structural equation modeling further demonstrates that the combined explanatory power of agency and experience exceeds single pathways (as in Study 2c), deepening understanding of the moral deficiency effect's complexity. Specifically, this research achieves causal inference of AI decision-making's moral deficiency mechanisms through multi-method validation combining experimental and questionnaire approaches. The 2×2 experimental design reveals interactions between decision-maker identity and mind dimensions: under AI conditions, agency manipulation significantly enhances moral responses, while human agents show no such effect. This "agency attribution bias" confirms mind perception theory's core assumption that humans are inherently endowed with complete mind schemas (Leyens et al., 2000), while AI's "subject-like" status is constructive and modifiable through mind dimension representation design (Bigman et al., 2023). By manipulating perceived agency/experience (high vs. control) and decision-maker (AI vs. human), we found that in control groups, decision-maker significantly predicted agency/experience levels, while AI group agency enhancement significantly improved moral responses, constructing a complete causal chain of "agent type $\rightarrow$ mind perception $\rightarrow$ moral response." This design overcomes traditional questionnaire methods' inherent limitations in causal inference, breaking previous research constraints of correlational analysis. Subsequent questionnaire methods simultaneously examined parallel mediation of agency and experience, revealing synergistic gain effects. In the gender discrimination scenario, the dual-pathway standardized coefficient was 0.398, 95% CI = [0.323, 0.467], jointly explaining 39.8% of variance—substantially more than single pathways (agency mediation alone: standardized coefficient = 0.214, 95% CI = [0.032, 0.383]; experience mediation alone: standardized coefficient = 0.184, 95% CI = [0.003, 0.373]), suggesting real-world moral judgment may follow dual mechanisms of "intentional action" and "emotional sensitivity." This advancement reveals AI decision-making's moral deficiency deep mechanism—people believe algorithms lack both autonomous intention (low agency) and ability to understand emotional harm (low experience), leading to insufficient moral responses. This resolves the fragmentation problem where different studies emphasized different mediators, forming a unified explanatory framework.

5.3 Combined Effects of Anthropomorphism and Expectancy Adjustment

This study innovatively proposes and validates a "anthropomorphism + expectancy adjustment" combined intervention strategy to mitigate AI's moral deficiency effect by enhancing agency and experience perception. Specifically, through randomized controlled experiments, we propose dual intervention strategies exploring how anthropomorphic design and expectancy adjustment can mitigate AI's moral deficiency effect. Results show that both single anthropomorphism intervention and expectancy adjustment have significant positive effects on enhancing moral responses and improving AI agency and experience perception; their synergistic effects are even more pronounced. These findings provide empirical foundations for constructing systematic intervention solutions.

The anthropomorphic design introduced in this study, by endowing AI with human-like images, voices, or behavioral characteristics, makes it easier for people to perceive high mind perception levels and view AI as a "complete subject" with moral responsibility (Lin et al., 2020; Gursoy et al., 2019; Melián-González et al., 2021). Two-way ANOVA results show anthropomorphism intervention achieved significant effects in enhancing moral response, agency, and experience, with effects further amplified under expectancy manipulation. This indicates anthropomorphic design can elevate moral response levels when facing unethical AI decisions. Moreover, mediation path analysis further reveals that anthropomorphism's enhancement of moral responses is achieved by strengthening individuals' perception of AI's agency and experience.

Beyond anthropomorphism intervention, expectancy adjustment as another intervention approach also demonstrates unique effects in eliminating AI's moral deficiency (Srinivasan & Sarial-Abi, 2021). Research shows that presetting appropriate expectations for AI behavior not only significantly enhances moral response levels but also validates this intervention mechanism through agency and experience mediation effects. Further moderated mediation tests indicate that the synergistic mechanism of anthropomorphism and expectancy adjustment primarily relies on dual enhancement of AI agency and experience perception. While existing research (Bigman et al., 2023; Hu et al., 2024; Xu et al., 2022) has examined anthropomorphism or expectancy adjustment's intervention effects from different perspectives, few studies have simultaneously examined both AI and human intervention angles to investigate combined effects. For example, Bigman et al. (2023) and Xu et al. (2022) only examined anthropomorphism's intervention effects; Hu et al. (2024) noted both approaches through literature review but did not propose combined intervention solutions. This study is the first to propose and examine dual-pathway intervention from AI design (anthropomorphism) and human cognition (expectancy adjustment), with experiments proving synergistic effects significantly stronger than single strategies. Through theoretical model innovation of intervention pathways, we advance from Bigman et al. (2023) and Xu et al. (2022)'s single free will or biased motivation

explanations to an "agency–experience" dual-pathway model, generating multi-dimensional intervention strategies that more effectively mitigate AI's moral deficiency phenomenon and providing innovative psychological pathways for AI ethics governance.

5.4 Limitations and Future Directions

Despite these contributions, this study has several limitations. First, there are scenario limitations in experimental design. Specifically, Studies 1 and 2 primarily examined discriminatory moral scenarios (e.g., gender, educational, and age discrimination), while Study 3 focused on autonomous driving moral dilemmas. Although these scenarios have significant social relevance, they only reflect specific types of moral problems. Notably, AI's real-world ethical challenges are more diverse, including privacy protection, life-rights trade-offs in medical decision-making, and public resource allocation. Public response patterns to AI decisions may differ significantly across these moral domains. For example, moral dilemmas involving personal safety may trigger stronger emotional responses, while privacy violations may focus more on responsibility attribution. Future research should expand scenario diversity to include environmental protection, educational equity, judicial sentencing, and other domains. This expansion can more comprehensively define the boundary conditions of AI's moral deficiency effect and provide richer empirical foundations for building more inclusive AI ethical governance frameworks.

Second, there are ecological validity concerns with experimental methods. This study primarily employed scenario simulation paradigms presenting moral decision-making situations through text descriptions and video materials. While this approach has clear advantages for variable control and experimental manipulation, its ecological validity has limitations. Laboratory responses may not fully reflect complex psychological processes in real scenarios. Real-world moral responses often occur in dynamically changing environments influenced by social norms, cultural backgrounds, and personal experiences. For example, in actual hiring scenarios, interactions between applicants and AI systems may last weeks, with moral cognition evolving over time—fundamentally different from immediate laboratory reactions. To improve external validity, future work could: conduct field experiments embedding research in real AI application scenarios like medical diagnostic systems or corporate automated hiring platforms; adopt longitudinal designs examining cumulative effects of long-term AI exposure on individual moral sensitivity; and combine multimodal data collection techniques (e.g., eye-tracking, physiological monitoring) to more comprehensively capture participants' response patterns in authentic contexts. These methodological innovations will significantly enhance research results' real-world explanatory power, providing more actionable scientific foundations for AI ethics governance.

Third, this study treats anthropomorphism and expectancy adjustment as independent intervention strategies, proposing potential solutions from AI charac-

teristics and human perspectives respectively. However, theoretically, anthropomorphism as a means of strengthening AI's human-like features may directly influence people's expectations of AI's mind capabilities. Perceived human-like features can facilitate human-computer interaction, prompting people to transfer social heuristic judgments to robot interaction contexts, thereby forming higher expectations (Duffy, 2003; Nass & Moon, 2000). However, although anthropomorphic design enhances surface similarity, AI's actual behavior often fails to fully meet user expectations, and this expectation gap may trigger significant negative emotions (Grazzini et al., 2023). Research shows anthropomorphism can enhance trust in technology, but when behavioral performance falls short of expectations, trust may transform into stronger disappointment and resistance (Waytz et al., 2014; Crolic et al., 2022). These clues suggest anthropomorphism and expectancy adjustment intersect in psychological mechanisms and are not completely independent. Therefore, the pathways of combined intervention strategies may be more complex than hypothesized in this study. Future research should examine how these strategies interact in long-term interactions across different individuals and contexts, thereby more comprehensively explaining how anthropomorphism and expectancy adjustment interact under various conditions and providing more targeted theoretical foundations and practical guidance for AI moral decision-making and human-computer interaction optimization.

In conclusion, this study's findings are: First, people's moral responses to AI's unethical decisions are significantly weaker than those to human decision-makers. Second, compared to human agents, agency and experience manipulations significantly enhance moral responses in AI decision-making contexts; moreover, perceived agency and experience serve as parallel mediators in AI's moral deficiency effect. Third, both single anthropomorphism intervention and expectancy adjustment have significant positive effects on enhancing moral responses and improving AI agency and experience perception; the synergistic effect of combined strategies is even stronger.

Note: Figure translations are in progress. See original paper for figures.