Cyberpsychology, Behavior, and Social Networking
Volume 00, Number 00, 2024

DOI: 10.1089/cyber.2023.0524

Psychological and Brain Responses to Artificial Intelligence's Violation of Community Ethics

Yue He, MS,[1–3] Ruolei Gu, PhD,[2,3] Guangzhi Deng, MS,[4] Yongling Lin, PhD,[5] Tian Gan, PhD,[6,7] Fang Cui, PhD,[1,8] Chao Liu, PhD,[5,9–11] and Yue-jia Luo, PhD[5,12]

Abstract

Human moral reactions to artificial intelligence (AI) agents' behavior constitute an important aspect of modern-day human–AI relationships. Although previous studies have mainly focused on autonomy ethics, this study investigates how individuals judge AI agents' violations of community ethics (including betrayals and subversions) compared with human violations. Participants' behavioral responses, event-related potentials (ERPs), and individual differences were assessed. Behavioral findings reveal that participants rated AI agents' community-violating actions less morally negative than human transgressions, possibly because AI agents are commonly perceived as having less agency than human adults. The ERP N1 component showed the same pattern with moral rating scores, indicating the modulation effect of human–AI differences on initial moral intuitions. Moreover, the level of social withdrawal correlated with a smaller N1 in the human condition but not in the AI condition. The N2 and P2 components were sensitive to the difference between the loyalty/betrayal and authority/subversion domains but not human/AI differences. Individual levels of moral sense and autistic traits also influenced behavioral data, especially on the loyalty/betrayal domain. In our opinion, these findings offer insights for predicting moral responses to AI agents and guiding ethical AI development aligned with human moral values.

Keywords: artificial intelligence, moral judgment, community ethics, event-related potential, N1

Introduction

Technological advancements have propelled the rapid autonomy growth of artificial intelligence (AI) systems, presenting an unprecedented challenge: AI's real-time decisions can contradict human commands or authority.[1,2] For instance, an autonomous driving system might redirect its user to safety, seemingly defying human control; In addition, AI security robots may disobey orders to harm unarmed individuals because of their programming.[3,4] Given the increasing human–AI social connections and human-like interpretation of AI behavior,[5,6] these actions might be morally judged as failures to fulfill one's expected responsibilities and obligations of a social role within a community (i.e., community ethics).[4,7,8]

[1]School of Psychology, Shenzhen University, Shenzhen, People's Republic of China.
[2]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, People's Republic of China.
[3]Department of Psychology, University of Chinese Academy of Sciences, Beijing, People's Republic of China.
[4]Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (BNU), Faculty of Psychology, Beijing Normal University, Beijing, People's Republic of China.
[5]State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, People's Republic of China.
[6]Department of Psychology, Zhejiang Sci-Tech University, Hangzhou, People's Republic of China.
[7]Research Institute on Aging, School of Science, Zhejiang Sci-Tech University, Hangzhou, People's Republic of China.
[8]Shenzhen Key Laboratory of Affective and Social Neuroscience, Magnetic Resonance Imaging, Center, Center for Brain Disorders and Cognitive Sciences, Shenzhen University, Shenzhen, People's Republic of China.
[9]Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, Beijing, People's Republic of China.
[10]Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing, People's Republic of China.
[11]National Demonstration Center for Experimental Psychology Education, Faculty of Psychology, Beijing Normal University, Beijing, People's Republic of China.
[12]Institute for Neuropsychological Rehabilitation, University of Health and Rehabilitation Sciences, Qingdao, People's Republic of China.

Traditional moral theories typically limit the objects of such moral judgment to humans only.[9,10] According to the classical Community-Autonomy-Divinity (CAD) triad theory, there are three distinct ethics, including autonomy (involving violations of individual rights, such as harm and fairness), community (involving violations of communal codes, such as duty and hierarchy), and divinity (involving violations of purity-sanctity), each of which evokes different moral emotions.[11,12] Empirical studies have consistently demonstrated that actions considered morally wrong when violating autonomy ethics remain so when committed by AI agents.[13–17] In contrast, human reactions to AI agents' violations of community ethics have been under-investigated, despite the increasing capability of AI agents to form relationships with humans as partners or followers.[13] Furthermore, while the CAD theory suggests that the judgment of violating one's duty and that of violating social hierarchy within a community serve similar adaptive functions,[12,18] the Moral Foundation Theory (MFT)—as an extension of the CAD theory—asserts that there are also clear differences between the two, that is, loyalty/betrayal versus authority/subversion.[19,20] The former domain focuses on the welfare of group members, whereas the latter one emphasizes respect for superiors.[19,21,22] Accordingly, the current study focuses on community ethics and examines the moral judgments of AI agents through the lens of the CAD theory; within the realm of community ethics, we distinguish between the loyalty/betrayal domain and the authority/subversion domain, in light of the MFT.

Previous studies about moral judgment toward AI agents have primarily relied on behavioral data, such as self-reports to questionnaires or open-ended questions.[13,23] However, researchers have indicated that people's moral judgments are largely determined by their implicit attitudes,[24] which may not be accessible through introspection.[25] This study used the event-related potential (ERP) technique based on time-locked electroencephalogram (EEG) signals, which can uncover people's implicit attitudes beyond self-reports.[26] Also, the exquisite temporal resolution of the ERP technique allows for the disentanglement of different mental processes underlying moral judgment that may overlap in the time domain.[27] Based on previous ERP studies on the moral judgment of humans, the early, fast, and automatic affect-laden moral intuition processes can be distinguished from the late, controlled, and deliberate moral reasoning processes using different ERP indexes.[28–30] According to these studies, this study selected the N1, P2, and N2 components as the ERP indexes of interest.

In this study, we presented participants with multiple hypothetical scenarios in which they were asked to make moral judgments about an actor (either a human or an AI agent) committing behaviors that violated community norms. The scenarios were designed to represent the domains of loyalty/betrayal, authority/subversion, or nonmorality (as a controlled condition). We collected both behavioral responses and ERPs associated with moral judgment. Our aim was to examine whether participants would judge AI and humans using the same moral standards regarding community-violating behaviors.

Moreover, we considered the potential role of individual differences, acknowledging the significant variability in moral standards and judgments among individuals.[31,32] First, we examined participants' levels of moral sense and moral sensitivity in their daily lives, as both factors could influence people's moral judgments of others.[33,34] In addition, we focused on characteristics related to autism and social withdrawal, as previous research has indicated that individuals with autism spectrum disorder or pathological social withdrawal often exhibit impaired moral judgment because of abnormalities in theory of mind or social perception.[35–37] Note that in the current study, these characteristics were investigated in a nonclinical sample because of practical restrictions. To our knowledge, the aforementioned factors have not been adequately addressed in previous studies dedicated to moral judgment concerning AI.

Materials and Methods

Participants

Before the experiment, we used G*power 3.1 software to estimate an appropriate sample size for this study.[38,39] According to the results of an *a priori* power analysis (within factors in repeated-measures F tests), 30 participants were required to reach a power value of 90% with the following parameters: an effect size of $f = 0.25$, $\alpha = 0.05$, a default measurement correlation of 0.5, and a nonsphericity correlation value (ε) of 1.[40]

This study selected college students as the research subjects because individuals with higher education are generally more knowledgeable about and interested in AI technology. Many previous studies on similar topics have also utilized college student samples.[41–47] During the recruitment stage, volunteers from Shenzhen University completed a basic demographic information questionnaire and were asked if they had previously participated in AI-related experiments. Those who confirmed previous participation were excluded to avoid the potential influence of prior experience. To account for possible dropouts or errors during the experiment, a total of 50 healthy participants were recruited. These participants were right-handed, had normal or corrected-to-normal vision, and reported no history of neurological or psychiatric disorders or head injury. Two of these participants were excluded from the formal experiment because they did not finish the task and another eight participants were excluded because their EEG data contained exceedingly large artifacts.[48] Consequently, the final sample consisted of 40 participants (20 female; age range: 18–27 years, mean ± standard deviation [*SD*] = 21.000 ± 2.324 years). This study was approved by the Ethics Committee of Shenzhen University. All participants provided written informed consent before the experiment.

Questionnaires

We used the loyalty/betrayal and authority/subversion domains of the Moral Foundations Questionnaire (MFQ) to assess moral sense,[49] Dispositional Moral Sensitivity Questionnaire to assess moral sensitivity,[50,51] the autism-spectrum quotient to assess autism-like characteristics,[52] and Chinese college students' social withdrawal questionnaire (CS-SWQ) to assess social withdrawal.[53] The reliability and validity of these scales have been supported by previous research.[53–57] See Supplementary Data for details.

Experiment stimuli and design

Vignettes for comparing moral evaluations between human and AI conditions. We selected vignettes from the loyalty/betrayal, authority/subversion, and nonmorality domains of the Moral Foundations Vignettes created by Clifford et al.[58] We then translated, adapted, and validated these vignettes and made the necessary modifications to ensure logical coherence in both human and AI conditions. Owing to space limitations, see Supplementary Data for details.

Task design. We used the section-by-section paradigm, which has been frequently used in many ERP studies investigating moral judgments.[59,60] Specifically, each vignette was presented as a combination of two sections: the context section providing background information (e.g., "An AI/Somebody ___ to a little girl.") and the keyword section describing the behavior (e.g., "talks dirty"). The context section consisted of 7–10 words, whereas the keyword section consisted of 2–3 words. To control the potential influence of AI appearance on participants' responses,[61,62] we refrained from using graphic or video material depicting an AI figure.

The formal task used a 2 (*actor*: AI or human) × 3 (*moral domain*: loyalty/betrayal, authority/subversion, or nonmorality) block design. The ERP experiment consisted of four blocks (one block with a mix of loyalty/betrayal and authority/subversion trials in which the actor was a human being, one block with a mix of loyalty/betrayal and authority/subversion trials in which the actor was an AI agent, and two nonmorality blocks, one with a human actor and the other with an AI agent, respectively), each comprising 60 trials, with a brief break between blocks. The order of blocks was pseudorandomly determined and counterbalanced across participants. The experimental stimuli were presented at a visual angle of 3.0° × 3.5°. Each block began with an instruction indicating the actors' identity (AI or human), along with a brief description emphasizing their efficiency in performing the described behavior. In AI blocks, the description also mentioned that the AI could either be an entity (e.g., robots) or not (e.g., chatbots).

Each trial (Figure 1) started with a 500 ms fixation cross to focus participants' attention. Next, the context section of a vignette was presented for 4,000 ms, followed by a random duration fixation cross (500–1,000 ms). After that, the keyword section of the vignette was displayed for 2,000 ms. Participants then provided ratings for either moral judgment or negative emotional response using a 5-point Likert scale. Moral judgment ratings reflected the actor's moral acceptability (1: totally immoral, 5: totally moral), whereas negative emotional response ratings indicated the intensity of negative emotion evoked by each vignette (1: no emotional response, 5: very strong negative emotional response). Each vignette was presented twice within the block, once for each rating. The vignette order was randomized to avoid consecutive ratings of the same vignette. In addition, the order of the moral judgment and negative emotional response ratings for each vignette was also randomized. The rating screen remained until a button was pressed. Finally, a 500 ms fixation cross appeared before the start of the subsequent trial.

Experimental procedure, EEG recording, preprocessing, analysis, and statistics

See Supplementary Data for details.

Results

Behavioral results

For brevity, only the significant findings are provided next. See Supplementary Data for nonsignificant results.

Questionnaires. See Table S3 of the Supplementary Data.

Moral rating. The main effect of *moral domain* was significant [loyalty/betrayal: 1.570, authority/subversion: 1.778, nonmorality: 3.083; $F(1, 78) = 96.730$, $p < 0.001$, $\eta_p^2 = 0.531$]. The two-way interaction of *actor* × *moral domain* was significant [$F(1, 78) = 3.320$, $p = 0.041$, $\eta_p^2 = 0.007$] (Figure 2A): the moral rating score was lower for the human actor than for the AI actor in both the loyalty/betrayal condition (human vs. AI = 1.497 vs. 1.643, $p = 0.001$) and the authority/subversion condition (human vs. AI = 1.675 vs. 1.778, $p = 0.013$) but not the nonmorality condition (human vs. AI = 3.065 vs. 3.100, $p = 0.638$).

Negative emotional rating. The main effect of *moral domain* was significant [loyalty/betrayal: 3.931, authority/

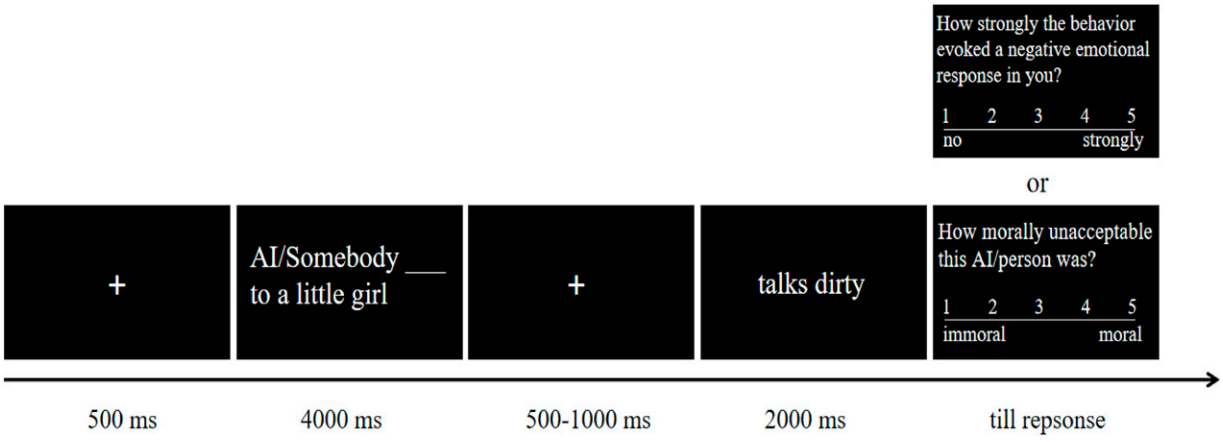

FIG. 1. Task design of an exemplar trial.

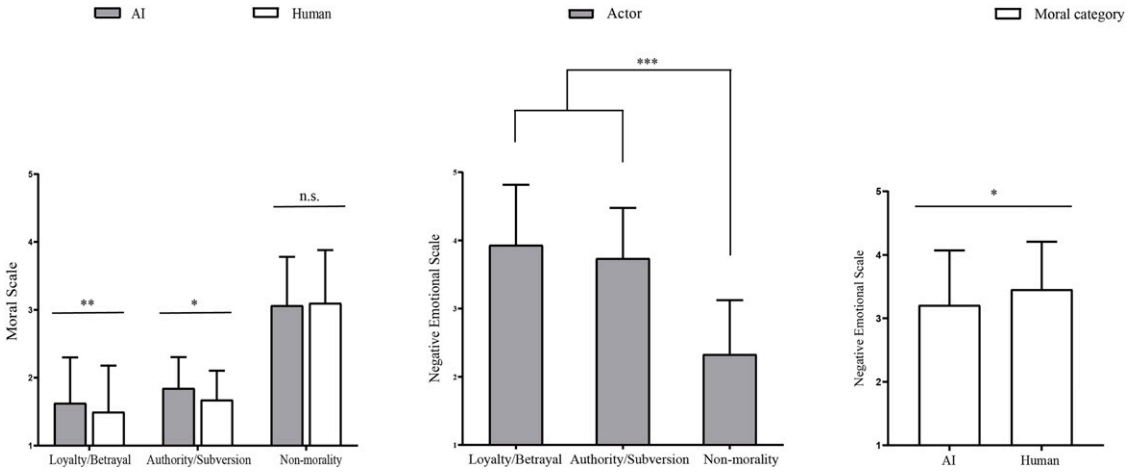

FIG. 2. **(A)** The results of moral judgment score in each condition. **(B)** The main effect of moral domain (loyalty/betrayal vs. authority/subversion vs. nonmorality) on self-reported negative emotional score. **(C)** The main effect of actor (AI vs. human) on negative emotional score. Error bars represent standard errors (n.s.: nonsignificant, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$). AI, artificial intelligence.

subversion: 3.737, nonmorality: 2.329; $F(1, 78) = 82.060$, $p < 0.001$, $\eta_p^2 = 0.443$] (Figure 2B). The main effect of *actor* was also significant [human: 3.456, AI: 3.208; $F(1, 39) = 6.450$, $p = 0.015$, $\eta_p^2 = 0.023$] (Figure 2C).

ERP results

N1. The main effect of *actor* was significant [human: –4.281 μV, AI: –3.520 μV; $F(1, 39) = 11.080$, $p = 0.002$, $\eta_p^2 = 0.018$]. The main effect of *moral domain* was significant [authority/subversion: –4.456 μV, loyalty/betrayal: –3.722 μV, non-morality: –3.524 μV; $F(1, 78) = 5.410$, $p = 0.006$, $\eta_p^2 = 0.020$]. The interaction of *actor* × *moral domain* was also significant [$F(1, 78) = 3.540$, $p = 0.034$, $\eta_p^2 = 0.015$] (Figure 3): when the participants judged betrayals, they showed a larger (i.e., more negative-going) N1 for the human actor (–4.265 μV) than for the AI actor (–3.180 μV, $p = 0.011$); the same was true when the

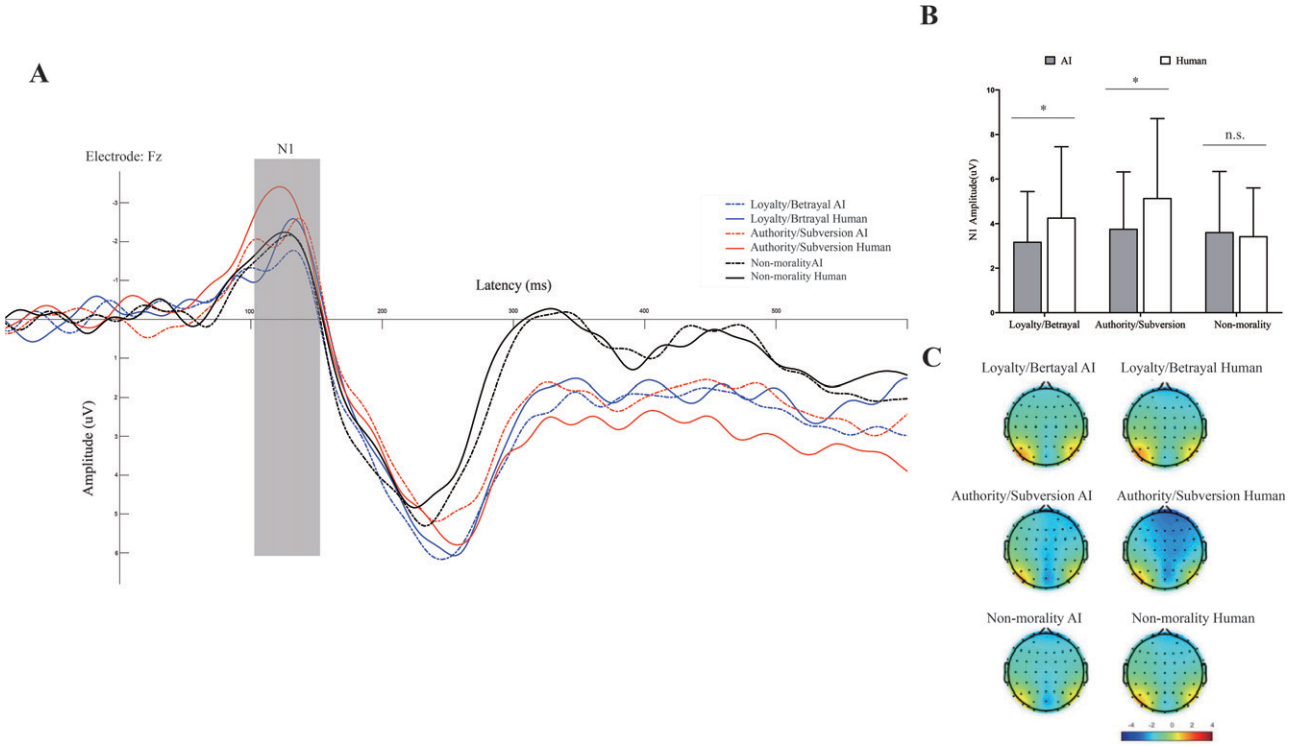

FIG. 3. **(A)** Grand average waveforms of the event-related potentials, in which the time window for N1 measurement is highlighted. These waveforms represent the data at the electrode site Fz. **(B)** The results of N1 peak amplitude in each condition. Error bars represent standard errors (n.s.: nonsignificant, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$). **(C)** The corresponding scalp topographies for all conditions.

participants judged subversions (human vs. AI = −5.145 μV vs. −3.767 μV, $p = 0.013$) but not in the nonmorality condition (human vs. AI = −3.433 μV vs. −3.614 μV, $p = 0.564$).

P2. The main effect of *moral domain* was significant [loyalty/betrayal: 5.351 μV, authority/subversion: 4.659 μV, nonmorality: 4.153 μV; $F(1, 78) = 12.559$, $p < 0.001$, $\eta_p^2 = 0.041$] (Figure 4A).

N2. The main effect of *moral domain* was significant [nonmorality: 0.280 μV, loyalty/betrayal: 2.091 μV, authority/subversion: 2.321 μV; $F(1, 78) = 18.557$, $p < 0.001$, $\eta_p^2 = 0.076$] (Figure 4B).

Correlation analyses

We examined the relationship between the questionnaires, behavioral indexes, and ERP indexes with two-tailed Pearson correlations analyses separately. These analyses were exploratory, using false discovery rate correction for multiple comparisons. The key findings were as follows: (1) MFQ score negatively correlated with moral rating for both human and AI actors in the loyalty/betrayal condition, and for human actor in the authority/subversion condition; and (2) CS-SWQ positively correlated with the N1 amplitude for the human actor in the loyalty/betrayal condition. See Supplementary Data for details.

Discussion

The present study focuses on investigating human reactions to AI agents' violations of community ethics. Overall, participants' behavioral and ERP responses in the moral judgment task differed depending on whether the moral actor being judged was a human or an AI agent, across different dimensions of community ethics (loyalty/betrayal vs. authority/subversion). Furthermore, individual levels of moral sensitivity and social withdrawal modulated some of these differences.

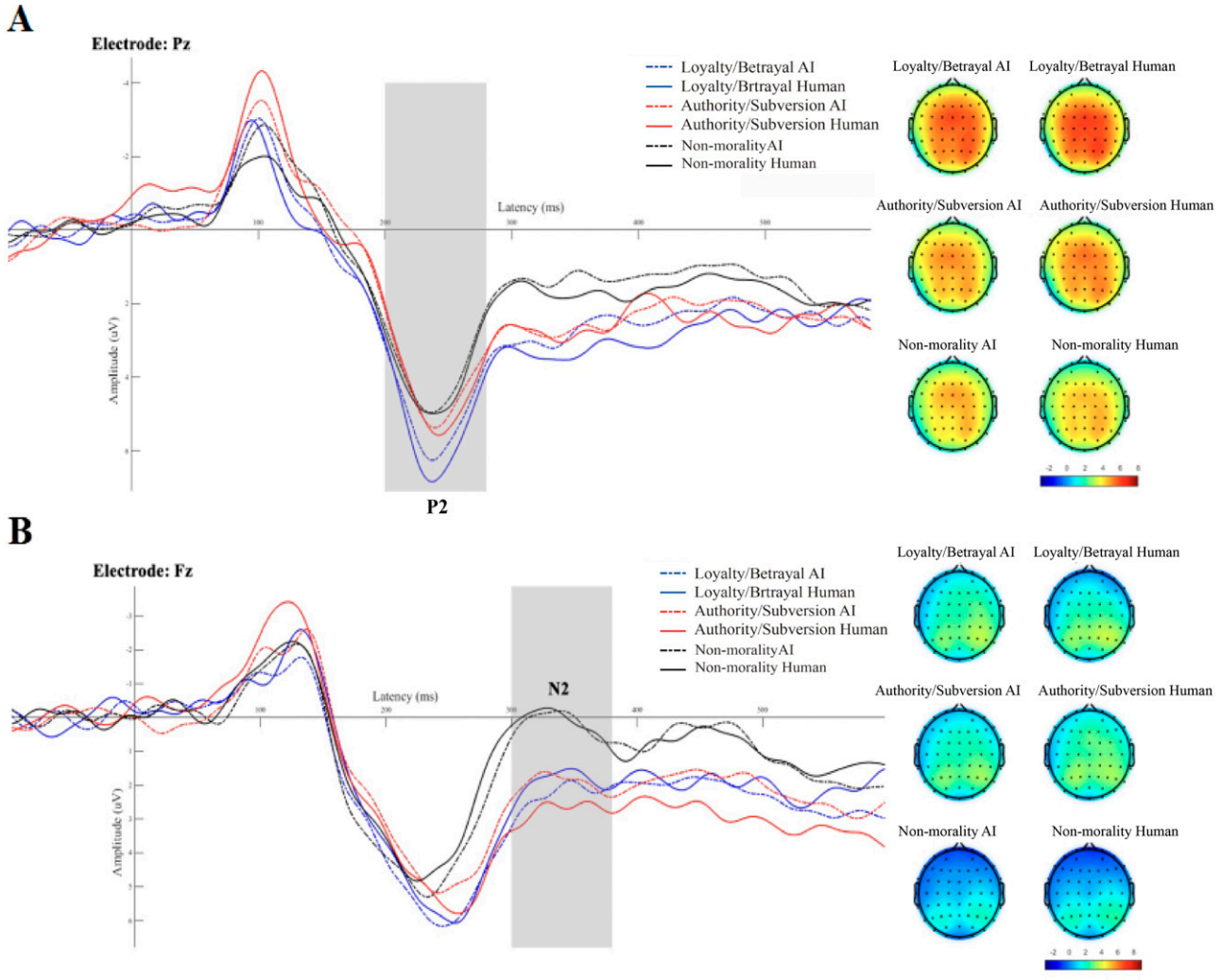

FIG. 4. **(A)** Grand average waveform of the event-related potentials, in which the time window of the P2 is highlighted. These waveforms represent the data at the electrode site Pz. The corresponding scalp topographies for all conditions are provided on the right. **(B)** Grand average waveforms of the event-related potentials, in which the time window of the N2 is highlighted. These waveforms represent the data at the electrode site Fz. The corresponding scalp topographies for all conditions are provided on the right.

Specifically, in both the loyalty/betrayal and authority/subversion domains, participants reported lower moral rating scores and stronger negative emotions when the moral actor being evaluated was a human compared with an AI agent. According to previous studies, a plausible explanation for our behavioral results is that AI is generally perceived as possessing less agency compared with human adults and consequently deemed less responsible for its moral decisions.[63–66] As a result, people tend to attribute less blame to AI agents than to humans in cases of ethical violations, including betrayals and subversions.[10] Based on the aforementioned findings, one might assume that different domains of community ethics are fundamentally similar when comparing between human and AI agents. However, the ERP results provided more nuanced details, revealing subtle differences across domains.

First, the amplitude of the N1 component was larger for humans compared with AI agents in both the authority/subversion condition and the loyalty/betrayal condition, but not in the nonmorality condition. These ERP results were consistent with the patterns observed in the moral rating scores. In the ERP literature, the N1 is regarded as a major index of the automatic and bottom-up stage of information processing.[67,68] The N1 component in moral studies might reflect initial moral intuitions that play a crucial role in moral judgment.[28–30,69] Accordingly, the current N1 results indicate that community-violating behaviors evoked weaker moral intuitions when the actor was an AI agent than a human being. In addition, subversions automatically elicited stronger moral intuitive responses than betrayals.

Second, we observed a gradual decrease in the P2 component from the loyalty/betrayal condition to the authority/subversion condition, and then to the nonmorality condition. The P2 component in moral studies has been associated with the processing of different features related to moral judgment; for instance, its amplitude distinguishes between instrumental and incidental dilemmas, as well as between shame-related and guilt-related dilemmas.[70,71] Accordingly, we suggest that although subversions initially elicited stronger automatic attention compared with betrayals (indicated by the N1), participants subsequently allocated more cognitive resources to delve into the features of betrayals compared with subversions (indicated by the P2). This was possibly because the loyalty/betrayal and authority/subversion domains are associated with different characteristic emotions: loyalty violations are expected to provoke rage, whereas authority violations tend to elicit resentment.[18,20] In our opinion, different moral emotions may interact with information processing in different ways (as reflected by ERPs), yet this variance does not necessarily affect behavioral responses.

Finally, the N2 component, which has been suggested to reflect the activity of the general system for conflict detection,[72,73] was largest in the nonmorality condition. In our opinion, this was because the vignettes in the nonmorality condition violated common sense (e.g., "On a hot summer afternoon, an Al/somebody wears a heavy coat"; see the Supplementary Data), whereas those in the loyalty/betrayal and authority/subversion conditions did not.

The importance of individual difference factors (as assessed by questionnaires) was evident in both the behavioral and ERP data. First, participants' self-reported moral sense in their daily lives (measured by the MFQ) predicted their behavioral moral rating scores in both the loyalty/betrayal domain and the authority/subversion domain when the moral actor being evaluated was a human being, which was not surprising. When the moral actor was an AI agent, only the MFQ in the loyalty/betrayal domain, but not that in the authority/subversion domain, predicted the moral rating score. In our opinion, this was because most people are unfamiliar with the situations in which AI agents serve as a lower rank in a hierarchical organization (e.g., military). As a result, they may find it more difficult to utilize their personal views of the authority/subversion norms to judge AI behavior in this domain.

Second, we found that a higher level of social withdrawal was associated with a reduced N1 response to human actor's betrayals. In contrast, the N1 response to AI betrayals showed no significant correlation with social withdrawal, which was in line with our recent finding that negative emotions affect the processing of social rewards provided by humans but not those provided by robots.[74]

Conclusions

Our findings reveal that people apply more lenient criteria when judging community-violating behavior by AI agents compared with humans, and that the difference between the loyalty/betrayal and authority/subversion domains could be observed at the brain level but not at the behavioral level. From a humanities perspective, these findings could help predict people's moral response toward AI agents who play a social role, thereby aiding in preparing for a future where AI agents actively participate in human life as social entities. From an AI science perspective, ethical AI developers may leverage these findings to design AI systems that better align with human moral values and preferences, fostering more effective and collaborative interactions between humans and AI. Nevertheless, the lack of actual experience among our participants regarding AI agents' community-violating behavior in real-life situations, particularly those related to the authority/subversion domain, may compromise the validity of our conclusion. We suggest that using virtual reality techniques to create immersive experiences in future research could address this limitation.

Acknowledgments

The authors thank Jing Yuan, Yaner Su, Ningning Zeng, Huihua Fang, and Yuqi Zhang for help with article revision. The authors used ChatGPT 3.5, Alibaba Cloud's Tongyi, and Google's Gemini to improve the writing.

Authors' Contributions

Y.H., R.G., Y.L., and F.C.designed the research; T.G. contributed to experimental materials; Y.H. conducted the experiment and collected the data; Y.H. and G.D. analyzed the data; Y.H. and R.G. wrote the article; C.L. and Y.-J.L. reviewed the article.

Ethical Approval

All procedures performed in this study were in accordance with the 1964 Declaration of Helsinki and its later

amendments or comparable ethical standards. The local ethics committee approved the experimental protocol.

Author Disclosure Statement

The authors declare no competing interests concerning the subject of this study.

Funding Information

This study was funded by the National Natural Science Foundation of China (32071083, 31920103009, 32020103008), the Major Project of National Social Science Foundation (19ZDA363, 20&ZD153), the Shenzhen-Hong Kong Institute of Brain Science—Shenzhen Fundamental Research Institutions (2022SHIBS0003), and the Scientific and Technological Innovation 2030-Major Projects (2021ZD0200500).

Supplementary Material

Supplementary Data

References

1. Birk A, Carpin S. Rescue robotics—A crucial milestone on the road to autonomous systems. Advanced Robotics 2006; 20(5):595–605.
2. Kaupp T, Makarenko A. Measuring human-robot team effectiveness to determine an appropriate autonomy level. In: 2008 IEEE International Conference on Robotics and Automation. (Mataric MJ, Schenker P, Schaal S, Sukhatme GS., eds.) IEEE: Pasadena, CA; 2008; pp. 2146–2151.
3. Louis M, Fernandez AA, Abdul Manap N, et al. Artificial intelligence: Is it a threat or an opportunity based on its legal personality and criminal liability? JISTM 2021;6(20):1–9.
4. Arnold T, Briggs G, Scheutz M. Only those who can obey can disobey: The intentional implications of artificial agent disobedience. In: International Conference on Autonomous Agents and Multiagent Systems. Lecture Notes in Computer Science. (Melo FS, Fang F., eds.) Springer: Cham; 2022; pp. 130–143.
5. Nass C, Moon Y. Machines and mindlessness: Social responses to computers. Journal of Social Issues 2000;56(1):81–103.
6. Catania F, Beccaluva E, Garzotto F. The conversational agent "emoty" perceived by people with neurodevelopmental disorders: Is it a human or a machine? Lecture Notes in Computer Science 2020;11970:5258.
7. Briggs G, Scheutz M. The case for robot disobedience. Sci Am 2017;316(1):44–47.
8. Bennett CC, Weiss B. Purposeful failures as a form of culturally-appropriate intelligent disobedience during human-robot social interaction. In: International Conference on Autonomous Agents and Multiagent Systems. Lecture Notes in Computer Science. (Melo FS, Fang F., eds.) Springer: Cham; 2022; pp. 84–90.
9. Russell PS, Piazza J, Giner-Sorolla R. CAD revisited: Effects of the word moral on the moral relevance of disgust (and other emotions). Social Psychological and Personality Science 2013;4(1):62–68.
10. Bigman YE, Gray K. People are averse to machines making moral decisions. Cognition 2018;181:21–34.
11. Shweder RA, Much NC, Mahapatra M, et al. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In: Morality and Health.1st ed (Brandt AM, Rozin P, eds.) Routledge; 1997; pp. 119–69.
12. Rozin P, Lowery L, Imada S, et al. The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). J Pers Soc Psychol 1999;76(4):574–586.
13. Wilson A, Stefanik C, Shank DB. How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations? Computers in Human Behavior Reports 2022;8:100229.
14. Malle BF, Scheutz M. When will people regard robots as morally competent social partners? 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE: Kobe, Japan; 2015; pp. 486–491.
15. Malle BF, Scheutz M. Moral competence in social robots. In: 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering. (Gardoni P, Hillerbrand R, Murphy C, Taebi B., eds) IEEE: Chicago, IL; 2014; pp. 1–6.
16. Voiklis J, Kim B, Cusimano C, Malle BF. Moral judgments of human vs. robot agents. In: 25th IEEE international symposium on robot and human interactive communication (RO-MAN) (Paiva A, Alves-Oliveira P, Tscheligi M, Giuliani M, Stollnberger G., eds.) IEEE: New York, NY; 2016; pp. 775–780.
17. Malle BF, Magar ST, Scheutz M. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In: Robotics and Well-being. Intelligent Systems, Control and Automation: Science and Engineering. (Ferreira MIA, Sequeira JS, Virk GS, Tokhi MO, Kadar EE., eds.) Springer: Cham; 2019; pp. 111–133.
18. Landmann H, Hess U. Testing moral foundation theory: Are specific moral emotions elicited by specific moral transgressions? Journal of Moral Education 2018;47(1):34–47.
19. Graham J, Haidt J, Koleva S, et al. Moral foundations theory: The pragmatic validity of moral pluralism. Advances in Experimental Social Psychology 2013;47:55–130.
20. Haidt J, Joseph C. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In: The Innate Mind. 3: Foundations and the future (Carruthers P, Laurence S, Stich S., eds.) Oxford University Press; 2008; pp. 367–391.
21. Haidt J. The new synthesis in moral psychology. Science 2007;316(5827):998–1002.
22. Graham J, Nosek BA, Haidt J, et al. Mapping the moral domain. J Pers Soc Psychol 2011;101(2):366–385.
23. Kahn PH, Kanda T, Ishiguro H, et al. "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. Dev Psychol 2012;48(2): 303–314.
24. Haidt J. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychol Rev 2001; 108(4):814–834.
25. Nisbett RE, Wilson TD. Telling more than we can know: Verbal reports on mental processes. Psychological Review 1977;84(3):231–259.
26. Lust SA, Bartholow BD. Self-reported and P3 event-related potential evaluations of condoms: Does what we say match how we feel? Psychophysiology 2009;46(2):420–424.
27. Amodio DM, Bartholow BD, Ito TA. Tracking the dynamics of the social brain: ERP approaches for social cognitive and affective neuroscience. Soc Cogn Affect Neurosci 2014;9(3):385–393.

28. Gui DY, Gan T, Liu C. Neural evidence for moral intuition and the temporal dynamics of interactions between emotional processes and moral cognition. Soc Neurosci 2016;11(4): 380–394.
29. Peng X, Lu J, Li L, et al. Three stages of perceiving consecutively moral behaviors: Neurophysiological effect of agent and valence on the moral judgments. Soc Neurosci 2020; 15(4):458–469.
30. Zhan Y, Xiao X, Tan Q, et al. Neural correlations of the influence of self-relevance on moral decision-making involving a trade-off between harm and reward. Psychophysiology 2020;57(9):e13590.
31. Graham J, Meindl P, Beall E, et al. Cultural differences in moral judgment and behavior, across and within societies. Curr Opin Psychol 2016;8:125–130.
32. Sturm RE. Decreasing unethical decisions: The role of morality-based individual differences. J Bus Ethics 2017; 142(1):37–57.
33. Johnson SG, Ahn J. Principles of moral accounting: How our intuitive moral sense balances rights and wrongs. Cognition 2021;206:104467.
34. Jagger S. Ethical sensitivity: A foundation for moral judgment. Journal of Business Ethics Education 2011;8(1): 13–30.
35. Akechi H, Kikuchi Y, Tojo Y, et al. Mind perception and moral judgment in autism. Autism Res 2018;11(9):1239–1244.
36. Fadda R, Parisi M, Ferretti L, et al. Exploring the role of theory of mind in moral judgment: The case of children with autism spectrum disorder. Front Psychol 2016;7:523.
37. Sigman M, Erdynast A. Interpersonal understanding and moral judgment in adolescents with emotional and cognitive disorders. Child Psychiatry Hum Dev 1988;19(1):36–44.
38. Faul F, Erdfelder E, Lang AG, et al. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2): 175–191.
39. Faul F, Erdfelder E, Buchner A, et al. Statistical power analyses using G*Power 3.1: TTests for correlation and regression analyses. Behav Res Methods 2009;41(4):1149–1160.
40. Vazire S. Editorial. Social Psychological and Personality Science 2016;7(1):3–7.
41. Eyssel F, Kuchenbrandt D. Social categorization of social robots: Anthropomorphism as a function of robot group membership. Br J Soc Psychol 2012;51(4):724–731.
42. Xu K, Lombard M. Persuasive computing: Feeling peer pressure from multiple computer agents. Computers in Human Behavior 2017;74:152–162.
43. Zhang Y, Wu J, Yu F, et al. Moral judgments of human vs. AI agents in moral dilemmas. Behavioral Sciences 2023; 13(2):181.
44. Fraune MR, Oisted BC, Sembrowski CE, et al. Effects of robot-human versus robot-robot behavior and entitativity on anthropomorphism and willingness to interact. Computers in Human Behavior 2020;105:106220.
45. Fuying Z, Yingying Y, Shining Z, et al. Should I blame the human or the robot? Attribution within a human–robot group. J Affect Disord 2021;255:1–9.
46. Kuchenbrandt D, Eyssel F, Bobinger S, et al. When a robot's group membership matters: Anthropomorphization of robots as a function of social categorization. Int J of Soc Robotics 2013;5(3):409–417.
47. Mirnig N, Stollnberger G, Miksch M, et al. To err is robot: How humans assess and act toward an erroneous social robot. Front Robot AI 2017;4:21.
48. Lin Y, Duan L, Xu P, et al. Electrophysiological indexes of option characteristic processing. Psychophysiology 2019; 56(10):e13403.
49. Graham J, Nosek BA, Haidt J, et al. Moral foundations questionnaire: 2008. Available from: http://www.moralfoundations.org/questionnaires
50. Zheng X, Cen G. A research on college students' dispositional moral sensitivity structure. Psychological Science (China) 2008;31(5):1026–1030.
51. Zhang W, Xiang Y. Reliability, validity and invariance of the moral sensitivity questionnaire in the China general social survey. Curr Psychol 2022;41(12):8646–8659.
52. Baron-Cohen S, Wheelwright S, Skinner R, et al. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J Autism Dev Disord 2001; 31(1):5–17.
53. Hao E, Gu C, Zhang F, et al. The effect of social withdrawal and social efficacy on internet relationship dependence among college students. China Journal of Health Psychology 2014;22(3):449–451.
54. Culiberg B, Cho H, Koklic MK, et al. The role of moral foundations, anticipated guilt and personal responsibility in predicting anti-consumption for environmental reasons. J Bus Ethics 2023;182(2):465–481.
55. Liu S, Yu C, Conner BT, et al. Autistic traits and internet gaming addiction in Chinese children: The mediating effect of emotion regulation and school connectedness. Res Dev Disabil 2017;68:122–130.
56. Zheng X, Cen G. A study on predictors of the relationship between implicit dispositional moral sensitivity and explicit dispositional moral sensitivity. Psychological Development and Education 2009;3:54–60.
57. Chen B, Guo Y, Liu W, et al. The effect of moral sensitivity on aggression among college students: The mediating role of self-control and emotion regulation strategies. Chinese Journal of Healthy Psychology 2018;26(1):93–98.
58. Clifford S, Iyengar V, Cabeza R, et al. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. Behav Res Methods 2015;47(4):1178–1198.
59. Yang Q, Li A, Xiao X, et al. Dissociation between morality and disgust: An event-related potential Study. Int J Psychophysiol 2014;94(1):84–91.
60. Wang Y, Dong Y, Yang Q, et al. Moral judgments by individuals with psychopathic traits: An ERP Study. Curr Psychol 2022;42(22):19101–19115.
61. Broadbent E. Interactions with robots: The truths we reveal about ourselves. Annu Rev Psychol 2017;68:627–652.
62. Rosenthal-Von Der Pütten AM, Krämer NC. How design characteristics of robots determine evaluation and uncanny valley related responses. Computers in Human Behavior 2014;36:422–439.
63. Gray HM, Gray K, Wegner DM. Dimensions of mind perception. Science 2007;315(5812):619.
64. Gray K, Young L, Waytz A. Mind perception is the essence of morality. Psychol Inq 2012;23(2):101–124.

65. Barlas Z. When robots tell you what to do: Sense of agency in human-and robot-guided actions. Conscious Cogn 2019; 75:102819.
66. Grynszpan O, Sahaï A, Hamidi N, et al. The sense of agency in human-human vs human-robot joint action. Conscious Cogn 2019;75:102820.
67. Luo W, Feng W, He W, et al. Three stages of facial expression processing: ERP Study with rapid serial visual presentation. Neuroimage 2010;49(2):1857–1867.
68. Näätänen R. The role of attention in auditory by event-related potentials and other brain measures of cognitive function. Behav Brain Sci 1990;13(2):201–233.
69. Yoder KJ, Decety J. Spatiotemporal neural dynamics of moral judgment: A high-density ERP study. Neuropsychologia 2014;60:39–45.
70. Sarlo M, Lotto L, Manfrinati A, et al. Temporal dynamics of cognitive-emotional interplay in moral decision-making. J Cogn Neurosci 2012;24(4):1018–1029.
71. Zhu R, Wu H, Xu Z, et al. Early distinction between shame and guilt processing in an interpersonal context. Soc Neurosci 2019;14(1):53–66.
72. Cohen MX. A neural microcircuit for cognitive conflict detection and signaling. Trends Neurosci 2014;37(9):480–490.
73. Yeung N, Cohen JD. The impact of cognitive deficits on conflict monitoring: Predictable dissociations between the error-related negativity and N2. Psychol Sci 2006;17(2):164–171.
74. Zhang D, Shen J, Li S, et al. I, robot: Depression plays different roles in human-human and human-robot interactions. Transl Psychiatry 2021;11(1):438.

Address correspondence to:
Prof. Ruolei Gu
Institute of Psychology
Chinese Academy of Sciences
Beijing 100101
China

E-mail: gurl@psych.ac.cn

Prof. Yue-jia Luo
Institute for Neuropsychological Rehabilitation
University of Health and Rehabilitation Sciences
Qingdao 266113
China

E-mail: luoyj@bnu.edu.cn