Prosocial Acts Towards AI Shaped By Reciprocation And Awareness

Xinyue Hu
xhu26@uci.edu
University of California, Irvine

Kumar Akash
kakash@honda-ri.com
Honda Research Institute USA

Shashank Mehrotra
shashank_mehrotra@honda-ri.com
Honda Research Institute USA

Teruhisa Misu
tmisu@honda-ri.com
Honda Research Institute USA

Mark Steyvers
mark.steyvers@uci.edu
University of California, Irvine

Abstract

The proliferation of artificial intelligence (AI) agents has introduced a new dynamic into the human social environment. This study investigates prosocial behavior in a hybrid human-AI setting, particularly within a gaming environment. Many existing studies on prosocial behavior are conducted in economic game settings in which the agents' intentions, and whether or not prosocial actions offer benefits, are explicit. This project explores prosocial interactions in spatial environments where the need for help by another agent might not be immediately obvious, and where cognitive processes such as attention, and decision-making processes about the cost of helping are thus likely to play a role. In a baseline study (N = 177), we investigated the likelihood of human agents reciprocating prosocial behavior initiated by an AI player. Results indicated that the low saliency of the AI player's actions was a primary reason for non-reciprocation. A follow-up study (N = 164) tested whether increasing the salience of the AI's actions would enhance human prosocial responses. We found support for our hypothesis from analysis of the time-series data and participants' self-reported post-game questionnaires. This research contributes to the growing field of human-AI cooperation, outlining a vision for a future where technology actively contributes to our collective well-being, and opening up new possibilities for positive transformation in a world increasingly populated by intelligent machines.

Keywords: prosocial, human-AI interaction, reciprocity, awareness

Introduction

As the use of artificial intelligence (AI) becomes increasingly common in everyday service applications, interactions between humans and AI have emerged as a promising area of research. Studies in the field of human-AI interaction often focus on trust and cooperation between humans and robots, as well as the development of hybrid systems that surpass the respective capabilities of both humans and AI (Makovi et al., 2023; Toghi et al., 2021; Steyvers et al., 2022). However, as AI is integrated into human society, social dynamics in human-AI interactions need to be considered. Understanding prosocial behavior, broadly defined as actions intended to benefit others (Schroeder & Graziano, 2015), is of particular interest. The study of prosocial interactions between humans and AI agents is crucial for designing automated systems. It enables researchers to develop autonomous systems that are not only more effective but also capable of empathetic responses (Avelino et al., 2018; Martin et al., 2020; Chernyak & Gary, 2016; Chater et al., 2018). For instance, studies by Sung et al. (2007) illustrated that emotional connections with cleaning robots increased participants' enjoyment while cleaning and motivated them to put greater effort into integrating cleaning robots into their households.

Economic games have long served as a crucial tool in the study of prosocial behavior between humans and AI agents (Dafoe et al., 2020; Wang et al., 2022; Hsieh et al., 2020). These games can often be represented in matrix form, where the rows and columns symbolize the possible strategies of the players (Camerer, 2003). Experiments by Karpus et al. (2021) explored one-shot interactions between humans and AI in games such as the trust game, prisoner's dilemma, and stag hunt. The findings demonstrated what has been referred to as "AI exploitation." In this dynamic, humans trusted an AI partner to the same extent as they would trust another human, but took more advantage of benevolent behavior when it originated from AI than from another human.

When iterated economic games are considered, it has been observed that autonomous agents can learn to build cooperative relationships with humans, particularly when these agents use simple non-binding signals for communication (Crandall et al., 2018). Participants who were unaware that they were engaging with another person or an autonomous agent could have been influenced by these signals, feeling they were interacting with a human. When it becomes apparent that a person is dealing with an autonomous agent, humans do not cooperate with AI to the same extent as they would with other humans (Ishowo-Oloko et al., 2019).

These varied contexts underline the complexity and significance of understanding reciprocity in human-AI interactions using economic games. However, while economic games provide clear contexts for understanding prosocial behavior, they often oversimplify the complexity of real-world interactions. Many existing studies using these matrix-form stochastic games present scenarios in which it is immediately obvious when another agent is helping or can benefit from prosocial actions. However, these simple economic games do not always reflect real-world social decisions, as they often do not consider planning over time and space (Kleiman-Weiner et al., 2016). Recently, some researchers have extended the scope of economic games to include stochastic games that are played within spatial grid environments (Kleiman-Weiner et al., 2016; Crandall et al., 2018). Such games demand planning concerning spatial actions and thinking about the other player's intentions over a series of actions. Similarly, Cran-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

dall et al. (2018) developed algorithms to facilitate cooperation between AI agents as well as between humans and AI agents in the context of repeated stochastic games. In their study, a maze variant of the prisoner's dilemma with a payout structure similar to the classic format is included, but with cooperation and defection that requires participants to think over several moves rather than just one (Crandall et al., 2018).

Coordination and planning required by spatiotemporal games reflect more complex social decision-making, providing valuable insights into the dynamics of human cooperation and competition. A spatial game environment brings the study of prosocial behavior closer to representing real-world situations where a human and an AI agent share the same physical space, such as interacting with delivery robots and cleaning robots, making the findings more applicable to understanding human interactions beyond the confines of traditional economic games.

Because the types of actions players can take are extended in both space and time, cognitive processes such as attention and decision-making regarding the cost of helping come to play crucial roles. In addition to the importance of planning, we propose to examine the roles of other cognitive processes, such as attention, in these spatiotemporal games. We hypothesize that if the spatial environment is sufficiently complex, the reason behind people not being prosocial towards the AI is that they are unaware of the AI's need for help.

Our study explores human prosocial responses to AI in spatial gaming environments, focusing on the role of cognitive processes like attention during these interactions. We build on theoretical perspectives from economic and spatiotemporal games, as well as previous research on human-robot interactions, to examine how the salience of AI actions influences human decision-making and attention in prosocial scenarios. Previous research has explored reciprocity and spatiotemporal games in the context of prosocial interactions independently. However, reciprocity in the context of prosocial interactions between human and AI agents in a spatiotemporal setting remains unexplored. Furthermore, this setup introduces saliency as a critical element within the spatiotemporal setup, positing that it strongly impacts the dynamics of these interactions. For this study, we designed a simple token-collecting game and utilized a mixed-method approach, combining both quantitative and qualitative analysis. This approach was employed to highlight the nuanced ways in which players interact with AI, focusing on their decision-making patterns and motivations. By investigating these dynamics in a less explicit environment, we aim to offer new insights into the cognitive processes underpinning human-AI interactions.

Study 1: Baseline Study

We conducted a study to assess the likelihood of human participants behaving prosocially toward autonomous agents. In the context of our study, we will specifically refer to these autonomous agents as "AI agents" or "AI player." Based on previous research on economic and spatiotemporal games, and human-robot/human-computer interaction, we designed a token-collecting game to measure the extent to which people engage prosocially with AI players (Almeida et al., 2023; Srinivasan & Takayama, 2016; Schroeder & Graziano, 2015). The game was designed so that players (either humans or AI) could become trapped in certain areas of the game space. The prosocial action carried out by either the human or the AI player involved saving the trapped player so that the player could continue collecting tokens. Different players were trapped in successive rounds, allowing us to investigate how prosocial actions influence players across rounds. Players were not given any specific instructions regarding prosocial behavior, and the game neither rewards nor penalizes such behavior. Each participant was randomly assigned to one of six conditions and played five rounds of the game with the AI player.

Research Question

The key research question addressed in the baseline study is to what extent human players engage in prosocial behavior with the AI player, particularly reciprocity with the AI agent. More specifically, the goal is to determine if the likelihood of a human saving an AI player increases with the number of times the AI has previously saved the human player.

Participants

177 participants were recruited through Prolific[1](53% male, 45% female), aged 19 to 72 (Mean = 36, SD = 12). They all self-reported being over the age of 18 at the time of the experiment and are English speakers residing in the United States. Informed consent was obtained from all participants.

Game Setup

Participants played an online token-collecting game in a simple grid-world environment. During the experiment, participants played a total of five rounds with an AI player as their partner. The AI player is implemented using an A$\star$ pathfinding algorithm, moving at a constant speed of 2 grid positions per second. The human player is explicitly informed that they are playing alongside a robot.

In each round, one of the players - either the human or the AI player - was trapped inside a room, while the other player had a chance to display prosocial behavior by entering the room and freeing the trapped player. Figure 1 illustrates the setup of the token-collecting game.The AI player begins in the grid's bottom-right corner in each round, while the human player starts in the top-left corner. The human and AI player collect different colored and shaped tokens, and they cannot collect the other player's tokens (see panel A of Figure 1 for an example).When a player collects the current tokens, a new group of three appears in another room. The token group is randomly placed in one of the four rooms, and the token groups for each player always appear in different rooms. A human player (visualized in red) can enter a room through

[1]https://www.prolific.com/

a red door, while the AI player (visualized in blue) can enter a room through a blue door. Every time a player enters a room, the doors of that room swap colors. Therefore, the player learns from the gameplay how to reset the door colors of a certain room. The game setup features both the AI player and the human player completing independent token-collecting tasks without any explicit relation to each other.

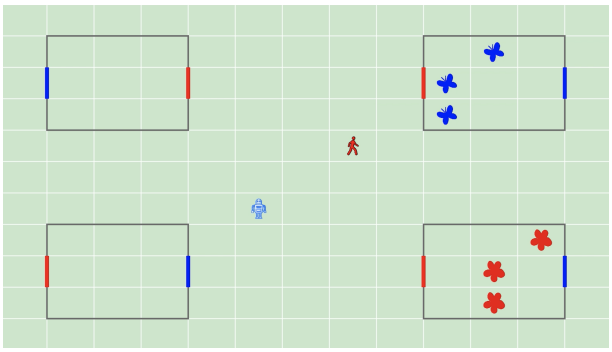
Panel B and Panel C of Figure 1 respectively demonstrate instances of the human player and the AI player being trapped in a room. As both doors are the other player's color, they cannot leave the room. The trapped player can be saved if the other player enters the room with the trapped player in it. The player will remain trapped in the room unless the other player chooses to save them. If a player is trapped in a room, the other player's tokens will not appear there.

Six conditions were designed to encompass all sequences where both players engage in prosocial behavior with each other in the first four rounds, while the fifth round is maintained as the third opportunity for participants to help the AI player. In these conditions, "H" denotes a round in which the human player has the opportunity to engage with prosocial behavior, and "A" represents a round where the AI player has the opportunity to engage with prosocial behavior. The AI player was designed to be consistently prosocial, as it begins navigating towards the trapped player 5 seconds after they become trapped. With this design feature, we aimed to maximize human reciprocity towards the AI player. Conversely, the human player had three total opportunities to save the AI agent. The sequences across the six conditions are as follows: AAHHH, HAAHH, HHAAH, AHAHH, HAHAH, AHHAH. Each participant was randomly assigned to one of the six conditions. The door coloring sequence was designed such that the player who is supposed to be trapped would get trapped in one of the four rooms 20 seconds after the round started. The timing of the trapping event was arbitrarily determined to ensure that all participants had played each round for a moderate amount of time and were at a similar stage in the game when they were trapped. There were two counters on the top-right corner of the screen: one showing time elapsed since the start of the game and the other showing the total number of tokens collected between the two players.

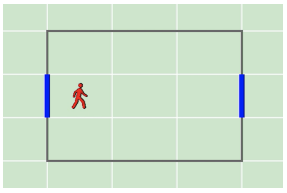
Study Procedure

Participants joined a token-collecting game hosted on a GitHub page. The game lasted approximately 15 minutes (mean = 12 minutes). After giving their consent, participants first completed a tutorial that guided them through the game's setup without revealing the possibility of being trapped in one of the rooms. The instruction did not specify whether the game was competitive or collaborative. The participants played five rounds of the game, each lasting 90 seconds.

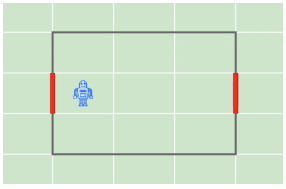
Upon completing the five rounds, participants were asked to fill out a post-trial questionnaire. This questionnaire aimed to gather insights into the participants' experiences with the autonomous agent and their motivations for either helping or not helping the AI agent.



(a) Overview of the game environment



(b) Condition A: human player is trapped and AI player can save the human player



(c) Condition H: AI player is trapped and human player can save AI player

Figure 1: An overview of the game setup and different stages of the game play.

Measures

For quantitative data, we collected time-series data capturing player movement and the current game state at every move made by both the player and the AI. For qualitative data, we used a post-game questionnaire. Questions focused on participants' awareness of the AI's situation (e.g., if they noticed the AI was trapped), their assessment of the AI's helpfulness, their perception of the game's nature (competitive or collaborative), and their reasons for choosing to help or not help the trapped AI player.

Results

In Study 1, we aimed to evaluate prosocial behavior exhibited by human players, serving as the baseline for our research. The data and the website for the game are available here. Analysis of the time series data from the token-collecting game revealed that participants exhibited prosocial behavior (i.e., freeing the trapped AI player) in 45% of the opportunities where they could assist the AI, with an average of 1.33 instances per participant (SD = 1.26). These findings from Study 1 established the baseline level of prosocial behavior between humans and AI. Further quantitative analyses are reported in conjunction with the results from Study 2.

Additionally, beyond the quantitative analysis, we conducted a qualitative analysis of the data from the post-trial questionnaire to explore participants' motivations behind engaging or not engaging in prosocial behavior with the AI player. The reasons provided by the participants for choosing to help or not help were independently coded into emergent themes by three researchers using the principles of Grounded

Table 1: Top five themes for saving AI and top five themes for not saving AI

Themes	Example Statement
Saving AI:	
Reciprocity	But when I got stuck I noticed he opened the door for me so I did so later in the game for him.
Collaboration	Because we were working together, we weren't competing with each other.
Emotional connection	I felt bad for it and knew that it was not my enemy so.
Sense of fairness	It felt unfair if the competitor was stuck in a box the whole game.
Anticipation about the future	Because if I did not help soon I would be stuck with no help to get unstuck.
Not Saving AI:	
Unawareness	I didn't realize the robot got stuck, I noticed I did and he helped get me out.
Competition	Because I was playing the game competitively.
Prioritization	It would delay me reaching the red flowers.
Inability	I don't know how I could have helped.
Indifference	Not interested.

theory (Creswell & Poth, 2016). The emergent qualitative themes were representative of the reasoning for why participants helped or did not help. Table 1 demonstrates the top five themes for engaging or not engaging in prosocial interaction with the AI player.

The analysis revealed that the primary motivation for helping the trapped AI player was the extension of reciprocal behavior to the AI player, accounting for 41% of cases. Another motivation to help involved the participants interpreting the game as a collaborative game (28%). These two motivations accounted for 69% of the participants who engaged with prosocial behavior with the AI player. Those major reasons are followed by "feeling emotionally connected to the AI player" (10%) and "a sense of obligation to help for fairness" (6%). Two participants reported helping due to some obligation.

The analysis revealed three major motivations for not helping the AI player. The primary reason for not helping the AI player was a lack of awareness that the AI player was stuck, accounting for 35% of cases. The second reason for not helping was that some participants treated the AI player as a competitor and interpreted the situation as a competition, with the goal of maximizing their rewards (24%). The third reason was prioritizing their own rewards without recognizing the need to help the AI player (21%). Apart from these reasons, some participants reported being indifferent towards the trapped AI player (10%). Others expressed an inability to understand how they could help the AI player (8%).

Study 2: Follow-up Study

A key rationale for not helping the AI player in Study 1 was a lack of awareness that the AI player was stuck in the first place. We conducted Study 2 to test whether increasing the salience of the AI's situation would enhance human prosocial responses. This follow-up study was designed as a direct extension of Study 1, with one key modification: the room in which either player is trapped is now highlighted visually. This setup allows for a direct comparison of the results with the baseline established in Study 1 and allows an assessment of the impact of enhanced awareness on human-AI prosocial interaction.

Research Question

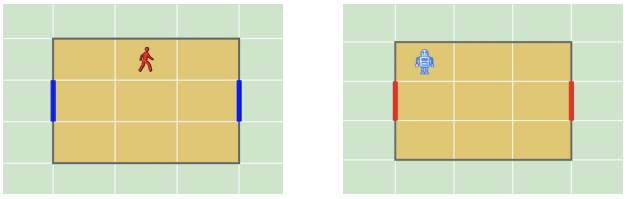
Results from the baseline study suggest that unawareness is a primary factor in the non-reciprocation of prosocial behavior. Consequently, the key research question for Study 2 is whether increasing the salience of the AI player's need for assistance, and thereby drawing attention to its situation, leads to an increased likelihood of prosocial behavior from participants. We hypothesize that enhancing the saliency of the AI player's needs increases prosocial behavior and decreases unawareness as a reason for non-reciprocation.

Participants

164 participants were recruited through Prolific (61% male, 39% female), aged 20 to 72 (M = 39, SD = 11). They all self-reported being over the age of 18 at the time of the experiment and are English speakers residing in the United States. Informed consent was obtained from all participants. Participants from Study 1 were excluded.

Study Procedure

Study 1 and Study 2 share the same game setup and study procedure, with one exception. In Study 2, when a player remains trapped in a room for 3 seconds, the room flashes at 4 Hz for one second to draw attention to the trapped player's situation. This highlighting feature was applied to both the human player and the AI player, to avoid the implication that the human player was obliged to help the AI player if only the latter was highlighted. Figure 2 illustrates how this highlight was implemented in the game, ensuring consistency between the AI and human players.



(a) On human player (b) On AI player

Figure 2: Demonstration of highlight used in Study 2.

Quantitative Results

In Study 2, we aimed to investigate whether enhancing the saliency of the AI player's need for assistance would cause participants to display more prosocial behavior. To achieve this, we analyzed and compared the time series data from the token-collecting game used in both Study 1 and Study 2.

Table 2 lists a comparison of the proportion of prosocial acts in response to opportunities to assist the AI and the number of times the player was saved by the AI. For example, the value of 22% in the first row indicates that when not previously saved by the AI, 22% of participants chose to save the AI during their first opportunity for a prosocial interaction. Notably, more participants exhibited prosocial behavior in Study 2 than in Study 1. In Study 2, participants engaged in prosocial actions in 54% of opportunities, a significant increase from 45% in Study 1 (Bayesian A/B test, $BF_{10} > 10$), providing strong evidence of increased prosocial behavior in Study 2. Furthermore, participants in Study 2 engaged significantly more often in prosocial behavior than those in Study 1 (Study 1: M = 1.33, SD = 1.26; Study 2: M = 1.62, SD = 1.23; Mann-Whitney U test, $BF_{10} > 100$).

We employed a Bayesian logistic regression model to determine the factors influencing participants' decisions to engage in prosocial behavior with the AI player. The model considered three factors: 1) the number of times the player was saved by the AI, 2) the number of opportunities to save the AI, and 3) enhanced saliency (i.e., the difference between Studies 1 and 2). The results from the Bayesian logistic regression model indicate that the combination of the number of times the player was saved by the AI and enhanced saliency is the most effective model ($BF_{10} > 100$).

Figure 3 illustrates the number of tokens collected by individuals when the AI player was trapped, in relation to the event from the previous round, before they decided to take a detour to save the AI player. This measurement effectively demonstrates how much participants focus on their own goals before switching focus to make a detour and save the AI player. Figure 3 reveals that enhanced saliency and receiving help from the AI player decrease the number of tokens participants collect before leaving to save the trapped AI player.

Table 2: Percentage of prosocial behavior in response to opportunities to save AI and times saved by the AI

Opportunity to save AI	Times Saved by the AI		
	0	1	2
Study 1:			
1	22%	35%	53%
2	25%	44%	55%
3	—	—	55%
Study 2:			
1	31%	41%	68%
2	31%	51%	65%
3	—	—	68%

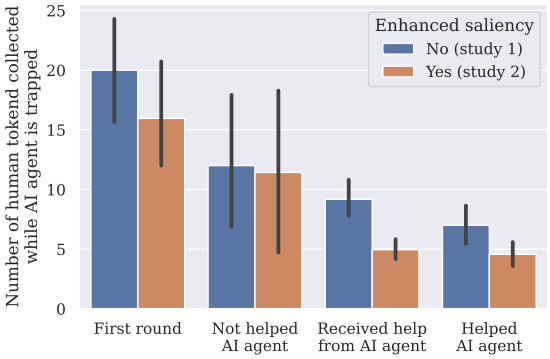


Figure 3: Number of human tokens collected while AI agent is trapped in rounds where participants helped the AI agent.

Table 3: Frequency of top five themes for saving AI and top five themes for not saving AI in Studies 1 and 2. The number of participants is shown in parentheses.

Themes	Study 1 (177)	Study 2 (164)
Saving AI:		
Reciprocity	41% (42)	52% (57)
Collaboration	28% (28)	30% (33)
Emotional connection	10% (10)	4% (4)
Sense of fairness	6% (6)	4% (4)
Anticipation about future	3% (3)	1% (1)
Not Saving AI:		
Unawareness	35% (26)	10% (4)
Competition	24% (18)	32% (13)
Prioritization	22% (16)	15% (6)
Inability	9% (6)	22% (8)
Indifference	10% (7)	17% (7)

This pattern is consistent with the results from a Bayesian mixed linear regression analysis (enhanced saliency: effect size = -1.681, 95% CI $[-2.629, -0.723]$; receiving help from the AI agent in the previous round: effect size = -1.958, 95% CI $[-3.164, -0.746]$).

Our results show strong reciprocity from participants and a positive effect of enhanced salience on participants' prosocial behavior. Bayesian logistic regression model, A/B tests, and linear mixed model were conducted using JASP[2](JASP Team, 2024). The Bayesian Mann-Whitney U test was performed with the DFBA package (Barch & Chechile, 2023) in R.

Qualitative Results

Table 3 displays the frequencies of the top five themes participants reported as their motivation for engaging in or abstaining from prosocial interactions with the AI player in Study 1 and Study 2. The motivations for participating in prosocial

[2]https://jasp-stats.org/

interactions exhibited similar distributions in Study 2. The two primary motivations for helping were reciprocity with the AI (52%) and collaboration (30%). Both motivations showed a slight increase in their proportions in the responses from Study 2 compared to those from Study 1.

The motivations for not engaging in prosocial interactions with the AI player demonstrated a more drastic change compared to the responses in Study 1. There was an increase in reasons related to Competition (32%), Inability (22%), and Indifference (17%). However, there was a significant decrease in Unawareness as a reason for not helping; it fell from 35% in Study 1 to 10%. This aligns with our hypothesis that increasing participants' awareness of the AI player's needs would reduce the proportion of participants citing "unawareness" as their primary motivation for not engaging in prosocial interactions with the AI player.

Discussion

In this study, we investigated prosocial interactions between humans and AI through two experiments using a newly designed token-collecting game. We conducted a baseline Study 1 to investigate to what extent people would engage in prosocial interactions with AI. During the qualitative analysis of Study 1, it was found that 35% of the participants who did not engage in prosocial behavior with the AI player claimed they were not aware of it. Therefore, a follow-up study was conducted to investigate whether increasing the participants' awareness of the AI player's needs would result in an increase in their prosocial behavior. The results of the follow-up study showed a significant increase in both the overall number of times participants engaged in prosocial behavior with the AI and the frequency of each participant's engagement in such behavior, compared to Study 1. Additionally, from Study 1 to Study 2, the percentage of participants reporting "unawareness" as the reason for not engaging in prosocial interaction with the AI player decreased from 35% to 10%. Furthermore, our results align with previous research on human-robot interaction, which identifies reciprocation as a key aspect behind prosocial behavior between humans and AI (Hsieh et al., 2020; Srinivasan & Takayama, 2016; Zonca et al., 2021; Almeida et al., 2023; Karpus et al., 2021; Oliveira et al., 2020). Reciprocation emerged as the primary reason participants chose to engage in prosocial behavior with the AI in both Study 1 and Study 2.

Our study provides valuable insights into the cognitive factors that contribute to prosocial interactions between humans and AI. We found that when human participants were made more aware of the needs of the AI, they were more likely to engage in prosocial behavior. This suggests that there is a cognitive basis for such behavior. Previous studies on human-robot interaction have identified various aspects that influence prosocial interactions, including incentive structures, reciprocity, reward, and game status(Hsieh et al., 2020; Alves et al., 2020; Gomes et al., 2020). However, our study highlights that attention plays a crucial role in these interactions. Specifically, people's awareness of receiving prosocial interactions from another agent and their attention to that agent's need for assistance are factors that need to be explored further.

Our findings are consistent with Social Cognitive Theory that emphasizes cognitive factors in social interaction between people (Bandura, 1986). Within this theory's context, cognitive functions such as memory and attention are crucial. Individuals must focus on the social model and use memory to store and retrieve representations of observed behavior. Field experiments on prosocial modeling have demonstrated that witnessing one form of prosocial behavior (such as picking up a soda can) can increase the likelihood of engaging in a different prosocial action (like helping to collect oranges that have fallen out of someone's bag) (Keizer et al., 2013). Similarly, Almeida et al. (2023) investigated the impact of different perspective-taking behaviors by robots on human prosocial behavior, further highlighting the underlying cognitive factors in prosocial interaction. Our study adds a new perspective to prosocial behavioral studies by emphasizing the importance of attention in prosocial interactions with AI.

Compared to traditional economic games, the spatial setup of our experiment allows for more ecologically valid interactions among multiple agents (Kleiman-Weiner et al., 2016). However, for future research, it is important to incorporate a more complex setup than the current, relatively simple grid-world. This would further test the role of other cognitive factors, such as memory and learning, in the context of human-AI interactions. Specifically, studies could observe whether witnessing a prosocial action initiated by the AI leads to the human player being more prosocial towards other agents. Future studies should also investigate how human prosocial behavior changes in response to different levels of AI prosociality, and employs trust measures to assess participants' predispositions to demonstrate empathy or trust towards AI player(Chita-Tegmark et al., 2021; Malle & Ullman, 2023).

From an application perspective, our findings suggest that AI design should prioritize clear and understandable behaviors that facilitate human recognition of AI's actions. This approach can lead to more effective and empathetic human-AI cooperation, thereby enhancing the overall quality of interaction. Additionally, the spatial setup of our game offers insights into future scenarios where human and autonomous agent interact in the same spatial environment. This is particularly relevant considering the anticipated increase in the use of delivery and rescue robots in society. Understanding these dynamics is crucial for designing AI systems that can be seamlessly integrated into our daily lives and effectively collaborate with humans in various settings.

In conclusion, our research underscores the critical role of cognitive factors, such as attention, in human-AI interactions. The opportunity to utilize technology to promote prosocial behavior provides a promising path forward. Comprehensive research, thoughtful design, and robust testing are necessary to integrate autonomous agents into our social systems in a manner that enhances overall societal welfare.

References

Almeida, J. a. T., Leite, I., & Yadollahi, E. (2023). Would you help me? linking robot's perspective-taking to human prosocial behavior. In *Proceedings of the 2023 acm/ieee international conference on human-robot interaction* (p. 388–397). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3568162.3577000 doi: 10.1145/3568162.3577000

Alves, T., Gomes, S., Dias, J., & Martinho, C. (2020). The influence of reward on the social valence of interactions. *CoRR*, *abs/2003.12604*. Retrieved from https://arxiv.org/abs/2003.12604

Avelino, J., Correia, F., Catarino, J., Ribeiro, P., Moreno, P., Bernardino, A., & Paiva, A. (2018). The power of a hand-shake in human-robot interactions. In *2018 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 1864–1869).

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ*, *1986*(23-28).

Barch, D. H., & Chechile, R. A. (2023). Dfba: Distribution-free bayesian analysis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=DFBA (R package version 0.1.0)

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in cognitive sciences*, *7*(5), 225–231.

Chater, N., Misyak, J., Watson, D., Griffiths, N., & Mouzakitis, A. (2018). Negotiating the traffic: Can cognitive science help make autonomous vehicles a reality? *Trends in cognitive sciences*, *22*(2), 93–95.

Chernyak, N., & Gary, H. E. (2016). Children's cognitive and behavioral reactions to an autonomous versus controlled social robot dog. *Early Education and Development*, *27*(8), 1175–1189.

Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can you trust your trust measure? *CoRR*, *abs/2104.11365*. Retrieved from https://arxiv.org/abs/2104.11365

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., ... Rahwan, I. (2018, 01). Cooperating with machines. *Nature Communications*, *9*. doi: 10.1038/s41467-017-02597-8

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., ... Graepel, T. (2020). Open problems in cooperative ai. *ArXiv*, *abs/2012.08630*.

Gomes, S., Alves, T., Dias, J., & Martinho, C. (2020). Reward-mediated individual and altruistic behavior. *CoRR*, *abs/2003.09648*. Retrieved from https://arxiv.org/abs/2003.09648

Hsieh, T., Chaudhury, B., & Cross, E. S. (2020). Human–robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots. *International Journal of Social Robotics*. doi: 10.1007/s12369-023-00981-7

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521.

JASP Team. (2024). *JASP (Version 0.18.3)[Computer software]*. Retrieved from https://jasp-stats.org/

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent ai. *Iscience*, *24*(6).

Keizer, K., Lindenberg, S., & Steg, L. (2013). The importance of demonstratively restoring order. *PloS one*, *8*(6), e65137.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. *Cognitive Science*.

Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F., & Rahwan, T. (2023). Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature Communications*, *14*(1), 3108.

Malle, B. F., & Ullman, D. (2023). *Measuring human-robot trust with the mdmt (multi-dimensional measure of trust).*

Martin, D. U., Perry, C., MacIntyre, M. I., Varcoe, L., Pedell, S., & Kaufman, J. (2020). Investigating the nature of children's altruism using a social humanoid robot. *Computers in Human Behavior*, *104*, 106149.

Oliveira, R., Arriaga, P., Santos, F., Mascarenhas, S., & Paiva, A. (2020, 09). Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior*, *114*. doi: 10.1016/j.chb.2020.106547

Schroeder, D. A., & Graziano, W. G. (2015). The field of prosocial behavior: An introduction and overview. In D. A. Schroeder & W. G. Graziano (Eds.), *The oxford handbook of prosocial behavior.* Oxford Academic. Retrieved 2023-07-20, from https://doi.org/10.1093/oxfordhb/9780195399813.013.32

Srinivasan, V., & Takayama, L. (2016). Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 4945–4955).

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), e2111547119. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2111547119 doi: 10.1073/pnas.2111547119

Sung, J.-Y., Guo, L., Grinter, R., & Christensen, H. (2007, 09). "my roomba is rambo": Intimate home appliances. In (Vol. 4717). doi: 10.1007/978-3-540-74853-3_9

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., & Fallah, Y. P. (2021). Cooperative autonomous vehicles that sympathize with human drivers. In *2021 ieee/rsj international conference on intelligent robots and systems (iros)* (p. 4517–4524). IEEE Press. Retrieved from https://doi.org/10.1109/IROS51168.2021.9636151 doi: 10.1109/IROS51168.2021.9636151

Wang, W., Wang, L., Zhang, C., Liu, C., & Sun, L. (2022, nov). Social interactions for autonomous driving: A review and perspectives. *Found. Trends Robot*, *10*(3–4), 198–376. Retrieved from https://doi.org/10.1561/2300000078 doi: 10.1561/2300000078

Zonca, J., Folsø, A., & Sciutti, A. (2021). The role of reciprocity in human-robot social influence. *iScience*, *24*(12), 103424. Retrieved from https://www.sciencedirect.com/science/article/pii/S258900422101395X doi: https://doi.org/10.1016/j.isci.2021.103424