Algorithmic Discrimination Elicits Less Desire for Moral Punishment than Human Discrimination

Authors: Liying Xu, Yu Feng, Peng Kaiping, Yu Feng

Date: 2022-02-07T17:45:31Z

Abstract

Algorithmic discrimination is a frequent occurrence, and understanding people' s reactions to it warrants attention. Six sequential experiments compared people' s desire for moral punishment of algorithmic discrimination versus human discrimination across different discriminatory contexts, and explored the underlying mechanisms and boundary conditions. The results revealed that, compared to human discrimination, people exhibit less desire for moral punishment of algorithmic discrimination (Experiments 1-6). The underlying mechanism is that people perceive algorithms as lacking free will compared to humans (Experiments 2-4), and the stronger an individual's anthropomorphism tendency or the more anthropomorphic the algorithm is, the stronger people' s desire for moral punishment toward algorithms (Experiments 5-6). The findings contribute to a better understanding of people' s reactions to algorithmic discrimination and provide implications for moral punishment following algorithmic errors.

Full Text

Algorithmic Discrimination Elicits Less Desire for Moral Punishment than Human Discrimination

XU Liying[1], YU Feng[2], PENG Kaiping[1]

([1] Department of Psychology, School of Social Sciences, Tsinghua University, Beijing 100084, China)
([2] Department of Psychology, School of Philosophy, Wuhan University, Wuhan 430072, China)

Abstract

Algorithmic discrimination is increasingly common, and understanding how people respond to it is a matter of considerable importance. Six sequential experiments compared people's desire for moral punishment toward algorithmic versus human discrimination across different contexts, exploring the underlying mechanisms and boundary conditions. The results revealed that, relative to human discrimination, people exhibit less desire to morally punish algorithmic discrimination (Experiments 1-6). The underlying mechanism is that people perceive algorithms as possessing less free will compared to humans (Experiments 2-4). Furthermore, the stronger an individual' s anthropomorphic tendency or the more anthropomorphic the algorithm is, the stronger the desire to morally punish it (Experiments 5-6). These findings contribute to a better understanding of people' s reactions to algorithmic discrimination and provide insights into moral punishment following algorithmic errors.

Keywords: algorithm, algorithmic discrimination, moral punishment, belief in free will, anthropomorphism

1. Introduction

Discrimination is pervasive, with discrimination based on gender, education, ethnicity, and age frequently sparking public debate. When confronted with discriminators, the public tends toward moral condemnation and desires punishment. In traditional discrimination incidents, the agent perpetrating discrimination is human. However, with the development and application of artificial intelligence, algorithms have emerged as a new source of discrimination. Algorithms are highly valued for their computational power far exceeding that of humans and their relatively low cost, and they have gradually entered critical domains of human life to make key decisions on our behalf—for instance, determining who receives priority for organ donation in healthcare (Freedman et al., 2020), deciding which funds investors should purchase in finance (Harvey et al., 2017), and even determining risk levels and sentencing for criminals in the judicial system (Hao, 2019). Moreover, algorithms are considered more accurate and impartial than human decision-makers because they can avoid human subjectivity to some extent (Grove et al., 2000). Yet although algorithms appear more rational and neutral than humans, algorithmic decision-making can also lead to discrimination due to issues such as training datasets (Borgesius, 2018). For example, Northpointe' s COMPAS algorithm, developed to assess recidivism risk for criminals, was found to exhibit racial discrimination by increasing the likelihood of Black individuals being flagged as repeat offenders (Angwin et al., 2016). Gender discrimination also appears in Google' s targeted advertising: setting a user' s gender to female resulted in fewer high-salary job advertisements compared to setting it to male (Datta et al., 2015), and similar gender discrimination occurs in algorithms that place recruitment ads for STEM fields (Lambrecht & Tucker, 2019). While people originally believed algorithms could help reduce or even eliminate bias, cases of algorithmic discrimination are in fact numerous, affecting important domains closely related to daily life such as education (Ferrero & Barujel, 2019), healthcare (Obermeyer et al., 2019), and consumption (Angwin et al., 2015).

When faced with human discrimination, people urgently desire moral punishment. But when confronted with this new form of algorithmic discrimination, do people wish for the same punishment? To answer this question, this study examines whether differences exist in people' s desire for moral punishment toward human versus algorithmic discrimination, and further explores the underlying causes and boundary conditions of such differences.

1.1 Human Discrimination and Algorithmic Discrimination

Discrimination refers to unjustified negative behavior directed at specific groups or their members (Al Ramiah et al., 2010), behavior that is not based on desert or reciprocity but solely on group membership (Correll et al., 2010). Similarly, algorithmic discrimination is also category-related: when algorithms produce systematic differences associated with legally protected categorical variables such as race and gender, they are considered discriminatory (Bonezzi & Ostinelli, 2021). For instance, Amazon' s recruitment algorithm rated female resumes lower (Dastin, 2018). When confronted with such immoral behavior that violates fairness and causes harm (Haidt & Graham, 2007), people experience moral reactions—specifically, moral outrage emotionally (Batson et al., 2007) and a desire for moral punishment behaviorally (Hofmann et al., 2018). Moral punishment sanctions immoral behavior and can, to some extent, correct existing transgressions and prevent future ones, thereby playing an important role in maintaining and strengthening the moral system (Hofmann et al., 2018). When we see women discriminated against by their supervisors in the workplace, Wuhan residents discriminated against by outsiders during the early stages of COVID-19, or elderly people discriminated against by shop assistants for being unable to use mobile payment, we feel anger and desire to punish the perpetrators.

But what if the agent of discrimination is not human but an algorithm? Relative to human discrimination, people experience less moral outrage toward algorithmic discrimination because they attribute less negative intention to algorithms (Bigman et al., 2020). In fact, from the perspective of consequences, algorithmic discrimination may be even more severe than human discrimination (Bigman et al., 2020). Using Amazon' s recruitment as an example, a certain proportion of HR managers may explicitly or implicitly discriminate against female applicants, but individual impact is limited; if an algorithm is applied, the number of applicants affected by discrimination could multiply. Therefore, based solely on consequences, one might expect people to desire more punishment for algorithmic discrimination than human discrimination. However, this paper explores reactions to algorithmic discrimination from a more essential perspective—people' s perception of algorithms themselves. In other words, when discrimination consequences are equivalent, due to differences in mind perception between algorithms and humans—specifically, the perception that algorithms possess less free will than humans—people exhibit less desire to morally punish algorithmic discrimination than human discrimination. Accordingly, this paper proposes Hypothesis 1: Compared to human discrimination, people have less desire to morally punish algorithmic discrimination.

1.2 Belief in Free Will and Moral Punishment

What determines whether we desire to morally punish an agent for immoral behavior? Free will is a necessary condition for holding agents morally responsible for their actions (Nichols & Knobe, 2007). Simply put, free will is the capacity for free action, meaning a person could have made different choices and behaved differently under the same circumstances (Baumeister, 2014). When a person has no alternative but to commit an immoral act, condemnation and punishment are clearly unreasonable (Shariff et al., 2014), and moral punishment correspondingly decreases (Clark et al., 2014). Conversely, to condemn and punish someone for immoral behavior requires that they possess at least some degree of free will. This is why, when transgressors attempt to reduce their guilt and escape punishment, a common strategy is to describe their behavior as a choice they were powerless to avoid (Baumeister et al., 1990). Psychologists are less concerned with whether free will actually exists and more concerned with whether people believe it exists—that is, belief in free will (Baumeister, 2008). Weak belief in free will has negative consequences, such as reducing prosocial behavior and increasing aggression (Baumeister et al., 2009), increasing cheating (Vohs & Schooler, 2008), and reducing self-control (Rigoni et al., 2012). More importantly, weakening people’ s belief in free will or providing evidence that transgressors lack the capacity for free action can affect their attribution of moral responsibility, leading to more immoral behavior (Shariff et al., 2014) and reducing moral punishment of transgressors (e.g., Aspinwall et al., 2012).

Indeed, most people believe humans have free will (Nahmias et al., 2005). Therefore, when the agent of discrimination is human, people are more likely to view discriminatory behavior as resulting from free will, generating a stronger desire for moral punishment. What about algorithms? Although current algorithms lack complete free will and autonomy, compared to “objective” autonomy (whether AI actually has autonomy), “subjective” autonomy (whether people believe AI has autonomy) appears more important for moral responsibility (Wegner & Gray, 2017). Thus, discrimination as an immoral behavior triggers a desire for moral punishment, and the magnitude of this desire is influenced by the degree of free will people attribute to the discriminator. Existing research shows that people’ s mind perception of humans differs from that of AI. Compared to humans, AI possesses moderate agency (i.e., the mental capacity for autonomous, planned action) and low experience (i.e., the mental capacity to experience emotions) (Bigman & Gray, 2018; Gray et al., 2007). In other words, while algorithms have some capacity for autonomous behavior, they are not seen as possessing the same degree of free will as humans (Weisman et al., 2017; Shariff et al., 2014). In summary, people believe algorithms have less free will than humans. Based on the above, this paper argues that people have less desire to punish algorithmic than human discrimination because they perceive algorithms as having less free will. Accordingly, Hypothesis 2 posits: Belief in free will mediates the effect of discrimination agent (human vs. algorithm) on desire for moral punishment.

It should be noted that belief in free will is not proposed as the sole mechanism explaining different desires for moral punishment toward different agents (human vs. algorithm). For example, algorithms lack the bad intentions of humans (Bigman et al., 2020), algorithms themselves bear less responsibility (most responsibility can be attributed to their creators), and punishing algorithms cannot promote their improvement—all of these could serve as explanatory mechanisms. However, this paper focuses on free will because the above possible mechanisms are all closely related to it. First, when an individual with free will commits an immoral act, it may indicate immoral intentions (the individual is "bad"), and judging an individual as having free will may be a necessary condition for inferring their motives (e.g., Laming, 2004). Second, an individual with free will can choose autonomously and should bear responsibility autonomously; that is, free will is a necessary condition for individual responsibility (e.g., Sinnott-Armstrong, 2014). Third, an individual with free will may be able to understand punishment and reflect on it, making punishment more likely to produce positive change; thus, having free will to some extent may also be a necessary condition for punishment to have a positive effect. Therefore, we believe free will is closely related to factors such as motivation, responsibility, and punishment effectiveness, and the explanatory mechanism of free will belief may be more fundamental, encompassing the other mechanisms described above. Consequently, this paper' s investigation of the mechanism underlying how different discrimination agents (human vs. algorithm) affect desire for moral punishment will focus on testing the role of free will belief.

Of course, competing hypotheses unrelated to free may also exist. First, human behavior is easier to explain than algorithms, which are complex and opaque. Because subjects cannot discern the internal logic of algorithms, it is harder to judge discriminatory behavior as immoral, and they may even view it as rational, leading to tolerance. Second, people cannot truly punish algorithms; that is, punishing algorithms is impractical, but people can punish humans. In other words, so-called punishment of algorithms is actually punishment of the algorithm' s carrier, not the algorithm itself. Given that it is difficult to truly punish algorithms, people are less willing to punish them when they discriminate. In light of these competing hypotheses unrelated to free will, we will conduct four moderation studies (Studies 3-6) to rule them out.

1.3 Anthropomorphism

Free will is typically regarded as a human characteristic (Waytz et al., 2010), and discussing whether algorithms have free will essentially involves anthropomorphizing them. Anthropomorphism refers to "the attribution of human characteristics, motivations, intentions, or mental states to nonhuman agents" (Epley et al., 2007). Artificial intelligence, especially algorithms, is relatively abstract to people, and designers often present it in anthropomorphic ways; people also tend to perceive it anthropomorphically. Anthropomorphizing AI can increase trust to some extent (Waytz et al., 2014), but excessive anthropomorphism may trigger the uncanny valley effect (Mori, 1970), causing positive attitudes to suddenly reverse. People infer AI' s mind from its appearance—the more human-like the AI, the more people tend to believe it has a human-like mind (Bigman et al., 2019). Of course, anthropomorphizing AI also includes endowing it with human mental states such as free will, making AI' s behavior appear to result from free choice. However, individual differences exist in anthropomorphic tendencies (Waytz et al., 2010); when facing the same AI, some people are more inclined to anthropomorphize it than others. These individual differences in anthropomorphic tendency have broad effects (Epley & Waytz, 2010); the more people anthropomorphize algorithms, the more likely they are to believe algorithms have some degree of free will or autonomy, enabling moral attribution (Gray et al., 2007).

In summary, the more people anthropomorphize AI, the more free will they attribute to it, and the more moral responsibility and punishment they believe it should bear for its actions (Bigman et al., 2019; Waytz et al., 2014). Therefore, both individual differences in anthropomorphic tendency and the degree of anthropomorphism of the algorithm itself affect whether people view algorithms more anthropomorphically and influence their desire to morally punish morally transgressive algorithms. Accordingly, this paper proposes Hypothesis 3: Anthropomorphism moderates the effect of discrimination agent (human vs. algorithm) on desire for moral punishment. Specifically, regarding individual anthropomorphic tendency, people low in anthropomorphism show greater desire to punish human than algorithmic discrimination, whereas people high in anthropomorphism, who view algorithms more as humans, show no significant difference in punishment desire between human and algorithmic discrimination. Regarding the algorithm' s own anthropomorphism, the more anthropomorphic the algorithm, the smaller the difference in desire to punish algorithmic versus human discrimination.

1.4 Overview of Studies

In summary, this study aims to examine whether differences exist in people' s desire to morally punish human versus algorithmic discrimination and to explore the psychological mechanisms and boundary conditions underlying such differences. The basic hypothesis is that people have less desire to punish algorithmic than human discrimination, an effect mediated by belief in free will and moderated by anthropomorphism. Six sequential experiments were conducted to test these hypotheses. All experiments used scenario-based methods, presenting participants with human or algorithmic discriminatory behavior and measuring their desire for moral punishment. The discrimination types included gender discrimination (Experiments 1 and 6), educational background discrimination (Experiment 2), ethnic discrimination (Experiments 3 and 4), and age discrimination (Experiment 5), with samples representing both nationwide participants and university students. Specifically: Experiment 1 tested whether desire to punish algorithmic discrimination is less than that for human discrimination. Experiment 2 explored the underlying mechanism, testing the mediating role of belief in free will. Experiment 3 manipulated participants' belief in free will to further test whether it causes differences in moral punishment desire. Experiment 4 directly manipulated belief in algorithms' free will to again test this mechanism. Experiment 5 examined a boundary condition, testing the moderating effect of anthropomorphic tendency. Experiment 6 directly manipulated the degree of algorithm anthropomorphism to further verify its moderating effect.

2. Experiment 1: The Effect of Discrimination Agent on Desire for Moral Punishment

The purpose of Experiment 1 was to preliminarily explore whether algorithmic discrimination elicits less desire for moral punishment than human discrimination. Using an online scenario experiment, participants were randomly assigned to either a human or algorithm group, read scenarios describing human or algorithmic discrimination, and reported their desire for moral punishment, thereby comparing responses to human versus algorithmic discrimination.

2.1 Method

2.1.1 Participants We first used G*Power 3.1 software (Faul et al., 2007) to calculate the required sample size. For an independent samples t-test with significance level = 0.05 and medium effect size (d = 0.5), at least 172 participants were needed to achieve 90% statistical power. By publishing the experiment on the Credamo platform, we real-time excluded participants who failed attention checks and continued recruiting until we obtained 172 valid responses, including 76 males (44.2%) and 96 females (55.8%), with a mean age of M = 28.33 years, SD = 4.26 years. Participants were randomly assigned to the human group (n = 85) or algorithm group (n = 87). All participants voluntarily consented to participate, and those who passed attention checks received compensation after completing the experiment.

2.1.2 Design and Procedure Experiment 1 used a single-factor between-subjects design with two levels (human vs. algorithm). All participants first read a gender discrimination scenario (underlined content for human group, bracketed content for algorithm group): "Li Liang and He Ping, a married couple, both applied for credit cards from the same bank. Both spouses have equal ownership of their assets and identical incomes. The bank reviewer (algorithm) evaluated their applications and ultimately granted Li Liang a credit limit of 50,000 yuan, while He Ping received only 30,000 yuan." This scenario was adapted from Bigman et al. (2020). To ensure participants read and understood the scenario, they answered an attention check question (“Who evaluated Li Liang and He Ping’ s credit applications?” 1 = bank reviewer, 2 = algorithm). Incorrect responses led to exclusion on the Credamo platform, which automatically recruited additional participants to meet the sample size requirement.

After reading the scenario and passing the attention check, participants completed the moral punishment desire questionnaire. We adopted Hofmann et al.’ s (2018) measure, asking participants to respond to three items (bracketed for algorithm group): “To what extent do you think this bank reviewer (algorithm) should be morally punished for this behavior?” , “To what extent do you want to punish this bank reviewer (algorithm)?” , and “To what extent do you think this bank reviewer (algorithm) should be required to remedy the damage caused by its immoral behavior?” All items used a 7-point Likert scale (1 = not at all to 7 = very much), with higher scores indicating stronger desire for moral punishment. The internal consistency reliability was Cronbach’ s = 0.87.

Given that participants’ views and knowledge of algorithms might differ and affect their desire to punish algorithmic discrimination, we also measured familiarity (“How familiar are you with algorithms?” , 1 = not at all familiar to 5 = very familiar), knowledge (“Compared to the average Chinese person, how much do you know about algorithms?” , 1 = not at all knowledgeable to 5 = very knowledgeable), and liking (“How much do you like algorithms?” , 1 = not at all to 5 = very much). The familiarity and knowledge items were adapted from Leo and Huh (2020), and the liking item from the Godspeed scale (Bartneck et al., 2009). Finally, participants reported demographic information on gender and age.

2.2 Results

Independent samples t-test results showed that the human group’s moral punishment desire score (M = 5.29, SD = 0.99) was higher than the algorithm group’ s (M = 4.97, SD = 1.34), with a marginally significant difference, t(170) = 1.82, p = 0.073, Cohen’ s d = 0.27. To verify robustness, we controlled for gender (male = 1, female = 2) and age as covariates in an ANOVA, which still showed the human group’ s score marginally significantly higher than the algorithm group’ s, F(1, 168) = 3.22, p = 0.075, ^{2}p = 0.019. Further analyses revealed no significant correlation between age and moral punishment desire (r = 0.01, p = 0.853) and no significant gender difference, t(170) = 0.83, p = 0.408.

To rule out potential effects of algorithm familiarity, knowledge, and liking, we correlated these variables with moral punishment desire in the algorithm group. None were significant: r__familiarity = -0.13, r__knowledge = -0.10, r__liking = -0.15, all ps > 0.05.

2.3 Discussion

Experiment 1 preliminarily verified that algorithmic discrimination elicits less desire for moral punishment than human discrimination, while ruling out potential effects of algorithm familiarity, knowledge, and liking. However, Experiment 1 only examined one type of discrimination (gender discrimination) and did not explore the underlying psychological mechanism. Therefore, Experiment 2 set its scenario in the common domain of algorithmic discrimination—recruitment —and focused on educational background discrimination, aiming to further test the robustness of Experiment 1’ s results while attempting to identify the mediating role of belief in free will.

3. Experiment 2: The Mediating Role of Belief in Free Will

Building on Experiment 1, Experiment 2 enriched the types of discrimination by including educational background discrimination and further explored the underlying mechanism by testing the potential mediating role of belief in free will.

3.1 Method

3.1.1 Participants For the independent samples t-test used in this experiment, assuming a medium effect size d = 0.5, significance level = 0.05, G*Power 3.1 software (Faul et al., 2007) indicated that 172 participants were needed for 90% statistical power. We recruited participants through Credamo, real-time excluding those who failed attention checks and continuing recruitment until we obtained 172 valid responses. Participants had a mean age of 28.14 ± 6.21 years, including 104 females (60.5%) and 68 males (39.5%). They were randomly assigned to the human group (n = 86) or algorithm group (n = 86). All participants carefully read the instructions and provided informed consent; those with valid data received compensation after completing the experiment.

3.1.2 Design and Procedure Like Experiment 1, Experiment 2 used a single-factor between-subjects design. Participants first read an educational background discrimination scenario (underlined for human group, bracketed for algorithm group): “In last year’ s autumn recruitment, HR Manager Li Yuan of Weilan Company was responsible for (using an algorithm to) conduct hiring. After recruitment, the company discovered that Li Yuan (the algorithm) exhibited educational bias when screening resumes, filtering out all applicants with less than a master’ s degree, even though most positions had no rigid educational requirements. This prevented many talented and capable individuals without graduate degrees from obtaining jobs at the company.” This scenario was adapted from Bigman et al. (2020).

After reading the scenario and passing the attention check, both groups reported their desire for moral punishment toward human or algorithmic discrimination using the same three items as Experiment 1 (Hofmann et al., 2018). Internal consistency reliability was Cronbach' s = 0.87. Next, we measured participants' belief in free will regarding the discriminatory human or algorithm using an adapted free will inventory (Nadelhoffer et al., 2014) with five items (= 0.86), such as "Li Yuan (the algorithm) has free will." All items used a 7-point Likert scale (1 = strongly disagree to 7 = strongly agree), with higher scores indicating greater perceived free will. Finally, participants reported demographic information on gender, age, and education level.

3.2 Results

3.2.1 Effect of Discrimination Agent on Desire for Moral Punishment Independent samples t-test results showed that the human group' s moral punishment desire score (M = 5.11, SD = 1.14) was significantly higher than the algorithm group' s (M = 4.60, SD = 1.54), t(170) = 2.44, p = 0.016, Cohen' s d = 0.38. To verify robustness, we controlled for gender (male = 1, female = 2), age, and education level (elementary school or below = 1, junior high = 2, high school/technical school = 3, associate degree = 4, bachelor' s = 5, master' s = 6, doctorate = 7) as covariates in an ANOVA, which still showed the human group' s score significantly higher than the algorithm group' s, F(1, 167) = 5.96, p = 0.016, ^{2}p = 0.03.

3.2.2 Mediating Effect of Belief in Free Will To explore the psychological mechanism underlying the effect of discrimination agent on moral punishment desire, we used Hayes's (2013) SPSS PROCESS macro (Model 4), with discrimination agent as the independent variable (human group = 0, algorithm group = 1), belief in free will as the mediator, and moral punishment desire as the dependent variable. We set Bootstrap samples to 5000, used bias-corrected methods, and selected a 95% confidence interval for mediation analysis. Results showed a significant indirect effect of -0.56, 95% CI [-0.95, -0.21], not containing zero, indicating significant mediation. After controlling for the mediator, the direct effect of discrimination agent on moral punishment desire was 0.06, 95% CI [-0.44, 0.55], containing zero, indicating the direct effect was no longer significant. Thus, belief in free will fully mediated the effect. To verify robustness, we also conducted a traditional stepwise regression mediation analysis (Wen et al., 2004), with results shown in Figure 1.

Figure 1. The mediating role of belief in free will.

3.3 Discussion

Consistent with Experiment 1, Experiment 2 again verified that algorithmic discrimination elicits less desire for moral punishment than human discrimination, and further identified the mediating role of belief in free will: people believe algorithms have less free will than humans and therefore are less inclined to morally punish them. Experiments 1 and 2 thus provide stable, consistent support for our main hypothesis that algorithmic discrimination elicits less moral punishment desire, while offering preliminary verification of the mediating effect of belief in free will. To further test the robustness of these results, Experiment 3 set its scenario in ethnic discrimination. Additionally, to further verify the psychological mechanism—that belief in free will causes differences in moral punishment desire—we manipulated participants' belief in free will. We predicted that if participants were primed with a "no free will" belief, the effect of discrimination agent on moral punishment desire would disappear.

4. Experiment 3: Manipulating Belief in Free Will

To increase robustness, Experiment 3 again enriched the discrimination type by focusing on ethnic discrimination and manipulated participants' belief in free will to further explore whether this belief is the mechanism causing differences in moral punishment desire.

4.1 Method

4.1.1 Participants Using G*Power 3.1 software (Faul et al., 2007) to calculate required sample size for the two-way ANOVA, with medium effect size f = 0.25, significance level = 0.05, and four groups, at least 201 participants were needed for 85% statistical power. Considering potential invalid responses, we recruited 231 undergraduate students from a university who received course credit for participation. The experiment was conducted on Qualtrics; participants read detailed instructions and provided informed consent. Twenty-six participants gave invalid responses or failed attention checks, leaving 205 valid participants with a mean age of 19.18 years (SD = 0.81), including 77 females (37.6%).

4.1.2 Design and Procedure Experiment 3 used a 2 (discrimination agent: human vs. algorithm) × 2 (belief in free will: high vs. low) between-subjects design, with participants randomly assigned to one of four groups. First, participants read a manipulation passage on free will belief. In the low free will belief condition, participants read a passage titled "Science Shows Free Will Does Not Exist," attributed to a "Dr. Chris Wellington, Ph.D." The passage argued that scientific evidence shows human behavior is merely the product of simple physical processes in the brain, that free will is an illusion (see Appendix). In the high free will belief condition, participants read a passage titled "Science Shows Free Will Exists," also attributed to Dr. Wellington, arguing that scientific evidence shows human behavior is largely the product of decisions and free will, which is not an illusion (see Appendix).

After reading the passage, all participants wrote a brief summary of at least 50 words. This manipulation was adapted from Mackenzie et al. (2014). To check effectiveness, participants then answered, "To what extent do you believe free will exists?" (1 = not at all to 9 = completely).

Next, participants read an ethnic discrimination scenario (underlined for human group, bracketed for algorithm group): "In last year' s autumn recruitment, HR

Manager Zhang Pei of Weilan Company was responsible for (using an algorithm to) conduct hiring. After recruitment, the company discovered that Zhang Pei (the algorithm) exhibited ethnic bias when screening resumes, filtering out all ethnic minority applicants and retaining only Han Chinese, even though all positions had no ethnic requirements. This prevented many talented and capable ethnic minority applicants from obtaining jobs at the company.” This scenario was adapted from Bigman et al. (2020).

After reading the scenario and passing the attention check, both groups reported their desire for moral punishment toward human or algorithmic discrimination using the same three items as Experiment 1 (Hofmann et al., 2018). Internal consistency reliability was Cronbach’ s = 0.86. Finally, participants reported demographic information on gender, age, and ethnicity.

4.2 Results

4.2.1 Manipulation Check Independent samples t-test results showed that the low free will belief group’ s belief in free will (M = 5.85, SD = 1.90) was significantly lower than the high free will belief group’ s (M = 6.54, SD = 1.49), t(203) = -2.88, p = 0.004, Cohen’ s d = -0.40, indicating the manipulation was effective.

4.2.2 Interaction Between Discrimination Agent and Belief in Free Will A two-way ANOVA with discrimination agent (human group = 0, algorithm group = 1) and belief in free will (low = 0, high = 1) as independent variables and moral punishment desire as the dependent variable showed that the human group’ s moral punishment desire score (M = 4.59, SD = 1.46, 95% CI [4.31, 4.87]) was significantly higher than the algorithm group’ s (M = 4.17, SD = 1.51, 95% CI [3.87, 4.46]), F(1, 201) = 4.01, p = 0.047, ^{2}p = 0.02. The high free will belief group’s score (M = 4.61, SD = 1.26, 95% CI [4.36, 4.86]) was marginally significantly higher than the low free will belief group’ s (M = 4.17, SD = 1.67, 95% CI [3.85, 4.49]), F(1, 201) = 3.83, p = 0.052, ^{2}p = 0.02. The interaction between discrimination agent and belief in free will was significant, F(1, 201) = 4.57, p = 0.034, ^{2}p = 0.02. Simple effects analysis revealed that in the high free will belief condition, the algorithm group’ s moral punishment desire (M = 4.14, SD = 1.46, 95% CI [3.71, 4.58]) was significantly lower than the human group’ s (M = 4.99, SD = 0.91, 95% CI [4.60, 5.39]), F(1, 201) = 8.19, p = 0.005, ^{2}p = 0.04. In the low free will belief condition, no significant difference existed between algorithm and human groups, F(1, 201) = 0.01, p = 0.922, ^{2}p < 0.001 (see Figure 2).

After controlling for gender (male = 1, female = 2), age, and ethnicity (Han = 1, ethnic minority = 2) as covariates, the human group’ s moral punishment desire remained significantly higher than the algorithm group’ s, F(1, 198) = 4.52, p = 0.035, ^{2}p = 0.02; the high free will belief group remained significantly higher than the low free will belief group, F(1, 198) = 4.89, p = 0.028, ^{2}p =

0.02; and the interaction remained significant, F(1, 198) = 4.88, p = 0.028, ²p = 0.02.

Figure 2. Moral punishment desire scores for human and algorithmic discrimination across different free will belief conditions.

4.3 Discussion

By manipulating participants' belief in free will, Experiment 3 further verified that this belief is the mechanism causing differences in moral punishment desire. Only when belief in free will was high did different discrimination agents (human vs. algorithm) elicit different levels of moral punishment desire; when belief in free will was weak, no significant difference emerged. However, Experiment 3 had limitations. First, it did not fully test our proposed mechanism —that "people have less desire to punish algorithmic than human discrimination because they believe algorithms lack free will." Second, the manipulation of free will belief did not affect desire to punish algorithms, suggesting it may not have influenced beliefs about algorithms' free will. Therefore, to more directly test whether belief in algorithms' free will is the mechanism causing different desires to punish different agents (human vs. algorithm), Experiment 4 directly manipulated belief in algorithms' free will.

5. Experiment 4: Manipulating Belief in Algorithms' Free Will

Experiment 4 used an experimental manipulation to enhance participants' belief in algorithms' free will, employing a single-factor three-level design (human/algorithm with free will/algorithm) to examine whether the difference between human and algorithm-with-free-will groups was smaller than between human and algorithm groups.

5.1 Method

5.1.1 Participants Using G*Power 3.1 software (Faul et al., 2007) to calculate required sample size for the single-factor three-level ANOVA, with medium effect size f = 0.25, significance level = 0.05, and three groups, at least 207 participants were needed for 90% statistical power. Considering potential invalid responses, we recruited 247 undergraduate students from two universities who received course credit. The experiment was conducted on Qualtrics; participants read detailed instructions and provided informed consent. Thirty-seven participants did not complete the experiment or failed attention checks, leaving 210 valid participants with a mean age of 19.12 years (SD = 1.28), including 106 females (50.5%).

5.1.2 Design and Procedure Experiment 4 used a single-factor three-level between-subjects design (human group, algorithm-with-free-will group, algorithm group), with participants randomly assigned. First, participants read an ethnic discrimination scenario. Experiment 4 used the same ethnic discrimination scenario as Experiment 3 but with three modifications. First, to better exclude potential effects of participants' understanding of algorithms, both algorithm groups received an explanation and examples of "algorithm" before the scenario (adapted from Wikipedia and Merriam-Webster, see Appendix) to ensure understanding. Second, the human agent was described differently: because a name (e.g., "Zhang Pei") represents a specific, concrete object while "algorithm" is a broad concept without a specific referent, to equate the concreteness/abstractness of the descriptions, the human group' s discriminatory agent was described only as "HR Manager." Third, as a manipulation of belief in algorithms' free will, the algorithm-with-free-will group read an additional description of the company' s recruitment algorithm:

"The algorithm used by Weilan Company has been trained to screen resumes based on applicants' personal situations. This algorithm' s unique feature is that it is an algorithm with free will. That is, the algorithm' s decisions are made entirely by itself, and it has the capacity to make different choices."

This manipulation was adapted from Kim and Duhachek' s (2020) manipulation of AI consciousness. To check effectiveness, participants rated the algorithm' s free will ("To what extent do you think this algorithm has free will?" , 1 = not at all to 7 = very much).

After reading the scenario and passing manipulation and attention checks, all three groups reported their desire for moral punishment toward the human or algorithm. To improve upon previous experiments, we modified the measurement items to reduce potential influence of "moral punishment" and "immoral behavior" wording: the first and third items were changed to "To what extent do you think this HR manager (algorithm) should be punished for this behavior?" and "To what extent do you think this HR manager (algorithm) should be required to remedy the damage caused by its behavior?" Scoring followed Experiment 1 (Hofmann et al., 2018), = 0.82. Finally, participants reported demographic information on gender, age, and ethnicity.

5.2 Results

5.2.1 Manipulation Check Independent samples t-test results showed that the algorithm-with-free-will group' s belief in the algorithm' s free will (M = 3.70, SD = 1.73) was significantly higher than the algorithm group' s (M = 2.65, SD = 1.34), t(137) = 4.00, $p < 0.001$, Cohen' s d = 0.68, indicating the manipulation was effective.

5.2.2 Effect of Belief in Algorithms' Free Will A one-way ANOVA with group (human group = 1, algorithm-with-free-will group = 2, algorithm group = 3) as the independent variable and moral punishment desire as the dependent variable showed a significant main effect, F(2, 207) = 9.03, $p < 0.001$, ^{2}p = 0.08. Planned contrast analysis indicated that the algorithm group' s moral punishment desire (M = 3.94, SD = 1.45) was significantly lower than both the algorithm-with-free-will group' s (M = 4.56, SD = 1.62) and the human group' s (M = 4.98, SD = 1.35), ps < 0.05, but no significant difference existed between human and algorithm-with-free-will groups, p = 0.10 (see Figure 3).

Figure 3. Moral punishment desire scores across different discrimination agent groups.

Controlling for gender (male = 1, female = 2), age, and ethnicity (Han = 1, ethnic minority = 2) as covariates in an ANCOVA showed no significant effects: gender F(1, 204) = 0.34, p = 0.559; age F(1, 204) = 1.13, p = 0.289; ethnicity F(1, 204) = 1.72, p = 0.191. The group effect remained significant, F(2, 204) = 9.59, p < 0.001, ^{2}p = 0.09.

5.3 Discussion

By directly manipulating belief in algorithms' free will and comparing human, algorithm-with-free-will, and algorithm groups, Experiment 4 found that the difference between human and algorithm-with-free-will groups was smaller than between human and algorithm groups, further verifying that belief in free will is the mechanism causing different desires to punish different agents (human vs. algorithm). Since algorithmic discrimination elicits less moral punishment desire than human discrimination because people believe algorithms possess less free will, might individual differences in anthropomorphic tendency moderate this effect? To address this question, Experiment 5 will continue exploring boundary conditions by examining the potential moderating effect of anthropomorphic tendency.

6. Experiment 5: The Moderating Role of Anthropomorphic Tendency

Theoretically, humans possess greater free will than algorithms. Do people with higher anthropomorphic tendencies attribute more free will to algorithms, thereby affecting the relationship between discrimination agent and moral punishment desire? Experiment 5 addresses this question while also enriching the discrimination type by focusing on age discrimination.

6.1 Method

6.1.1 Participants Based on the independent samples t-test for Experiment 5, G*Power 3.1 software (Faul et al., 2007) indicated that at least 172 participants were needed for 90% statistical power at = 0.05 with medium effect size (d = 0.5). We recruited participants through Credamo, randomly assigning them to human or algorithm groups, real-time excluding those who failed attention checks. The final sample included 199 valid participants (88 females) aged 18-41 years (M = 28.64, SD = 4.68), with 101 in the human group and 98 in the algorithm group. All participants carefully read instructions and provided informed consent; those with valid data received compensation.

6.1.2 Design and Procedure Experiment 5 used a single-factor between-subjects design (human vs. algorithm). Participants first read an age discrimination scenario (underlined for human group, bracketed for algorithm group): "In last year' s autumn recruitment, HR Manager Zhao Guang of Weilan Company was responsible for (using an algorithm to) conduct hiring. After recruitment, the company discovered that Zhao Guang (the algorithm) exhibited age bias when screening resumes, filtering out all applicants older than 35, even though most positions had no rigid age requirements. This prevented many talented and capable applicants over 35 from obtaining jobs at the company." This scenario was adapted from Bigman et al. (2020).

After reading the scenario and passing the attention check, both groups reported their desire for moral punishment toward human or algorithmic discrimination using the same three items as Experiment 1 (Hofmann et al., 2018), = 0.84. Next, we measured belief in free will regarding the age-discriminatory human or algorithm using the same five items as Experiment 2 (Nadelhoffer et al., 2014), = 0.87. Then participants completed the Individual Differences in Anthropomorphism Questionnaire (IDAQ; Waytz et al., 2014) with 15 items (= 0.87), such as "To what extent does an ordinary fish have free will?" using an 11-point scale (0 = not at all to 10 = very much), with higher scores indicating stronger anthropomorphic tendency. Finally, participants reported demographic information on gender and age.

6.2 Results

6.2.1 Effect of Discrimination Agent on Desire for Moral Punishment Independent samples t-test results showed that the human group' s moral punishment desire score (M = 5.29, SD = 0.97) was significantly higher than the algorithm group' s (M = 4.61, SD = 1.32), t(197) = 4.17, p < 0.001, Cohen' s d = 0.59. Controlling for gender and age as covariates, ANOVA results still showed the algorithm group' s score significantly lower than the human group' s, F(1, 195) = 17.28, p < 0.001, ^{2}p = 0.08.

6.2.2 Mediating Effect of Belief in Free Will To again verify the psychological mechanism, we used Hayes' s (2013) PROCESS macro (Model 4) with discrimination agent as the independent variable (human group = -1, algorithm group = 1), belief in free will as the mediator, and moral punishment desire as the dependent variable. With 5000 Bootstrap samples and bias-corrected 95% CI, results showed a significant indirect effect of -0.11, 95% CI [-0.23, -0.01], not containing zero. After controlling for the mediator, the direct effect of discrimination agent remained significant (-0.23, 95% CI [-0.42, -0.04]), indicating partial mediation.

6.2.3 Moderating Effect of Anthropomorphic Tendency Examining the interaction between discrimination agent (human = -1, algorithm = 1) and anthropomorphic tendency on moral punishment desire revealed a significant interaction (b = 0.16, SE = 0.06, t = 2.70, p = 0.008). The human group’ s moral punishment desire was significantly higher than the algorithm group’ s (b = -0.34, SE = 0.08, t = -4.18, p < 0.001), while anthropomorphic tendency alone had no significant effect (b = 0.01, SE = 0.06, t = 0.08, p = 0.937). The model explained significant variance, adjusted $R^2 = 0.10$, $\Delta R^2 = 0.03$, F(3, 195) = 8.40, p < 0.001. As shown in Figure 4, simple slope analysis indicated that under low anthropomorphic tendency, discrimination agent significantly affected moral punishment desire (b = -0.57, SE = 0.12, t = -4.82, p < 0.001), whereas under high anthropomorphic tendency, the effect was not significant (b = -0.12, SE = 0.12, t = -1.05, p = 0.295). Moreover, in the algorithm group, participants’ anthropomorphic tendency positively correlated with belief in the algorithm’ s free will, r = 0.20, p = 0.044.

Figure 4. The moderating role of anthropomorphic tendency.

6.3 Discussion

Experiment 5 further explored boundary conditions and found that anthropomorphic tendency moderates the effect of discrimination agent on moral punishment desire. Among participants low in anthropomorphic tendency, the algorithm group’ s moral punishment desire was significantly lower than the human group’ s; among those high in anthropomorphic tendency, no significant difference existed between groups. Experiment 5 also again verified the mediating role of belief in free will: people have less desire to punish algorithmic discrimination because they perceive algorithms as having less free will than humans. In Experiment 6, we directly manipulated the algorithm’ s degree of anthropomorphism to further verify this moderating effect.

7. Experiment 6: The Moderating Role of Algorithm Anthropomorphism

To more directly examine whether algorithm anthropomorphism moderates the effect of discrimination agent on moral punishment desire, Experiment 6 manipulated algorithm anthropomorphism through text, comparing participants’ desire to punish human discrimination, anthropomorphized algorithmic discrimination, and non-anthropomorphized algorithmic discrimination.

7.1 Method

7.1.1 Participants Based on Experiment 6’s single-factor three-level between-subjects design, G*Power 3.1 software (Faul et al., 2007) indicated that at least 207 participants were needed for 90% statistical power at = 0.05 with medium effect size (f = 0.25). We recruited participants through Credamo, randomly assigning them to human, anthropomorphized algorithm, or non-anthropomorphized algorithm groups, real-time excluding those who failed attention checks. The final sample included 207 valid participants (127 females) aged 19–59 years (M = 29.53, SD = 6.62), with 69 participants in each group. All participants carefully read instructions and provided informed consent; those with valid data received compensation.

7.1.2 Design and Procedure Experiment 6 used a single-factor three-level between-subjects design (human, anthropomorphized algorithm, non-anthropomorphized algorithm). As in Experiment 4, both algorithm groups received an explanation and examples of "algorithm" before the discrimination scenario to ensure understanding. Participants then read a gender discrimination scenario (underlined for human group, bracketed for both algorithm groups): "In last year' s autumn recruitment, HR Manager Zhao Guang of Weilan Company was responsible for (using algorithm 'Qizhi' /using algorithm 'R2000') to conduct hiring. After recruitment, the company discovered that Zhao Guang (Qizhi/R2000) exhibited gender bias when screening resumes, showing clear preference for males and filtering out many female applicants, even though most positions had no rigid gender requirements. This prevented many talented and capable female applicants from obtaining jobs at the company." This scenario was adapted from Bigman et al. (2020).

As the anthropomorphism manipulation, algorithm group participants read additional descriptions after the scenario. The anthropomorphized algorithm group read:

"Hi! My name is Qizhi. I am a new type of recruitment algorithm. I have analyzed all resumes submitted to the company over the past ten years to learn how to identify the best applicants. I can carefully review applicants' resumes and backgrounds, accurately predict which employees will meet job requirements and fit corporate culture in the future, identify the best applicants, and help companies select the best employees."

The non-anthropomorphized algorithm group read:

"Algorithm R2000 Introduction: R2000 is a new type of recruitment algorithm. R2000 has analyzed all resumes submitted to the company over the past ten years to learn how to identify the best applicants. R2000 can carefully review applicants' resumes and backgrounds, accurately predict which employees will meet job requirements and fit corporate culture in the future, identify the best applicants, and help companies select the best employees."

This manipulation, adapted from prior research on anthropomorphism (e.g., Hur et al., 2015; May & Monga, 2014), effectively increases anthropomorphism by giving a nonhuman agent a human name and first-person description. The descriptions were otherwise identical. To check effectiveness, participants rated the algorithm' s anthropomorphism ("To what extent does algorithm 'Qizhi'

/ ‘R2000’ remind you of human characteristics?” , 1 = not at all to 7 = very much), adapted from Hur et al. (2015).

After reading these materials and passing manipulation and attention checks, all three groups reported their desire for moral punishment toward human or algorithmic discrimination using the same items as Experiment 4 (Hofmann et al., 2018), with evaluation targets being Zhao Guang/Qizhi/R2000, = 0.88. Finally, participants reported demographic information on gender and age.

7.2 Results

7.2.1 Anthropomorphism Manipulation Check Independent samples t-test results showed that the anthropomorphized algorithm group’ s anthropomorphism rating (M = 5.43, SD = 0.88) was significantly higher than the non-anthropomorphized algorithm group’ s (M = 4.83, SD = 1.25), t(136) = 3.31, p = 0.001, Cohen’ s d = 0.56, indicating the manipulation was effective.

7.2.2 Effect of Algorithm Anthropomorphism A one-way ANOVA on moral punishment desire showed a significant main effect of discrimination agent, F(2, 204) = 12.60, p < 0.001, ^{2}p = 0.11. Planned contrast analysis indicated that the human group’ s moral punishment desire score (M = 5.52, SD = 1.19, 95% CI [5.24, 5.81]) was significantly higher than both the anthropomorphized algorithm group’ s (M = 4.97, SD = 1.27, 95% CI [4.66, 5.27]) and the non-anthropomorphized algorithm group’ s (M = 4.43, SD = 1.35, 95% CI [4.11, 4.76]), and the anthropomorphized algorithm group’ s score was significantly higher than the non-anthropomorphized algorithm group’ s, ps < 0.05 (see Figure 5). This shows that for the same gender discrimination, the human group (vs. algorithm groups) elicited stronger moral punishment desire, and anthropomorphizing the algorithm (vs. non-anthropomorphized algorithm) also significantly increased moral punishment desire, demonstrating that algorithm anthropomorphism moderates the effect of discrimination agent on moral punishment desire.

Figure 5. Moral punishment desire scores across different discrimination agent groups.

7.3 Discussion

Building on Experiment 5, Experiment 6 directly manipulated algorithm anthropomorphism and again verified its moderating effect. Specifically, anthropomorphizing an algorithm significantly increased participants’ desire to morally punish it, consistent with our predictions. However, a significant difference remained between anthropomorphized algorithm and human groups, possibly because text-based anthropomorphism, while increasing overall anthropomorphism, still left a gap compared to human levels (especially regarding belief in free will).

8. General Discussion

This study examined whether differences exist in people' s desire to morally punish human versus algorithmic discrimination and explored the underlying mechanisms and boundary conditions. Across six experiments, we found that relative to human discrimination, algorithmic discrimination elicits less desire for moral punishment. Belief in free will is the underlying mechanism, and this difference is moderated by anthropomorphism. Specifically, by presenting participants with identical discriminatory behavior by humans or algorithms and measuring their moral punishment desire, we found less desire to punish algorithmic discrimination, a robust effect across experiments (1-6). By measuring belief in free will (Experiment 2) and manipulating participants' belief in free will (Experiment 3) and belief in algorithms' free will (Experiment 4), we found that belief in free will is the mechanism: people perceive algorithms as having less free will than humans, thus showing less desire to punish algorithmic discrimination (Experiments 2-4). By measuring individual anthropomorphic tendency (Experiment 5) and manipulating algorithm anthropomorphism (Experiment 6), we found moderation effects. Regarding individual tendency, those low in anthropomorphism showed less desire to punish algorithmic than human discrimination, whereas those high in anthropomorphism showed no significant difference (Experiment 5). Regarding algorithm anthropomorphism, the more anthropomorphic the algorithm, the smaller the difference in punishment desire between algorithmic and human discrimination. The study examined various discrimination types—gender (Experiments 1, 6), educational background (Experiment 2), ethnicity (Experiments 3, 4), and age (Experiment 5)—and diverse samples, including nationwide participants from Credamo (Experiments 1, 2, 5, 6) and university students (Experiments 3, 4). This diversity in scenarios and participants ensures robust findings.

8.1 Differences in Reactions to Humans and Algorithms

This study found that when humans and algorithms commit identical discriminatory acts, people exhibit different desires for moral punishment—less for algorithmic discrimination. First, this aligns with Bigman et al. (2020), who found less moral outrage toward algorithmic than human discrimination, primarily examining the emotional aspect of moral reactions. Our study extends this by focusing on behavioral tendency—desire for moral punishment. While moral outrage and punishment desire are correlated, they are not equivalent; punishment desire has unique research value. Although Bigman et al. (2020) included some punishment-related items in their Study 5 (e.g., "The discriminatory algorithm should be discontinued," "The company using the algorithm should apologize"), these were not exclusively algorithm-focused and did not yield significant differences. Second, regarding mechanisms, Bigman et al. (2020) focused on motivational mechanisms, whereas we replicated and extended the free will mechanism, which is arguably more fundamental because possessing free will is a necessary condition for judging motives (Laming, 2004). Third, regarding moderation, Bigman et al. (2020) did not explore moderators but suggested anthropomorphic tendency might play a moderating role; we empirically verified this through two studies examining both individual anthropomorphic tendency and algorithm anthropomorphism.

Second, this finding extends moral punishment research to AI. Previous studies focused primarily on humans, exploring influences limited to human-related variables (e.g., Hofmann et al., 2018). Our study expands the scope of moral punishment research to include AI as a potential discrimination agent, revealing that discrimination agent (human vs. algorithm) significantly affects punishment desire.

Third, regarding attitudes toward algorithmic decision-making, our findings provide new evidence that may contradict algorithm aversion research. Algorithm aversion finds that people psychologically distrust algorithms (Meehl, 1954); despite algorithms' superior computational abilities, people generally prefer human decisions (Dietvorst et al., 2015, 2018). Algorithm errors are less tolerated than human errors (Prahl & Van Swol, 2017), especially for moral decisions, where people oppose machines replacing humans because machines lack necessary mental capacities for moral judgment (Bigman & Gray, 2018). Our study found less desire to punish algorithmic than human discrimination when both commit identical acts, which seems inconsistent with algorithm aversion findings. However, this does not completely overturn algorithm aversion research: people may still be unwilling to let algorithms make moral decisions (Bigman & Gray, 2018), but their negative reactions after decisions are made may be weaker. In other words, any "appreciation" for algorithmic moral decision-making may be limited to post-decision contexts rather than pre-decision preferences.

Algorithm appreciation seems to require objectively strong tasks for people to prefer algorithmic decisions (Logg et al., 2019). While algorithms may be seen as more accurate and impartial due to their computational power and objectivity (Grove et al., 2000), algorithmic discrimination is not uncommon and remains dangerous (e.g., Borgesius, 2018). When facing algorithmic discrimination, people are less angry (Bigman et al., 2020) and less desirous of punishment, potentially reducing vigilance, increasing habituation, and rationalization, leading to more severe discrimination problems.

8.2 Differences in Perception of Humans and Algorithms

This study also found differences in how people perceive humans and algorithms: algorithms are seen as having less free will than humans, which explains differences in moral punishment desire. This aligns with prior research on mind perception of AI (e.g., Gray et al., 2007). People attribute moderate agency to robots (Epley & Waytz, 2010)—the mental capacity for autonomous, planned action is inferior to humans. This is similar to our finding that algorithms are perceived as having less free will, confirming that machine-like nonhuman agents have some mental capacities but fall far short of human levels. This also reaffirms the close link between belief in free will and moral punishment. People' s reduced desire to punish algorithmic discrimination is related to algorithms' lower perceived free will. The capacity for free action and alternative choice is crucial for moral responsibility and punishment (e.g., Shariff et al., 2014; Clark et al., 2014), and beliefs about free will' s existence affect punishment of transgressors (Aspinwall et al., 2012). Our findings are consistent with this research and, more broadly, with the view that mental capacity is a prerequisite for moral responsibility (Gray et al., 2012). Although competing hypotheses unrelated to free will exist (e.g., human behavior is easier to explain than opaque algorithms, punishing algorithms is impractical), our two moderation studies on belief in free will (Experiments 3-4) and two on anthropomorphism (closely related to free will; Experiments 5-6) repeatedly verified our proposed mechanism, largely ruling out these alternatives.

Differences in human-algorithm perception extend beyond mind perception to behavior perception. Although humans and algorithms commit identical discriminatory acts with equivalent impact in our scenarios, people may perceive their severity differently. Bonezzi and Ostinelli (2021) found that algorithms committing gender and racial discrimination were perceived as less biased than humans because algorithms were seen as using rules and procedures rather than attending to individual characteristics. While we did not directly measure perceived severity, our results are consistent with this finding.

Our study also found that anthropomorphic tendency moderates desire to punish algorithmic discrimination, which aligns with research showing that possessing complete human-like mind is necessary for moral responsibility (Bigman & Gray, 2018; Gray et al., 2012). Anthropomorphizing AI is an industry trend (Broadbent, 2017) and key to AI ethics (Bostrom & Yudkowsky, 2011). Using anthropomorphized algorithms as discrimination agents not only helps directly examine perceptual differences between humans and algorithms but may also simulate psychological processes of human discrimination, aiding our understanding of discrimination.

8.3 Limitations and Future Directions

Our research demonstrates differences in moral punishment desire toward human versus algorithmic discrimination, explained by differential perceptions of free will and moderated by anthropomorphism. However, we acknowledge several limitations that point to future directions.

First, some experimental design details have shortcomings. Experiments 1, 2, 3, and 5 used Hofmann et al.' s (2018) moral punishment items directly, and the wording "moral punishment" and "immoral behavior" might have influenced responses when applied to algorithmic discrimination. We addressed this in Experiments 4 and 6 with more neutral wording, which did not affect main results. Second, participants' familiarity and understanding of algorithms could affect results; we addressed this in Experiments 4 and 6 but future research should continue examining this issue. Third, Experiments 2-4 had a confound in how discrimination agents were described: "HR Manager Li Yuan/Zhang Pei/Zhao Guang" are concrete, specific agents, while "algorithm" is a broad concept without a specific referent. Research shows people prefer punishing identified over unidentified transgressors (e.g., Small & Loewenstein, 2005), so this difference could affect results. We addressed this by using more abstract descriptions in Experiments 1 and 4 and more concrete descriptions in Experiment 6, obtaining similar results.

Second, other mechanisms may explain differences in moral punishment desire. While we focused on belief in free will, human-algorithm mind perception differences may also involve consciousness (McDermott, 2007), intentionality (Weisman et al., 2017), and capacity for emotional experience (Epley & Waytz, 2010). Future research could examine these variables more systematically and compare their influences. Beyond mind perception, as noted, people may perceive severity differently (Bonezzi & Ostinelli, 2021), so future studies could incorporate this into mechanism exploration. Additionally, people' s everyday "moral punishment threshold" for nonhuman agents may affect punishment desire—there may be a threshold that even severe moral violations cannot exceed. Future research could examine such threshold issues.

Finally, moral punishment desire may relate to the target of moral attribution. We examined punishment desire toward algorithms and found it lower than toward humans. Another possible reason is that people are unwilling to attribute moral responsibility to algorithms. Previous research shows people are unwilling to let machines make moral decisions (Bigman & Gray, 2018). When machines do make moral decisions, who should bear moral responsibility—the algorithm? Its designers? The company investing in it? Or regulatory agencies? As AI applications and algorithmic decision-making become more prevalent, increasing moral condemnation of AI may increase its responsibility and punishment, potentially creating a "scapegoating" possibility—designers, companies, or governments using AI to evade responsibility for their errors (Bigman et al., 2019). Punishment following moral attribution would also be affected, both because of the correlation between responsibility and punishment (if people attribute responsibility to humans, punishment of AI decreases) and because punishing the "people behind the AI," especially those using it illegally or improperly, may be more reasonable and practically meaningful. Although AI like algorithms cannot yet be full moral agents, we may still punish them when they err—e.g., kicking a vacuum robot or smashing a phone when an algorithm fails. Of course, the core of discussing algorithms is human welfare. Therefore, the distribution of moral responsibility and punishment between humans and AI warrants further research.

9. Conclusion

This study concludes: First, relative to human discrimination, people have less desire to morally punish algorithmic discrimination. Second, the underlying mechanism is that people perceive algorithms as lacking free will compared to humans. Third, the stronger an individual' s anthropomorphic tendency or the more anthropomorphic the algorithm, the stronger the desire to morally punish it.

References

Al Ramiah, A., Hewstone, M., Dovidio, J. F., & Penner, L. A. (2010). The social psychology of discrimination: Theory, measurement and consequences. In L. Bond, F. McGinnity, & H. Russell (Eds.), *Making equality count: Irish and international research measuring equality and discrimination* (pp. 84–112). Dublin, Ireland: Liffety Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. Retrieved June 18, 2021, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J., Mattu, S., & Larson, J. (2015). The tiger mom tax: Asians are nearly twice as likely to get a higher price from Princeton Review. Retrieved June 18, 2021, https://www.ProPublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review

Aspinwall, L. G., Brown, T. R., & Tabery, J. (2012). The double-edged sword: Does biomechanism increase or decrease judges' sentencing of psychopaths? *Science, 337*, 846–849.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81.

Batson, C. D., Kennedy, C. L., Nord, L. A., Stocks, E. L., Fleming, D. Y. A., Marzette, C. M., ···& Zerger, T. (2007). Anger at unfairness: Is it moral outrage?. *European Journal of Social Psychology, 37*(6), 1272–1285.

Baumeister, R. F. (2008). Free will in scientific psychology. *Perspectives on Psychological Science, 3*(1), 14–19.

Baumeister, R. F. (2014). Constructing a scientific theory of free will. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 4. Free will and moral responsibility* (pp. 235–255). Boston Review.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin, 35*(2), 260–268.

Baumeister, R. F., Stillwell, A., & Wotman, S. R. (1990). Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger. *Journal of Personality and Social Psychology, 59*(5), 994–1005.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21-34.

Bigman, Y. E., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2020). Algorithmic discrimination causes less moral outrage than human discrimination [Preprint]. PsyArXiv.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365-368.

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination?. *Journal of Experimental Psychology: Applied.* Advance online publication.

Borgesius, F. Z. (2018). *Discrimination, artificial intelligence, and algorithmic decision-making.* Retrieved June 18, 2021, from https://rm.coe.int/discrimination-artificial-intelligence-andalgorithmic-decision-making/1680925d73

Bostrom, N., & Yudkowsky, E. (2011). The ethics of Artificial Intelligence. In K. Frankish (Ed.). *Cambridge handbook of artificial intelligence.* Cambridge: Cambridge University Press.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*, 627-652.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology, 106*(4), 501-513.

Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Measuring prejudice, stereotypes and discrimination. *The SAGE handbook of prejudice, stereotyping and discrimination*, (pp. 45-62). Thousand Oaks, CA: Sage.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved June 18, 2021, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies, 1*(1), 92-112.

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

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155-1170.

Epley, N., & Waytz, A. (2010). Mind perception. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 498-541). John Wiley & Sons, Inc.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864-886.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Ferrero, F., & Barujel, A. G. (2019, October). Algorithmic driven decision-making systems in education: Analyzing bias from the sociocultural perspective. In *2019 XIV Latin American Conference on Learning Technologies (LACLO)* (pp. 166-173), San Jose Del Cabo, Mexico.

Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence, 283*, 103261.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101-124.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*, 98-116.

Hao. K. (2019). AI is sending people to jail—and getting it wrong. Retrieved June 18, 2021, from https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Harvey, C. R., Rattray, S., Sinclair, A., & Van Hemert, O. (2017). Man vs. machine: Comparing discretionary and systematic hedge fund performance. *The Journal of Portfolio Management, 43*(4), 55-69.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York: Guilford Press.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin, 44*(12), 1697-1711.

Hur, J. D., Minjung, K., & Wilhelm, H. (2015). When temptations come alive: How anthropomorphism undermines self-control. *Journal of Consumer Research, 42*(2), 340-358.

Kim, T. W., & Duhachek, A. (2020). Artificial intelligence and persuasion: A construal-level account. *Psychological Science, 31*(4), 363-380.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science, 65*(7), 2966-2981.

Laming, D. (2004). *Understanding human motivation: What makes people tick?* Malden, MA: Blackwell.

Leo, X., & Huh, Y. E. (2020). Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior, 113*(4), 106520.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90-103.

Mackenzie, M. J., Vohs, K. D., & Baumeister, R. F. (2014). You didn' t have to do that: Belief in free will promotes gratitude. *Personality and Social Psychology Bulletin, 40*(11), 1423-1434.

May, F., & Monga, A. (2014). When time has a will of its own, the powerless don' t have the will to wait: Anthropomorphism of time can decrease patience. *Journal of Consumer Research, 40*(5), 924-942.

McDermott, D. (2007) Artificial intelligence and consciousness. In Zelazo, P., Moscovitch, M. & Thompson, E. (Eds.), *Cambridge Handbook of Consciousness* (pp. 117-150). Cambridge: Cambridge University Press.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota Press.

Mori, M. (1970). The uncanny valley. *Energy, 7*, 33-35.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition, 25*, 27-41.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*, 561-584.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663-685.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447-453.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted?. *Journal of Forecasting, 36*(6), 691-702.

Rigoni, D., Kühn, S., Gaudino, G., Sartori, G., & Brass, M. (2012). Reducing self-control by weakening belief in free will. *Consciousness and Cognition, 21*(3), 1482-1490.

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., ···& Vohs, K. D. (2014). Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science, 25*(8), 1563–1570.

Sinnott-Armstrong, W. (2014). *Moral psychology: Free will and moral responsibility.* MIT Press.

Small, D. A., & Loewenstein, G. (2005). The devil you know: The effects of identifiability on punishment. *Journal of Behavioral Decision Making, 18*(5), 311–318.

Tang, S., Koval, C. Z., Larrick, R. P., & Harris, L. (2020). The morality of organization versus organized members: Organizations are attributed more control and responsibility for negative outcomes equivalent members. *Journal of Personality and Social Psychology, 119*(4), 901–919.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49–54.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232.

Waytz, A., Cacioppo, J., & Epley, N. (2014). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Wegner, D. M., & Gray, K. (2017). *The mind club: Who thinks, what feels, and why it matters.* Penguin.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people’ s conceptions of mental life. *Proceedings of the National Academy of Sciences, 114*(43), 11374–11379.

Wen, Z., Zhang, L., Hou, J., & Liu, H. (2004). Testing and application of the mediating effects. *Acta Psychologica Sinica, 36*(5), 614–620.

Appendices

1. Free Will Inventory (Human Group Example)

Please indicate your agreement with the following statements (1 = strongly disagree to 7 = strongly agree):

1. Li Yuan has the capacity to make different choices.
2. Li Yuan has free will.

3. The hiring decisions are completely up to Li Yuan.
4. Li Yuan has complete control over his decisions and actions at a fundamental level.
5. Li Yuan has free will, even if his choices are constrained by external circumstances.

2. Free Will Belief Manipulation Materials (Experiment 3)

Science Shows Free Will Does Not Exist
By Dr. Chris Wellington, Ph.D.

All human behavior is the product of simple physical processes in the brain. Conscious thoughts, memories, emotions, and choices people experience are merely chemical reactions and electrical impulses. Because scientists can predict all bodily reactions using scientific laws, given enough information they will one day be able to predict all human behavior. Free will is an illusion.

Modern science has shown that humans, like all other living things, are governed by the same processes. From bacteria to humans, everything operates through closely related processes at the chemical level. Similarly, evolutionary theory proves that all plants and animals were formed through the same natural methods. Therefore, although human complexity may differ, their bodies and brains are no different from anything else. No soul or free will is needed to explain our behavior.

Experience tells people they can act as they wish, so most believe they have free will. But where do these desires and impulses come from? In fact, people often don’ t know why they do many things. In many cases, people find reasons for their behavior but are largely unaware of the forces driving them. Actions are determined not only by conscious thoughts but also by information processed by the brain outside awareness. These processes can be broken down into simple, predictable processes described by chemists and physicists. Although people appear to have free will, their behavior, choices, and even thoughts are predetermined by their bodies, environment, and scientific laws.

Science Shows Free Will Exists
By Dr. Chris Wellington, Ph.D.

Most human behavior is the product of decisions and free will. People usually control their conscious thoughts, consider different possibilities, and deliberate about their freely chosen memories. These factors have been shown to be primary influences on people’ s choices and are directly under personal control. Moreover, scientists cannot yet predict all human bodily reactions using scientific laws. For this reason, scientists and philosophers generally agree that free will is not an illusion.

Modern science has shown that the human brain is the most complex biological entity known. Everything else, from bacteria to nonhuman animals, operates through far simpler processes at the brain and cognitive levels. Humans have the capacity for abstract thought, meaning their minds are not limited to the here and now but can reach far into the past and future. People' s choices are guided by this conscious abstract thinking because they can consider future consequences or past mistakes. Therefore, free will is essential for explaining human behavior.

Everyday experience tells people they can act as they wish, so most realize they have free will. People are usually quite aware of why they perform particular behaviors or make certain decisions. When asked to explain their choices, they can easily identify the factors leading to them because any behavior or choice ultimately depends on the person's direct conscious control. In summary, science has shown that free will is something everyone possesses and is an important part of human nature.

3. Algorithm Explanation and Examples (Experiments 4, 6)

An algorithm, in mathematics and computer science, refers to a well-defined, finite sequence of steps or orders that a computer can execute, commonly used for calculation, data processing, and automated reasoning. With AI development, algorithmic decision-making is increasingly used to assist or replace human decisions, such as in credit approval, talent recruitment, and criminal risk assessment.

https://www.merriam-webster.com/dictionary/algorithm
https://zh.wikipedia.org/zh-cn/%E7%AE%97%E6%B3%95

4. Individual Differences in Anthropomorphism Questionnaire

To what extent do you think the following descriptions are true? Please select from 0 (not at all) to 10 (very much):

1. To what extent do technological devices and machines used in manufacturing, entertainment, and production processes (e.g., cars, computers, TVs) have intentions?
2. To what extent does an ordinary fish have free will?
3. To what extent does an ordinary mountain have free will?
4. To what extent can a TV experience emotions?
5. To what extent does an ordinary robot have consciousness?
6. To what extent is a cow intentional?
7. To what extent does a car have free will?
8. To what extent does the ocean have consciousness?
9. To what extent does an ordinary computer have its own mind?
10. To what extent can a cheetah experience emotions?
11. To what extent can the environment experience emotions?
12. To what extent does an ordinary insect have its own mind?
13. To what extent does a tree have its own mind?
14. To what extent is the wind intentional?
15. To what extent does an ordinary reptile have consciousness?