Algorithmic Discrimination Causes Fewer Positive Expectations of Punishment Effects Than Human Discrimination

Yueqi Guo[1,a,*]

[1]Zhejiang Sci-Tech University, Hangzhou, 310018, China
a. yueqi0427@gmail.com
**corresponding author*

Abstract: People have different attitudes toward algorithmic discrimination than human discrimination. This study collected 179 data through an online experiment, comparing people's expectations of the effect of punishing algorithms and humans due to discriminatory behaviors in recruitment. It turns out that people have fewer positive expectations about the effects of punishing algorithms than punishing humans. This may be because people don't trust algorithms. The findings contribute to a better understanding of people's responses to algorithmic discrimination and provide new evidence for algorithm aversion.

Keywords: algorithm, algorithm discrimination, moral punishment, punishment effect

1. Introduction

Although times have progressed, our lives are still full of stereotypes, prejudice, and discrimination. Regardless of the decision's importance, people will inevitably use the stereotypes formed over the years to judge and make biased decisions. Prejudice is an inherent negative characteristic of human beings that must be overcome. Prejudice can hurt people's lives. For example, it can interfere with employment decisions, affect the quality of education or health care people receive, and it is still a serious and common problem [1]. People cannot eliminate their prejudices, but with the advancement of technology, people see another way to solve this problem. Algorithms are increasingly used in decision-making, not only in daily life, such as what movie to watch and what product to buy but also in many critical decisions, such as loan approval and employment recruitment [2]. Algorithmic decision-making has apparent benefits. It can consider more factors than humans and avoid human subjectivity to a certain extent. People think that algorithms are more accurate, fair, rational, and neutral decision-makers than humans [1, 3]. Unfortunately, however, as these systems became commercialized, it was discovered that these applications contained some deviations. Prejudice, or discrimination, is a negative attitude or behavior toward a particular category of individuals or groups that unconsciously influences the decision-making process, leading to differential treatment of others. Like people, algorithms can also be affected by bias to make discriminatory behaviors, and their decisions will be biased toward specific groups of people [1-3].

Algorithms can be racially discriminatory, such as the COMPAS algorithm developed by Northpointe, which US courts use to make pretrial detention and release decisions. It incorrectly predicted that African Americans have a higher risk of recidivism. Similarly, AI systems are biased against darker-skinned contestants in beauty pageants, and facial recognition software overpredicts Asian blinks [2]. Algorithms also have gender discrimination. For example, Google's targeted advertising algorithm will show fewer advertisements related to high-paying jobs for women. A similar situation exists in the recruitment advertising algorithm in the STEM field [3].

Why might an algorithm be biased, even though it is believed to be more impartial, trustworthy, and objective than humans? This is because the algorithm's design, data, and the rules by which the algorithm processes and applies the data are all created by biased humans, so algorithms are as biased as humans [4]. There are two potential sources of unfairness during algorithm training: data bias and algorithm design [2]. Take sexism as an example. Numerous studies have shown that people's gender-stereotyped expectations can affect how we judge women's and men's abilities. People judge academic performance, resumes, cover letters, design creativity and more based on whether the person being judged has a female or male name. These evaluation differences caused by gender stereotypes will significantly impact the career development and income levels of men and women. They may accumulate into profound gender inequality throughout life [5]. Algorithms may be designed by programmers who are sexist, or algorithms may be designed to mimic existing sexist human decisions, or algorithms may learn data from existing sexist choices, all of which lead to algorithms also being sexist [6].

When people face the immoral behavior of discrimination, they will have a moral reaction, emotional anger, and tend to condemn morally, hoping that the discriminator will be punished [3]. Previous studies have found that compared with human discrimination, people feel less moral outrage and moral punishment desire for algorithmic bias, with motivational attribution and free will beliefs mediating [3, 6]. But the penalty effect may be another explanatory factor, and the punishment of the algorithm does not promote its progress. Moral punishment is to sanction unethical behavior and deter future unethical behavior. People hope to identify and punish the perpetrators of unethical behavior. Punishment is the corrective measure necessary to maintain the integrity of the established moral system [7]. Algorithms lack complete thinking, and people have different perceptions of humans and algorithms. Mind perception is related to Experience and Agency; people think that algorithms lack emotional experience, compassion, and the psychological ability to plan actions independently [8, 9]. It is believed that algorithms have less free will than humans, so punishing algorithms cannot make them understand and reflect on punishment. Punishment for unethical behavior of algorithms may not promote positive changes [3]. Perhaps because punishment has lost its positive facilitation, people think less about punishing algorithms. Accordingly, this paper hypothesizes that people have less positive expectations about the effects of punishing algorithmic discrimination than of punishing human discrimination.

The gender discrimination of algorithms in recruitment has been a broad concern, such as Amazon's recruitment algorithm will lower the score of resumes containing the word "female" [10]. This study chooses gender discrimination in recruitment as the experimental situation. It explores the difference between people's expectations of the effects of punishing algorithmic discrimination and human discrimination through online experiments and filling out questionnaires.

2. Materials and Methods

2.1. Participants

This study first used G*Power3.1 software [11] to calculate the required sample size. For the two-factor ANOVA used in this experiment, the medium effect size f=0.25, the significance level α=0.05, and the number of groups is 6. At least 178 participants are needed to achieve 85% statistical power. This study recruited participants through the Credamo platform. Considering that there may be invalid data that did not fill in seriously or failed the attention check, a total of 230 pieces of data were collected. After deleting invalid data and participants who did not answer seriously, the remaining 179 were valid Data, and the effective recovery rate was 77.83%. The mean age of effective participants was 29.61 years old (SD=6.84 years old), and 108 were women (60.3%).

2.2. Procedure

This experiment is a 2 (discriminatory behavior subject: human/algorithm) × 3 (punishment effect learning: effective/ineffective/no-learning) experimental design among participants. All participants are randomly assigned to one of the six groups.

First, participants are randomly divided into two groups (humans/algorithms) who will see the material about human/algorithmic sexist hiring decisions in the experiment. In the first stage, people will be randomly divided into three groups (effective/ineffective/no-learning), the no-learning group will directly enter the second stage, and people in the effective group and ineffective group will see material about A/B/C/D four companies were gender discriminatory in their hiring last year. Material adapted from the study by Xu et al. [3].

After reading, the participants were asked whether they would punish the discriminatory person/algorithm, for example, “If you could choose to punish the manager of company A, would you choose to do so?” If the participants chose No, the two groups, everyone will see a prompt: They are not penalized, and this year’s hiring is still discriminatory, such as “You didn’t punish the manager. He still has a gender bias in this year’s hiring screening”; if participants choose Yes, the effective group will see prompt: This year’s hiring is no longer discriminatory, such as “You punished the manager. He is no longer gender biased in this year’s hiring screening,”; while the ineffective group will see the prompt: This year’s hiring is still discriminatory, such as “ You punished the manager. He still has a gender bias in this year’s hiring screening.” After four rounds of A/B/C/D learning, the two groups entered the second stage.

In the second stage, participants read the discrimination materials of E company, and I used three self-made questions to measure their expectations for the punishment effect, “To what extent do you think the manager/algorithm will change?”, “To what extent do you think your punishment will have an impact on the manager/algorithm?”, “How likely do you think the manager/algorithm is to be still discriminatory?” All three questions are 7-point Likert scale (1=very unlikely, 7=very likely), the third question is reverse scoring, and the average score of the three questions is used as the score of expectations for the punishment effect. Higher scores indicate that participants are more optimistic about expectations for the punishment effect. The internal consistency reliability of this measure is Cronbach α=0.95.

Participants then read more detailed material on sexism in hiring, adapted from Bigman et al.[6]. After reading, participants filled out the Moral Punishment Desire Questionnaire and the Free Will Belief Questionnaire, adapted from Xu et al.[3]. The Moral Punishment Questionnaire includes three items (α=0.59), all of which are scored on a 7-point Likert scale (1=not at all, 7=very much), for example, “To what extent do you want to punish this algorithm?” Higher scores indicate a greater desire to punish the human/algorithm in the situation morally. The free will belief questionnaire includes five items (α=0.88), all of which are scored by 7 points Likert scale (1=strongly disagree, 7=strongly agree), for example, “algorithms have free will.” Higher scores indicate that participants believe the human/algorithm has more free will. Finally, participants reported four demographics on their gender, age, ethnicity, and education level.

3. Results

Discriminatory behavior subjects (human group=1, algorithm group=2) and punishment effect learning (effective group=1, ineffective group=2, no-learning=3) are used as independent variables, and the expectation of punishment effect is used as the dependent variable for ANOVA. The data results show that the score of expectations for the punishment effect of the human group (M = 4.33, SD = 1.74, 95% CI [3.97, 4.69]) is significantly higher than that of the algorithm group (M = 4.05, SD = 1.95, 95% CI [3.63], 4.47]), the difference was marginally significant, F(1, 173) = 3.36, p = 0.068, η2p = 0.02. The main effect of penalty effect learning was significant, F(2, 173) = 60.33, p < 0.001, η2p = 0.41. The interaction between discriminatory agents and learning of punishment effects was marginally significant, F(2, 173) = 2.75, p = 0.067, η2p = 0.03. Further simple effects analysis using the Bonferroni method correction found that in the human group, there were significant differences among the three punishment effect learning groups, and the score of expectations for the punishment effect in the ineffective group (M = 3.11, SD = 1.44, 95% CI [2.62, 3.61]) were significantly lower than the effective group (M = 5.73, SD = 1.01, 95% CI [5.36, 6.10]) and no-learning group (M = 4.29, SD = 1.55, 95% CI [3.67, 4.92]]), the no-learning group was also significantly lower than the effective group, ps<0.05, and in the algorithm group, the score of expectations for the punishment effect of the ineffective group (M = 2.20, SD = 1.31, 95% CI [1.68, 2.71]) was significantly lower than the effective group (M = 5.18, SD = 1.56, 95% CI [4.59, 5.76]) and no-learning group (M = 4.59, SD = 1.60, 95% CI [3.99, 5.19]), ps< 0.001, but there was no significant difference in expectations for the punishment effect between the effective group and the no-learning group, p=1.000 (see Figure 1).

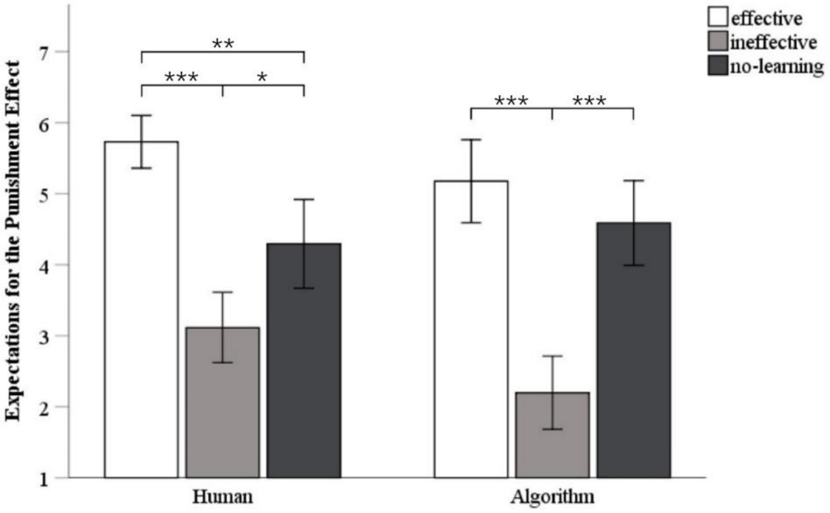

Figure 1: Expectations for the punishment effect by an algorithm versus a human.

Note. Error bars reflect standard errors. * p < 0.05, ** p<0.01, *** p<0.001.

Discriminatory behavior subjects (human group=1, algorithm group=2) and punishment effect learning (effective group=1, ineffective group=2, no-learning=3) were used as independent variables, and free will belief was used as the dependent variable for ANOVA. The data results show that the free will belief score of the human group (M = 4.90, SD = 1.13, 95% CI [4.67, 5.13]) is significantly higher than that of the algorithm group (M = 3.55, SD = 1.33, 95% CI [3.27, 3.83]]), F(1, 173) = 52.14, p < 0.001, η2p = 0.23. The main effect of penalty effect learning was insignificant, F(2, 173) = 0.04, p = 0.966, η2p < 0.001. The interaction between discriminatory actors and learning of punishment effects was insignificant, F(2, 173) = 0.63, p = 0.532, η2p = 0.01.

When free will belief, age, sex (male = 1, female = 2), ethnicity (Han = 1, minority = 2), education level (primary school and below = 1, junior high school = 2, public high school/technical secondary school/ Technical school/vocational high school = 3, junior college = 4, undergraduate = 5, graduate student = 6, doctoral student = 7) were used as covariates, the results of ANOVA show that there is no significant difference between the human group and the algorithm group in expectations for the punishment effect, F(1, 168) = 1.79, p = 0.183, η2p = 0.01, the main effect of penalty effect learning is still significant, F(2, 168) = 59.60, p < 0.001, η2p = 0.42. The interaction between discriminatory behavior subjects and punishment effect learning is still marginally significant, F(2, 168) = 2.85, p = 0.061, η2p = 0.03, and further simple effects analysis is consistent with those mentioned above.

Finally, the discriminatory behavior subject (human group=1, algorithm group=2) and punishment effect learning (effective group=1, ineffective group=2, no-learning=3) are used as independent variables, and moral punishment desire is used as the dependent variable for ANOVA. The data show no significant difference in moral punishment desire between the human and algorithm groups, F(1, 173) = 0.96, p = 0.329, η2p = 0.01. The main effect of penalty effect learning was insignificant, F(2, 173) = 2.19, p = 0.115, η2p = 0.03. The interaction between discriminatory actors and learning of punishment effects was insignificant, F(2, 173) = 1.65, p = 0.194, η2p = 0.02.

4. Discussion

In this study, people were relatively neutral on expectations for the punishment effect when they were not manipulated. The expectations for the punishment effect of the no-learning group was 4.45, and there was no significant difference between the human group and the algorithm group. Different things happen when people realize that their punishment is effective or ineffective: When people learn that their punishment is ineffective, expectations for the punishment effect are more negative, both for humans and algorithms, whereas when they know that their punishment is effective, the expectations for the punishment effect on humans will be more positive, but there will be no change on the algorithm. This shows that people were not sure whether their punishment was adequate. After four rounds of learning the effect of punishment, their attitude towards humans changed with the experimental manipulation. When the experiment told them that the punishment given to the algorithm was not effective, they believed it, but they did not believe that the experiment told them that the punishment given to the algorithm was effective, and they remained skeptical. In other words, compared with the effect of punishing humans, people have less positive expectations of the effect of punishing algorithms. To a certain extent, people have more negative expectations of the impact of punishing algorithms. This result is consistent with the hypothesis proposed in this paper.

In this study, there is no significant difference in people’s expectations of the effect of punishing humans and algorithms after controlling the belief in free will. However, in further simple effect analysis, people’s expectations of the impact of punishing algorithms are still more negative to a certain extent, suggesting that other factors influence people’s expectations about the effects of punishment from humans and algorithms. Trust is one possible explanation. Research on algorithm aversion has found that people distrust algorithms, and even when algorithms perform better than humans, people still prefer human decisions over algorithmic decisions [3]. People believe that algorithms lack the mental ability to make moral decisions and oppose algorithms replacing humans in ethical choices [9]. When people receive wrong advice, they are less likely to adopt algorithmic advice than human advice, and their trust in human direction will be more resilient, and their trust in algorithms will be weaker. One possible explanation is that participants believed that the automated algorithm would surely make the same mistake next time, while humans could notice and correct their errors [12].

5. Limitations

This study also repeated the previous research content and found that the participants did think that the algorithm had less free will, which is consistent with previous research. Still, there is no difference in the moral punishment desire of the participants between humans and algorithms. This is inconsistent with previous studies. This nonsignificant result likely reflects a type II error or the possibility that the less punishment desire to algorithm does not exist in this context. In the study of Bigman et al. [6], there were also cases where the results were not significant; after they expanded the sample size to eight hundred four participants, the result was significant. The large sample size may cause this significance, and more research is needed to explore this in the future. Of course, there is another possible explanation, perhaps because the discrimination situation in this study is relatively single, involving only gender discrimination. Participants in this study had a high desire to punish, with a mean of 6. Maybe no matter who made this sexist hiring decision, people were outraged and wanted to retaliate, which led to a non-significant result. Subsequent research can change the discrimination situation to explore.

6. Conclusion

Using an online experiment, this study finds that people have less positive expectations about the effects of punishing algorithms than the effects of punishing humans. This provides new evidence for algorithmic aversion. Future research should explore the mechanism explaining this finding and clarify the influence of factors such as free will beliefs, trust, and discrimination situations on the desire to punish algorithms.

References

[1] *Howard, A., Borenstein, J. (2018) The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. Sci Eng Ethics, 24(5):1521-36.*

[2] *Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021) A Survey on Bias and Fairness in Machine Learning. Acm Computing Surveys, 54(6).*

[3] *Xu, L., Yu, F., Peng, K. (2022) Algorithmic discrimination causes less desire for moral punishment than human discrimination. Acta Psychologica Sinica, 54(9).*

[4] *Ayre, L., Craner, J. (2018) Algorithms: avoiding the implementation of institutional biases. Public Library Quarterly, 37(3):341-7.*

[5] *Ellemers, N. (2018) Gender Stereotypes. Annu Rev Psychol, 69:275-98.*

[6] *Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A., Gray, K. (2022) Algorithmic discrimination causes less moral outrage than human discrimination. J Exp Psychol Gen.*

[7] *Hofmann, W., Brandt, M.J., Wisneski, D.C., Rockenbach, B., Skitka, L.J. (2018) Moral Punishment in Everyday Life. Pers Soc Psychol Bull, 44(12):1697-711.*

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

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

[10] *Dastin, J. (2018) Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazoncom-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-toolthat-showed-bias-against-women-idUSKCN1MK08G*

[11] *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):175-91.*

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