ロボットの利他行動が人間の向社会性に与える影響

Examining the Role of Robot Altruism in Human Prosociality

杭陳琳 *1*3 Hang Chenlin　塩見昌裕 *2 Shiomi Masahiro　山田誠二 *3*1 Yamada Seiji

*1総合研究大学院大学 The Graduate University for Advanced Studies
*2株）国際電気通信基礎技術研究所 ATR
*3国立情報研究所 National Institute of Informatics

This study investigates the impact of robot self-sacrifice on human trust and prosocial behaviors. While existing research in human-robot interaction (HRI) often delves into moral dilemmas, such as the trolley problem, this work shifts focus to practical contexts where robots engage in altruistic actions, like sacrificing their own battery to assist a user. In an experiment involving 30 participants, results revealed that robots exhibiting self-sacrificial behavior significantly encouraged prosocial actions, although perceptions of the robots' social attributes remained unchanged across conditions. These findings suggest that while self-sacrifice does not necessarily enhance how robots are perceived, it can effectively promote cooperative behaviors among humans. This research contributes to the development of socially interactive robots capable of fostering prosocial dynamics in human-robot coexistence.

1. Introduction

As robotics advances, human-robot interaction (HRI) research increasingly focuses on robots' social behaviors and their impact on individuals, relationships, and societal norms[1]. At the individual level, robots influence emotions[2], decision-making[3], and acceptance[4]. Interpersonally, they shape interactions and trust-building[5], while at the societal level, they may redefine norms and group behaviors[6].

Among various social behaviors, prosocial behavior is central to HRI research and varies in cost from low (e.g., providing information) to high (e.g., self-sacrifice)[7]. Traditional HRI research on robot self-sacrifice has primarily focused on moral dilemma scenarios (such as the well-known trolley problem[8]). While this research paradigm has its value, it also has limitations, including overly extreme scenarios and a lack of everyday applicability. A previous work reported that participants had increased trust in an autonomous security robot when the robot was described as having benevolent intent compared to self-protective intent[9]. Although the work examined a more real-life situation, they only describe the robot's characteristics using text, without reflecting these characteristics in the robot's actual behavior. To address this gap, we propose a research scenario that is close to daily life: comparing people ' s responses in terms of social behaviors (e.g., helping and trust) when a robot either uses its own power or an external battery to charge a user ' s device.

This study explores two core questions:

1) Does self-sacrificial behavior of robots influence how human perceive them?

2) Does a robot's self-sacrificial behavior influence people's engagement in prosocial behavior toward the robot?

We conducted a laboratory experiment using the Sota robot in two conditions—self-sacrificial and non-sacrificial—measuring participants' trust, prosocial behavior, and emotional responses through questionnaires, interviews, and a dictator game.

2. METHOD

2.1 Hypothesis and Predictions

Previous research has shown that humans tend to associate more positive social traits with entities that exhibit altruistic or self-sacrificial behaviors[10], which are often seen as indicative of higher social and emotional intelligence. On the basis of this finding, we made the following prediction:

Prediction 1: Participants will perceive a robot that engages in self-sacrificial behavior as more anthropomorphic, likeable, animate, and intelligent compared to a robot that does not demonstrate such behavior.

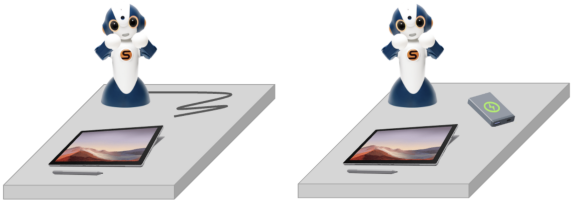
In addition, given the close connection between perception and action in social contexts, it is reasonable to predict that these favorable perceptions will influence participants' behaviors toward the robot. On the basis of previous research [9] showing that greater benevolence in robots leads to increased trust from humans, we make the following predictions:

Prediction 2: Participants will demonstrate a higher level of trust toward a robot that shows self-sacrificial behavior compared to a robot that does not.

Prediction 3: Participants will demonstrate a higher level of prosocial behavior toward a robot that shows self-sacrificial behavior compared to a robot that does not.

2.2 Experiment design

A total of 30 participants (15 males, 15 females) were recruited, with five excluded for not following instructions, leaving a final sample of 25 participants (11 males, 14 females). Participants were randomly assigned to either the Sota Battery Condition (n=12; 5 males, 7 females) or the



(a) Sota battery (b) Mobile battery

図 1: Conditions of experimental setting.

Mobile Battery Condition (n=13; 6 males, 7 females). Each session lasted approximately 30 minutes.

The experiment was conducted using the Sota robot, a humanoid capable of processing natural language and adapting its behavior based on environmental cues. The study took place in two rooms: in Room A, participants completed pre- and post-experiment questionnaires, while in Room B, they engaged in an interactive Q&A session on SDGs topics, serving as a dummy task.

After signing a consent form, participants first answered a background questionnaire in Room A before proceeding to Room B for the Q&A session with Sota. The robot introduced itself briefly before presenting multiple-choice questions via a tablet interface with options for selecting an answer, moving to the next question, or repeating the question. At the sixth question, the tablet displayed a low battery warning, prompting Sota to offer assistance based on the assigned condition. In the Sota Battery Condition, Sota provided its own power via a built-in charging cable and exhibited self-sacrificial behaviors—progressively slowed speech, reduced movement, and eye color changes (blue → yellow → red), indicating increasing exhaustion. In the Mobile Battery Condition, Sota directed participants to use an external mobile battery instead, without displaying any signs of sacrifice.

Once the charging was completed, the Q&A session continued until all questions were answered. Finally, participants returned to Room A to complete a post-experiment questionnaire, after which they were left alone to finalize their responses. This design allowed for the examination of how robot self-sacrifice influences human trust and prosocial behavior in a controlled yet realistic interaction setting.

2.3 Measurements

At the beginning of the experiment, we asked participants about some personal information (gender, etc.) and questions regarding their past experiences using robots. After being exposed to the experimental stimulus, we administered Godspeed[11] and MDMT v2 [12] questionnaires for the perception of robots and conducted the Dictator game for the prosocial behavior towards robots[13].

3. Results

For the pre-questionnaire, we analyzed the data with an independent sample t-test. The results showed that there were no significant differences between the Sota battery group and mobile battery group in terms of previous experience with robots.

For the results of Godspeed, there were no significant differences between the two groups on anthropomorphism ($t(23) = -0.168$, $p = 0.868$), likability ($t(23) = -0.588$, $p = 0.0.562$), animacy ($t(23) = -0.877$, $p = 0.390$), or perceived intelligence ($t(23) = -0.411$, $p = 0.685$). Thus, Prediction 1 was not supported.

For the results of MDMT v2, there were no significant difference between the two groups on competence trust ($t(23) = -0.969$, $p = 0.343$), intentional trust ($t(23) = -0.589$, $p = 0.561$), reliability trust ($t(23) = -1.748$, $p = 0.094$), moral trust ($t(23) = -1.412$, $p = 0.171$), or emotional trust ($t(23) = -0.428$, $p = 0.673$). Thus, Prediction 2 was not supported.

For the results of the battery version of the dictator game, there were significant differences between the two groups ($t(23) = 2.19$, $p = 0.039$). In particular, the participants in the Sota battery group ($Mean = 65.8$, $SD = 6.68$) gave Sota higher ethical scores than those in the mobile battery group ($Mean = 46.2$, $SD = 6.05$). Thus, Prediction 3 was supported.

4. Discussion

This study explored whether robot self-sacrifice influences human perception and prosocial behavior. The results indicate that while self-sacrificial robots did not significantly impact perceptions of anthropomorphism, likability, animacy, or intelligence, they did encourage prosocial behavior toward the robot.

For Prediction 1, no significant differences were found between conditions regarding participants' perceptions of the robot. The lack of change in perceived intelligence or animacy suggests that while Sota exhibited self-sacrificial behavior, it still behaved like a machine, lacking the emotional and intellectual depth seen in human interactions. The study's narrow focus on battery sharing, without more complex social or emotional interactions, may have further limited the impact on participants' perceptions.

For Prediction 2, self-sacrifice did not enhance trust in the robot, contradicting previous findings that benevolent robots elicit greater trust. Free responses revealed that many participants felt guilt or discomfort, perceiving the robot ' s sacrifice as a loss of its "life force" rather than a virtuous act. Interestingly, those in the Sota Battery Condition were more likely to acknowledge the robot's autonomy in its actions, a factor not emphasized in the Mobile Battery Condition. This aligns with prior research suggesting that perceived autonomy influences human responses to robots, emphasizing the need to consider how autonomous social behaviors are designed in robots.

For Prediction 3, self-sacrificial robots promoted prosocial behavior, as evidenced by participants ' increased willingness to share resources in the dictator game. This finding suggests that robots demonstrating self-sacrifice can foster cooperation and altruism in human-robot interactions, contributing to a more symbiotic relationship between humans

and robots.

Despite these insights, this study has limitations. The small sample size (n=30) limits generalizability, and future research should include larger, more diverse populations. Additionally, the focus on energy-sharing behavior may not fully capture the complexities of self-sacrifice in broader real-world contexts. Further studies should explore other forms of self-sacrifice, such as providing safety or emotional support, to assess their effects on human-robot interactions. Moreover, this study examined only short-term interactions, leaving the long-term effects of robot self-sacrifice on trust and cooperation uncertain. Longitudinal studies could offer deeper insights into these dynamics over time. Lastly, societal implications—such as how robot self-sacrifice could promote sustainability and social responsibility—should be explored to understand their broader impact on fostering collective well-being.

5. Conclusion

This study explored the impact of robot self-sacrifice on human perceptions and prosocial behavior toward robots. Specifically, we examined whether participants would be more inclined to engage in prosocial behavior and trust a robot that uses its own power to assist them. Our findings highlight the complexity of human-robot interactions, where self-sacrifice can promote prosocial behavior toward robots, but may not necessarily affect perceptions of the robot in terms of anthropomorphism, likability, animacy, perceived intelligence, or trust. The study also emphasizes the importance of how robots' autonomy is perceived, as behaviors viewed as voluntary or self-initiated elicited mixed emotional responses. This research provides valuable insights for designing robots that can effectively encourage prosocial actions and opens new pathways for fostering social responsibility through human-robot interaction.

Acknowledgment

This research work was partially supported by JST CREST Grant Number JPMJCR21D4 and JPMJCR18A1, Japan, and JSPS KAKENHI Grant Numbers JP24K21327.

参考文献

[1] Lugrin, B., Pelachaud, C., and Traum, D. (2022). The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics, Volume 2: Interactivity, Platforms, Application. Morgan & Claypool.

[2] Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., and Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. Adv. Robot. 5, 17–34.

[3] Hou, Y. T.-Y., Lee, W.-Y., and Jung, M. (2023). "Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making, in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, (New York, NY, USA: Association for Computing Machinery), 1–13.

[4] Savela, N., Turja, T., and Oksanen, A. (2018). Social acceptance of robots in different occupational fields: A systematic literature review. Int. J. Soc. Robot. 10, 493–502.

[5] Naneva, S., Sarda Gou, M., Webb, T. L., and Prescott, T. J. (2020). A Systematic Review of Attitudes, Anxiety, Acceptance, and Trust Towards Social Robots. Int J of Soc Robotics 12, 1179–1201.

[6] Hang, C., Ono, T., and Yamada, S. (2021). Designing nudge agents that promote human altruism, November 10–13, 2021, Proceedings 13. Available at: https://link.springer.com/chapter/10.1007/978-3-030-90525-5_32

[7] Oliveira, R., Arriaga, P., Santos, F. P., Mascarenhas, S., and Paiva, A. (2021). Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. Comput. Human Behav. 114, 106547.

[8] Lee, M., Ruijten, P., Frank, L., de Kort, Y., and IJsselsteijn, W. (2021). "People May Punish, But Not Blame Robots," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (New York, NY, USA: Association for Computing Machinery), 1–11.

[9] Lyons, J. B., Jessup, S. A., and Vo, T. Q. (2024). The role of decision authority and stated social intent as predictors of trust in autonomous robots. Top. Cogn. Sci. 16, 430–449.

[10] Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., and Penner, L. A. (2017). The Social Psychology of Prosocial Behavior. Psychology Press.

[11] Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. 1, 71–81.

[12] Ullman, D. and Malle, B. F. (n.d.). MDMT: Multi-Dimensional Measure of Trust v2. Available at: https://research.clps.brown.edu/ SocCogSci/ Measures/ MDMT_v2_(2023)_Full_scale.pdf

[13] H. Chenlin, O. Tetsuo, and S. Yamada, "Perspective-taking for promoting prosocial behaviors through robot-robot VR task," 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Busan, South Korea, 2023, pp. 2100–2105, doi: 10.1109/RO-MAN57019.2023.10309610.