The Effect of Agent-Based Feedback on Prosociality in Social Dilemmas

Jennifer Renoux[*]
Örebro University
Örebro, Sweden
jennifer.renoux@oru.se

Filipa Correia
Interactive Technologies Institute
Lisbon, Portugal
filipacorreia@tecnico.ulisboa.pt

Joana Campos
INESC-ID and Instituto Superior Técnico, University of Lisbon
Lisbon, Portugal
joana.campos@tecnico.ulisboa.pt

Lucas Morillo-Mendez
Örebro University
Örebro, Sweden
lucas.morillo@oru.se

Neziha Akalin
Jönköping University
Jönköping, Sweden
neziha.akalin@ju.se

Fernando P. Santos
University of Amsterdam
Amsterdam, The Netherlands
f.p.santos@uva.nl

Ana Paiva[†]
INESC-ID and Instituto Superior Técnico, University of Lisbon
Lisbon, Portugal

ABSTRACT

Tackling many of humanity's contemporary challenges requires individuals to cooperate in so-called collective risk dilemmas, i.e. scenarios where cooperation is costly yet required to reach collective targets and prevent catastrophic outcomes. It remains a scientific challenge to understand which external incentives enable cooperation and whether that can be facilitated through socially interactive agents. In this paper, we evaluate human cooperation in the presence of an artificial virtual agent. We developed a game called The Pest Control, in which five players attempt to maximize their earnings while avoiding being infested by a spreading pest. Controlling the pest requires costly public good contributions, yet free-riding on the efforts of others leads to maximum individual payoffs. We conducted an online experiment and analyzed the data of 265 participants, where we manipulated the feedback strategy of the virtual agent in a between-subject design. Our results suggest that feedback highlighting salient elements of the game increases participants' cooperation, while feedback regarding the consequences of actions slightly promotes selfish behaviors. Our study provides insight into how future artificial agents and AI systems could be designed to promote cooperation in complex social dilemmas by leveraging different strategies.

KEYWORDS

Prosociality, Collective Risk Dilemma, Human-Agent Interaction, Game Theory, Asymmetric Risk

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

ACM Reference Format:
Jennifer Renoux, Filipa Correia, Joana Campos, Lucas Morillo-Mendez, Neziha Akalin, Fernando P. Santos, and Ana Paiva. 2025. The Effect of Agent-Based Feedback on Prosociality in Social Dilemmas. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

From vaccination to climate change, averting catastrophic events often requires individuals to cooperate for the collective or public good. However, individual behavior is driven by cost-benefit evaluations that may prevent one from selecting actions that increase collective benefits at individual costs [20]. In addition, not all actors in a society are submitted to the same risks and the contribution of individual under low or no risk is necessary to prevent catastrophic outcomes on individuals under high risk [31]. A central question surrounding these types of dilemmas is *what mechanisms (social or technological) can be created in order to make individuals subdue their selfish interests and promote or sustain prosocial behavior?*

Prosocial action involves voluntary acts that are intended to benefit others while incurring costs for the self, and without any guarantees of future reward [3, 9]. Examples of altruistic cooperation gestures include donating money to charity, donating blood, or sharing resources. Although these cooperative acts are common in our society, it is still being determined how they evolved [28], how they can be incentivized, and what is the role of technology in promoting or sustaining them [27]. Previous research attributes the willingness to choose prosocial actions to the interaction between dispositional and situational causes [3], which supports using artificial agents to persuade individuals to cooperate [26, 27]. As suggested by Paiva et al. [27], an artificial agent could make information more salient, enforce norms or create empathetic relations with humans, among other possibilities. Regarding artificial agents' ability to drive humans towards more prosocial actions, a recent survey highlights mixed results [26], where only 52% of 23 studies reported positive effects from interacting with AI systems (mostly

[*] Jennifer Renoux, Filipa Correia, and Joana Campos have contributed equally
[†] Ana Paiva has not been actively working on the work since April 5th, 2024.

robots), and other pointed either towards no effect or mixed results. Very little is known about *how* artificial social agents can persuade humans to be prosocial towards others and, in particular, how *virtual agents* could be leveraged.

This paper addresses this challenge with an empirical evaluation of virtual agents as assistants in a social dilemma, assessing the effectiveness of different strategies on people's prosocial actions. In particular, we took inspiration from feedback strategies that have shown potential in the literature of interpersonal relationships and decision-making tasks. Feedback is conceptualized as knowledge about *results* of a behavior and the *process* of engaging in that behavior [15], and is able to draw attention to something self-relevant [16]. Its adoption by artificial agents has shown to be efficient in regulating behaviors [25], but its effect during decision-making tasks involving social dilemmas is still limited.

We developed two feedback strategies for a virtual agent, modeled after effective communication strategies previously observed in human players during public good games [17]: Problem Awareness (PA) and Player Strategies (PS). Koessler et al. [17] have first analyzed the communication patterns among humans, identifying four relevant ones, which included the PA, as attempts to find a common understanding of the problem at hand, and PS, as the discussion of the possible ways to tackle the problem. Then, the authors empirically tested the access to each type of communication strategy among humans, and some of their combinations (such as PA+PS), and found them to be efficient in improving collaboration [17].

Therefore, we investigated the following research question: **are feedback strategies naturally used by human players in a public good game efficient when given by an artificial agent?**

To answer this research question, we developed the *Pest Control Game*, based on the scenario proposed by Reeves et al. [31]. The game captures the dilemma of cooperating to avert collective losses in a spatial setting where risk is asymmetric [24, 31]. The game is a public good game in which five players attempt to gather as many coins as possible while preventing their farm from being infested by a spreading pest. The pest may spread every year (i.e. game round) if not controlled, and farmers can spend coins to reduce the chance of pest spreading. Throughout the game, the virtual agent is a *non-playing character* that prompts information about game mechanics (PA) and/or possible strategies a human player could employ (PS), making action consequences more salient. Supported by the results from interpersonal groups [17], we hypothesized that the agent's presence and the type of information it conveys will persuade players to contribute more to the farmers' collective.

We conducted a between-subjects experiment (N = 265) manipulating the feedback strategy of the virtual agent, comparing the strategies PA, PS, and PA+PS against a control condition in which the agent provided no feedback. Results showed a positive effect of the PA strategy. When the virtual agent raised awareness about the social dilemma, participants contributed more immediately after having received that feedback. However, this effect did not sustain throughout the whole game, suggesting a short-term positive effect of this strategy.

2 RELATED WORK

Games are important paradigms for studying human behavior. Well-defined economic games, such as trust and dictator games, have been extensively used to study the effects of small variations on individuals' behavior in laboratory experiments [1, 6, 13, 33]. More recently, researchers have been exploring what could be the role of artificial agents in these social dilemmas and how they should behave so that humans increase and sustain levels of cooperation in groups. Works can be divided into two main categories: *acting in the social world* and *providing information* to assist decision-making. Shirado and Christakis [35] explored the impact of adding individual actors to small networks and making them interact with their neighbors in a public goods game. They found that agents using simple decision-making models can act locally in the network and help humans to develop more cooperative actions. In another public goods game where players were aware that other players were artificial agents, Tulli et al. [37] found that transparency following agent's actions did not influence the cooperative choices of players, but the action themselves did as people were more cooperative towards cooperative agents.

Transparency[1] and feedback are two different mechanisms for *providing information* to humans and trying to induce a positive behavior change. Transparency is regarded as a fundamental mechanism for collaboration and mutual awareness between humans and agents [32]. However, its applicability may depend on the characteristics of the environment (e.g. risk asymmetry). Its explanatory nature describing the inner workings of autonomous agents (as in [37]) may support belief formation about the interacting parties, yet how people use this information may facilitate or impede cooperation [12]. It may encourage, for instance, *free-riders* in social dilemmas.

Artificial agents that provide feedback or nudges are designed *to create a social or individual benefit* without acting directly in the environment. This type of technology makes it possible to collect behavioral data, deliver dynamic information about goal-achievement status, and use social cues in order to elicit social responses from humans [18]. Agents that provide feedback show promising results in increasing individuals' motivation and helping people to achieve their goals [21], but it is missing from the literature how we can leverage forms of feedback provided by artificial agents to enhance prosocial action. As previously stated, feedback has been conceptualized in the literature as knowledge about *results* of a behavior and the *process* of engaging in that behavior [15]. Our work attempts to study whether exposing individuals to factual information conveyed by an artificial agent – attempting to exert *informational social influence* [7] [2]– throughout a social game will make people engage in prosocial action. We base our assumption on a large body of research that studies the use of environment descriptions [3] in our society to promote or inhibit a behavior [5, 14, 23, 39] and argue that artificial agents could actively describe critical elements of social interactions to promote prosocial action.

[1] Here *transparency* refers to the knowledge available after an action has been taken

[2] *Informational social influence* refers to one accepting information from another as evidence about reality.

[3] In particular the use of signs, such as symbols or prompts, as a device to communicate directives

In social dilemmas, people also provide directives for action, when communication is allowed, and that has been widely known as a solution to cooperation [2]. One of the underlying reasons is that communication helps individuals coordinate beliefs [2], but it is still unclear *what* information people exchange and *how* and the different types of information impact decisions. Koessler et al. [17] looked into social dilemmas with free form communication and extracted four factors that foster cooperation in those settings. Regarding information that is shared they identified that people tend to clarify the dilemma structure and the game mechanics (*Problem awareness*), eliciting the various strategies and their consequences (*Identification of strategies*).

In our work, we do not investigate the use of an interactive agent as a member of the group; instead, we focus on an actor that conveys factual information about the environment within a social dilemma that lacks communication. We explore how having an agent that provides information about the dilemma structure, the game mechanics, possible strategies, and their consequences in a complex setting – similarly to what humans do –, affect how individuals construct their belief systems when making decisions [17] and their tendency to act prosocially.

3 THE PEST CONTROL GAME

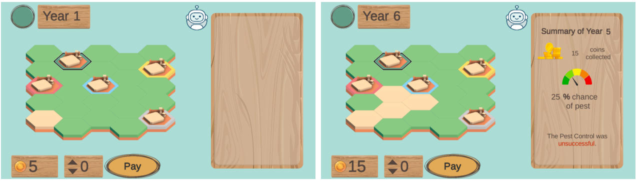
3.1 Overview

The pest control problem [31] is a game theoretic model of cooperation with risk asymmetry. N farms are situated on a 1-dimensional lattice point. In the first point, a pest infestation directly threatens the neighboring farmer. To prevent the pest from spreading, farmers can contribute to a collective fund each year (i.e. each round). The number of coins gathered by the collective will predict the likelihood of the pest control to be successful, calculated following the pest control function:

$$p(c) = \frac{kc}{1 + kc} \tag{1}$$

where c is the total amount of coins gathered by the collective this round, and $k > 0$ is a fixed parameter that quantifies how easy it is to control the pest. Reeves et al. studied the pest control problem from a game-theory perspective and concluded that, due to the risk asymmetry (i.e., farmers located farther away from the pest face a lower risk of infection), in the Nash equilibrium, the farmers closest to the pest pay significantly more to the collective [31].

Based on Reeves et al.'s version of the pest control problem we created a public good game called the *Pest Control Game*[4]. In the *Pest Control Game*, five players situated in a 2-dimensional grid attempt to gather as many coins as possible while preventing their farm from being infested by the spreading pest. Each player starts the game with five coins in their wallet, and the pest is located on the tile at the bottom left corner of the game board (Figure 1a). Each year, players first select how many coins they want to contribute to the collective fund. The sum of all the coins determines the probability of the pest control success according to Equation 1. If the pest control succeeds, the pest does not spread. If it fails, the pest spreads to an adjacent tile, selected randomly. If the pest spreads in a player's farm, this player is eliminated from the game and loses all their previously accumulated resources. Here lies the dilemma: investing in pest control averts collective losses, although players maximize their payoffs if they do not invest and remain safe through the contributions of others. Finally, all players still active in the game at the end of the turn (i.e., who did not see their farm infested by the pest) receive five coins in their wallet, and a new turn begins. The game ends at the end of year 15 or earlier if all players lose their farms.



(a) (b)

Figure 1: The *Pest Control Game* interface on year 1 (a) and 6 (b). The area on the right-hand side of the game is used to display a summary of what happened in the previous round.

3.2 Choosing the Risk Level

The main factor influencing the game and the prosocial behavior of participants is how controllable the pest spread is. On one hand, if the pest is too easy to control, participants have little reason to act prosocially as a small amount of money collected would suffice to stop the spread. On the other hand, if the spread is too difficult to control, participant's prosociality becomes irrelevant, and they may decide that contributing is not worth it. We conducted simulations to figure out an appropriate value for the pest controllability (k in Equation 1). We decided to choose $k = 0.2$ as the value given the best balance of risk and potential impact given the number of rounds in the game (15), size of the environment, and the endowment (5 coins) provided to players at each round. Indeed, higher values of k present too sharp a slope and allow for too many players to free-ride without consequences. On the other side, lower values of k does not allow for successful pest control even if all players contribute a significant amount of their endowment to the collective. In this setup, we also considered that a fair action would be a contribution of 3 coins, bringing the total amount of coins gathered to the collective to 15 if all players play fair, and therefore the chance of the pest control to be successful to 0.75. In this setup, whether the pest spreads or not follows a binomial distribution with a probability $p = 0.25$. Therefore, the expected amount of pest spread during whole game if all players always play fair is $E = n * p$ where $n = 15$ and $p = 0.25$, i.e. $E = 3.75$. Given that the closest farms are 2 and 3 tiles away from the pest, this means that some players are at risk but relatively safe given that other players do not free-ride.

3.3 Pat the Bot

Pat the Bot is an artificial non-playing agent who provides information about the game to the players. It is represented by an avatar to have a social presence. We chose a minimalistic appearance to avoid as much as possible interaction from social cues. Therefore,

[4] Implementation available at https://github.com/jrenoux/pest-control-game-source and playable demo version available at https://jrenoux.github.io/pestcontrolgame/demo/ (use 123456789123456789123456 as ID)

Pat the Bot has two representations: one "silent" when it does not give feedback and one "talking" when the feedback chat is activated (see Figure 2). In alternating rounds, Pat provides feedback about the *process* of engaging in prosocial behavior as well as expository information of the task itself, as studied in [15]. We designed Pat the Bot to expose the same two categories of information as Koessler et al. [17]:

(a) Silent state.

(b) Talking state.

Figure 2: Pat the Bot in silent and talking states

- **Problem Awareness (PA):** Information about the social dilemma dynamics. It highlights important concepts of the social dilemma: *Bonus Payment*, *Collective* and *Pest Spread*.
- **Player Strategies (PS):** Information about financial consequences of individual and joint courses of actions, including factual information about the *Joint Contributions* and *Individual Contributions* in terms of consequences, and the actual impact of *Players Actions*.

Pat the Bot gave feedback to the players at the end of the year, after the player selected their contribution and the pest spread had been performed. Each of the 6 PA and PS concepts were associated to 2 utterances, with a complete list given in Table 1. The sentences were designed based on the work of Kluger and DeNisi [16], which identifies 3 moderators of effect: standards, goals, and attention. As described in [15], *Standards* refer to personal goals or comparisons with past behaviors; *Goals* sets the importance of the feedback for the individual; and *Attention* directs to focal task goals. Feedback should focus on standards connected to self-related goals [15]. Two things are important to note regarding the utterances. First, even though some concepts relate to the same element (Collective and Joint Contributions for instance both relate to the amount of coins gathered), the utterances are formed to convey different information in the two types of communication. In PA (Collective), the utterances have been created to put the accent on the game mechanics, while in PS (Joint Contributions), the utterances have been created to put the accent on the consequence of certain actions. Second, all the information given by Pat the Bot is information that the player should already know (from the game's explanation or tutorial) but not available otherwise through the board.

4 EXPERIMENTAL SETUP

Participants were invited to individually play a game of pest control with four other players, which the participant was led to believe were humans. However, we decided to control for the impact of how other players play the game, by having scripted artificial agents performing the decisions. The agents were contributing a mostly fair amount, 3 coins per turn on average, with minor variations each turn and adapting to their distance to the pest, i.e. agents close or adjacent to a pest tile would contribute more. This behavior allows free riding for the human participant. We also controlled the pest pattern and spread of the game, which were entirely scripted, meaning the participants' actions had no impact on the game outcome and only impacted how much coins remained in their wallet at the end of the game. In our scenario, the participant "wins" the game, i.e. their farm survives the pest. However, one of the artificial agents sees its farm being taken by the pest in round 7. Table 2 summarizes the pest spread speed, the events happening during the game, and the rounds where Pat the Bot gave feedback. Participants were randomly assigned to one of four conditions:

- **Control Condition:** The participant did not receive any information from Pat the Bot and was not even made aware of its existence.
- **PA Condition:** The participant only received information of the *Problem Awareness* type, and received two utterances for each PA concept displayed in a random order.
- **PS Condition:** The participant only received information of the *Player Strategies* type, and received two utterances for each PS concept displayed in a random order.
- **PA+PS Condition:** The participant received information from both *Problem Awareness* and *Player Strategies* types, and received one utterance for each PA and PS concepts, chosen randomly and displayed in a random order.

4.1 Experimental Protocol

Participants were recruited online through the Prolific[5] platform with the screening criteria of being fluent in English. The participants were first presented with a summary of the study, which gave them basic information while omitting the fact that other players were artificial agents and that the study game was fully scripted. They were also told that the "money" (in the form of coins) they gathered during the game would be converted into a bonus payment, to increase self-interest. Upon accepting to participate, the participants were redirected towards a Labvanced[6] document that automatically assigned them to one of our four conditions. They were then directed toward the game itself.

The participants were first shown a tutorial explaining the key components of the game and played a tutorial game, during which the pattern of the pest spread was scripted to ensure that participants were as little primed as possible. However, the probability of the pest spreading was controlled by the participant's contribution to allow them to evaluate the impact of their action. The participants were informed that during the test game, other players were artificial agents who would contribute exactly as much as them. Upon completing the tutorial game, the participants were then told that they were connected with other human players to play the study game. Upon completing the study game, the participants were redirected toward the Labvanced document to answer a questionnaire containing attention checks, demographic questions, two questions related to their comprehension of the game's rules, and one related to their impression of other players (whether they were humans or not). After completed the questionnaires, participants were redirected towards a debriefing page where they were informed about the deception (the fact that other agents were not humans and

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

[6] https://www.labvanced.com/

Type	Concept	Utterance
PA	Bonus Payment	(1) If the pest does not get to a player, all the money in their wallet will be converted into a bonus in the end.
		(2) When a player reaches the final year without being caught by the pest, the coins in their wallet are converted into a bonus.
	Collective	(3) The more farmers contribute to the collective, the lesser the probability of pest spreading.
		(4) The number of coins collected each year impacts the probability of pest spreading.
	Pest Spread	(5) When the pest arrives to a farm, it gets destroyed and that farmer loses the game. The remaining farmers then need to contribute more to the collective to prevent pest spreading.
		(6) A farmer loses the game when the pest gets to their tile. Having less farmers means that the individual contributions to the collective must be higher to maintain a low risk of pest spreading.
PS	Joint Contrib.	(1) Note that if each farmer contributes 3 coins per year, the risk of pest spreading will reduce to 25%.
		(2) Note that no contribution to the collective will increase the probability of pest spreading.
	Players actions	(3) The collective has gathered $< collectiveAVG >$ coins, on average, per year.
		(4) Each farmer has contributed $< farmerContributionAVG >$ coins, on average, per year.
	Individual Contrib.	(5) Your contribution directly affects the risk of pest spreading and at the same time your final bonus.
		(6) How much you contribute each year affects both your bonus and the probability of pest spreading.

Table 1: Sentences given by Pat the Bot during the game. The values of $< collectiveAVG >$ (resp. $< farmerContributionAVG >$) represents the average amount of coins collected by the collective (resp. average contribution per farmer) up to the point where this utterance is sent. These amounts are calculated during the game for each participant.

Year	Pest spread?	Event	Feedback?
1	Yes		Yes
2	Yes	Blue is threatened	No
3	No		Yes
4	No		No
5	Yes	Red is threatened	Yes
6	No		No
7	Yes	Blue is infested	Yes
8	Yes	Grey is threatened	No
9	No		Yes
10	No		No
11	Yes		Yes
12	No		No
13	No		No
14	Yes	Yellow is threatened	No
15	Yes		No

Table 2: Summary of the pest spread, outcomes, and feedback. "Threaten" means that the pest becomes adjacent to a player's farm, and "Infest" means that the pest infests the location of a player's farm, thus eliminating the player from the game. Feedback is given at the end of the year after the player selected their contribution and the pest spread is completed.

that the game was scripted). Participation in Prolific was reviewed based on the attention checks according to Prolific Guidelines. Participants who failed had their submission rejected and did not get compensated. Participants whose submissions were accepted were paid GBP 3.75, plus a bonus payment of GBP $0.02 \times final\_wallet$, where $final\_wallet$ is the number of coins the participant gathered at the end of the study game. The maximum bonus payment that a participant could receive was GBP 1.6. The Ethical Committee of the Instituto Supérior Tecnico approved this protocol.

4.2 Hypothesis

Koessler et al. [17] suggested a positive effect of each strategy on cooperation with the following order : Control < PA < PS < PAPS. Based on these findings, we tested the following hypothesis :

H1: The different types of feedback from the artificial non-playing agent positively affect the participants' prosocial actions compared to the control condition. More specifically, we expect the cooperation to increase according to the following order: Control < PA < PS < PAPS.

4.3 Measures

We measured prosocial actions with two metrics: the **final wallet**, which is the number of coins remaining in the wallet of the participant at the end of the game; and the **average contribution after a feedback round**, which is the average number of coins spent in a single round and included data of the rounds directly following feedback in the experimental conditions, see Table 2. A lower final wallet indicates that the participant has contributed more during the game, and constitutes a more long-term measurement of prosocial actions within our experiment. The average round contribution refers to a more immediate behavior of the participants.

4.4 Sample

Sample size calculation indicated that 231 participants would allow for a power of 0.9, assuming a medium effect size. We obtained 363 unique valid submissions, out of which 18 got removed due to various issues (e.g. test experiments, submission not completed, game data corrupted, duplicate submission). In addition, 5 participants got excluded for failing the attention checks, and 75 got removed from the data for failing on items showing that they might not have understood the game's rules. The final sample included 265 participants (104 Female, 156 Male, 5 Non-binary), divided between conditions as follows:

- **Control:** 51 participants (19 Female,31 Male, 1 Non-binary)
- **PA:** 76 participants (34 Female, 40 Male, 2 Non-binary)
- **PS:** 53 participants (16 Female, 36 Male, 1 Non-binary)
- **PA+PS:** 85 participants (35 Female, 49 Male, 1 Non-binary)

The participants' ages ranged from 18 to 64 ($M = 29, SD = 9.46$). To evaluate if participant's beliefs about other players changed their way of playing, we added in the post-game questionnaire the question: "Were the other players' decisions controlled by the computer?", where possible answers were "Yes", "No", and "Not sure". We then looked at the average final wallet per possible answer and did not find any difference, thus suggesting that participants did not change their overall way of playing even if they thought the other players were artificial agents. Therefore, we included the whole sample in our analysis.

4.5 Analysis

We used the non-parametric Kruskal-Wallis test to compare differences among the four conditions, due its robustness. Pairwise comparisons used the Mann-Whitney U test[7].

5 RESULTS

For the final wallet metric, we did not find a statistically significant difference between our four conditions ($H(3) = 5.720, p = .126$). Figure 3 shows a graphical representation of the average final wallet value per condition ($M_{Ctrl} = 35.69, SD_{Ctrl} = 18.05$; $M_{PA} = 30.51, SD_{PA} = 16.43$; $M_{PS} = 36.85, SD_{PS} = 18.97$; $M_{PA+PS} = 32.54, SD_{PA+PS} = 16.93$). For the second measure of average contribution after a feedback round, we found a statistically significant difference between our four conditions ($H(3) = 18.767, p < .001$). Figure 4 shows a graphical representation of the average final wallet value per condition ($M_{Ctrl} = 3.22, SD_{Ctrl} = 1.70$; $M_{PA} = 3.62, SD_{PA} = 1.70$; $M_{PS} = 3.13, SD_{PS} = 1.61$; $M_{PA+PS} = 3.37, SD_{PA+PS} = 1.63$). The additional pairwise comparisons indicate participants contributed more immediately after receiving a PA feedback, compared to all other conditions (PA vs. Ctrl: $U = 61072.5, Z = -2.968, p = .003, r = 0.3$, medium effect; PA vs. PS: $U = 60349.5, Z = -4.049, p < .001, r = 0.4$, medium effect; PA vs. PA+PS: $U = 106722.5, Z = -2.252, p = .024, r = 0.2$, small effect). The pairwise conditions between PA+PS and PS was also statistically significant ($U = 73974.5, Z = -2.176, p = .030, r = 0.2$, small effect).

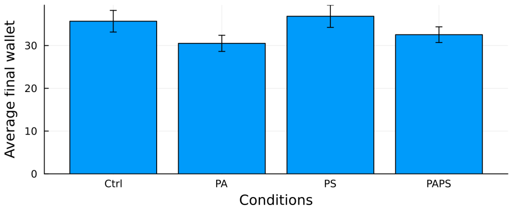


Figure 3: Average amount on the final wallet per condition. The error bars represent the standard error of the mean.

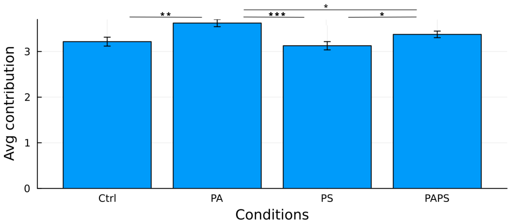


Figure 4: Average number of coins contributed to the collective immediately after a feedback round. Error bars represent the standard error of the mean.

6 A POSTERIORI EXPLORATORY ANALYSIS

Although we found partial support to our hypothesis, when considering the PA strategy, we decided to explore in depth the impact of PS type of feedback. In particular, we wondered how each of the feedback sentences affected participants' prosocial actions. As explained before, we created 6 feedback sentences for the PA strategy and 6 for the PS strategy (see Section 3.3). Participants in conditions PA and PS were exposed to all the possible feedback sentences in a randomized order. However, participants in the PA+PS condition received 3 random sentences of PA type and 3 random sentences of PS type. As a result, for this exploratory analysis, we created new boolean independent variables to mark whether each participant had (or not) received each possible feedback sentence by the virtual agent. Then, we compared the final wallet amount of participants according to each independent variable, i.e. having or not received each feedback sentence.

For the PA feedback sentences, the results show that, among the 6 possible sentences, 3 of them had a statistically significant impact on the final wallet, specifically the PA-3 ($U = 6700, p = .002$), PA-5 ($U = 6872.5, p = .005$), and PA-6 ($U = 7411.5, p = .04$). We did not find a significant difference of the presence of feedback sentences PA-1 ($U = 8308.5, p = .51$), PA-2 ($U = 7600, p = .08$), and PA-4 ($U = 8119, p = .32$). It is worth noting that for each possible PA sentence, the average amount in final wallet was lower for participants that received the sentence compared to those that did not receive it (see Figure 5). In other words, all the created PA sentences show at least a tendency to increase people's contributions to the collective (leaving them with less money in their final wallet).

[7] Data and code for the full analysis are available at https://osf.io/axnjk/

For the PS feedback sentences, we only found a statistically significant difference for sentence PS-6 ($U = 9424.5, P = .04$). This difference shows that participants that received feedback sentence PS-6 had less prosocial actions (i.e. contributing less to the collecting and keeping more money on their individual final wallet), than participants who did not receive the PS-6 sentence (contrarily to the effect of PA sentences). We did not find significant differences for the other PS sentences, specifically PS-1 ($U = 8517, p = .46$), PS-2 ($U = 7936, p = .92$), PS-3 ($U = 8911, p = .13$), PS-4 ($U = 8449.5, p = .66$), nor PS-5 ($U = 8849.5, p = .22$). It is also noteworthy that, when observing Figure 5, there is a tendency that the presence of PS feedback sentences led participants to be less prosocial (contrarily to the observed effect of the PA feedback sentences).

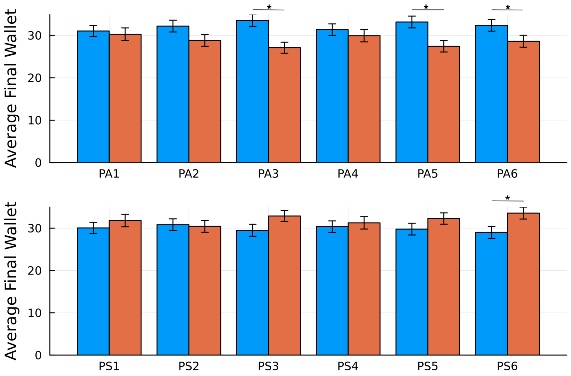


Figure 5: Average final wallet for each of the sentence, grouped by whether or not the participant received a given sentence (left = did not receive, right = received). Error bars represent the standard error of the mean.

We conducted a post-hoc power analysis to estimate the power of each reported effect. For the pairwise comparisons, which were more susceptible to being underpowered, we obtained the following results. PA-Ctrl: power=0.75 PA-PS: power=0.76 PA-PA+PS: power=0.44 PS-PA+PS: power=0.38. This suggests the first two comparisons have significant power, while the other two are underpowered. Nevertheless, we would like to emphasize that the discussion of our results (Section 7) is mostly focused on the first two differences.

7 DISCUSSION

7.1 Discussion of the Findings

The existing literature shows discrepancies in the efficiency of feedback strategies [17, 18, 21, 37], namely both on feedback-giving agents outside of public good games and on feedback given by humans playing public good games. As a result, we pioneered the investigation of feedback strategies provided by a socially interactive non-playing agent in a public good game to test its efficiency in promoting prosociality. Drawing on Koessler et al. [17]'s findings about how people communicate to promote greater cooperation in groups, we integrated these strategies into an artificial feedback-giving agent and hypothesized that participants' prosocial behaviors in each of our conditions would vary in the following order: Control < PA < PS < PAPS.

Feedback on *problem awareness* (PA) is expository of important concepts of the game: player's *wallet*, the importance of the *collective*, and the impact of *pest spread*. **Our results indicated that PA feedback was effective in improving prosocial behaviors, but only immediately after receiving the feedback.** Although participants generally contributed more to the collective in the following round of a feedback by the virtual agent, such prosocial acts did not sustain throughout the whole game and, therefore, did not affect the final amount kept in participants' individual wallet. The positive effect of PA feedback strategy partially supported our hypothesis when comparing the PA and Control conditions, suggesting that feedback strategies that highlight the collective issues can be effective. However, considering how challenging it can be to sustain such effectiveness over time, we would like to discuss on important considerations to inform future studies on feedback-giving agents to foster prosociality. First, one critical aspect that may play a role in these outcomes is the level of agency attributed to the agent. The lack of social cues of our agent may have affected the success of its feedback strategies [21]. Second, our agent was not a player in the public good game, and previous work concluded that the type of communication employed in the Problem Awareness type is efficient when used among peers, but not when given by an expert [4]. Pat the Bot might have been seen as an expert by the players due to its non-playing nature and its extensive knowledge of the game. Additionally, all the information provided by Pat the Bot was something the player should have already known from the game instructions. Players may have felt it was repeating information they already knew and disregarded it. However, artificial agents are becoming more integrated into society and assisting in decision-making. Therefore, it is important to discuss the findings of this study in the context of design of these technologies. Finally, the number of times the virtual agent provided feedback may have also been insufficient to sustain the effect over throughout the entire game. We recommend future studies further explore the agent's interactivity, how humans perceive the agent's role, and an increased amount of feedback reinforcement provided by the agent.

Problem strategies (PS) focus on game mechanics, players' actions and individual gains without evoking norms. **We did not find support that the PS feedback type improves prosocial behaviors compared to both the Control and PA conditions.** In fact, our exploratory analysis suggests that this type of feedback might have had a negative impact on participants' level of cooperation. Most feedback sentences used by the agent showed a tendency to negatively affect prosocial acts, and the last sentence was even able to produce a significant difference. We noticed that the specific sentence PS-6 highlights the tension between the pest spreading and the individual bonus. This finding can be counterintuitive given the fact that such type of feedback emerges naturally between human participants [17]. We designed our sentences to convey three moderators of effect (Section 3.3): standards, goals, and attention. However, the feedback is not triggered by a player action, and thus does not induce a feedback-standard gap[8]. Furthermore, our manipulation of the PS feedback integrated concepts on individual gains that might have highlighted self-interest behaviors when mentioning the bonus. Future studies can further explore improved manipulations of PS feedback strategies, trying to highlight all concepts in the same feedback sentence (instead of in difference sentences), or trigger each sentence in accordance to the human's previous action.

Finally, our PAPS condition mixed feedback of both PA and PS types. **We did not find support that PAPS feedback can promote prosocial actions compared to the Control, PA, nor to the PS conditions.** Considering the PAPS feedback was a combination of both PA and PS, we believe these results balance, to a certain extent, the outcomes obtained by the agent that employs individually the PA and PS feedback types, as discussed above.

7.2 Implications and limitations of the design

Several of the decisions made during the experiment design likely had an impact on the results obtained and must be discussed. In particular, we will discuss the choice of the gamified setting, the spatiality of the game, and the choice to include a visualization of the feedback-giving agent (Pat the Bot).

First, our study uses a gamified setting in which participants interact through a web-based platform, with static virtual partners that do not adapt their behavior to the player's. This may have impacted the prosociality of participants, though we did not find any evidence that our participants played differently based on their belief that the other players were humans or not (Section 4.4).

Second, we design the game to take place in a 2-dimensional version of the pest control problem explored by Reeves et al. [31], thus introducing spatiality, rarely explored in public good games. The main characteristic of the pest control problem (and therefore the Pest Control Game we developed) is the presence of risk inequality, represented by the varying distances between each farmer and the pest. This asymmetry reflects societal dynamics, where individuals face differing circumstances that shape their group decisions. Unequal exposure to the pest creates diverse incentives, opening up opportunities for exploitation and thus clear acts of prosociality. As noted in [30], spatial representations help participants understand the context of their actions and facilitate deeper insights into cooperative and competitive dynamics. Although we acknowledge the spatial properties are not being fully explored in our experiment, their complexity paves the floor for new research avenues.

Finally, we chose to include Pat the Bot, compared to an alternative design where the interface display the same messages without the visual presence of the agent. Drawing on prior research on prosocial and honest behaviors influenced by the presence of agents [29, 34], we speculate that the results of such an alternative design differs. Indeed, these findings suggest that the perception of being observed by a visually present agent significantly impacts behavior, thus justifying our decision to include it in our design.

8 CONCLUSION

In this paper, we investigated the impact of feedback from an artificial agent on human prosocial behavior. Our primary goal was to examine whether two types of feedback, commonly found in human-human interactions, could be effective when delivered by an artificial agent. We found no significant differences between the designed conditions in the players' final wallets (*long-term effect*), which can be attributed to several factors. These factors highlight key design challenges (**DC**) for an agent aiming to make elements of the social game and the consequences of the agent's actions more salient in a social dilemma. We found that delivering feedback throughout an interaction, without considering the timing or frequency, may not be sufficient (**DC1**). Our results indicated that PA feedback was effective in improving prosocial behaviors immediately after receiving the feedback (*short-term effect*) (**DC2**). We also found that delivering feedback focusing on self and other gains, without evoking social norms, shows a negative impact on participants' level of cooperation (**DC3**). From a design perspective, agents aimed at enhancing cooperation among humans should foster more transparent and accountable interactions. This involves highlighting both static and strategic elements of the interaction and finding ways to address **DC2** and **DC3**.

The game we present is framed as a pest control scenario, but the mechanisms for eliciting cooperation can apply to various real-world interaction paradigms involving cooperation dilemmas in spatial settings, especially those with asymmetric risk exposure. Pest control on a farm is one example of a broader category of dilemmas in risk management, where cooperation among individuals with differing risk exposure—such as proximity to water sources in drought or flood mitigation—is essential for collective success and preventing future losses. Other real-world examples of spatial cooperation dilemmas with heterogeneous agents include the challenge of forest cleaning and wildfire management, green technology adoption in an urban setting [10], preventing illegal logging [19] and poaching [11], or even the global challenge of investing in climate change mitigation and adaptation infrastructure [24].

This paper explores how AI agents' communication affects human behavior. While our use case may yield positive societal changes, similar research can also be used for negative purposes, such as scams or misinformation. Therefore, the benefits of this fundamental research must be weighed against its risks. We believe that understanding human interactions with AI systems can enhance positive applications and help us better identify harmful influences, enabling more informed decisions both individually and collectively. For a deeper discussion on the topic, we refer to to [22] and [36]. In addition, the development of systems using the results of our research must be carried out within ethical and responsible development processes [8]. Methodological approaches such as Design for Values [38] can be useful when designing such systems and considering potential ethical issues.

ACKNOWLEDGMENTS

This work has been supported by the European Commission, under the Horizon HumanE AI Network project, grant agreement 952026, the Horizon AI4Europe project, grand agreement 101070000, by CRAI C628696807-00454142 (IAPMEI/PRR) and by FCT under 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, 10.54499/ UIDB/ 50009/2020, and 2022.00816.CEECIND/CP1713/CT0013.

[8] A feedback-standard gap occurs when the behavior differs from the standard [15]

REFERENCES

[1] Konstantinos C. Apostolakis, Kyriaki Kaza, Athanasios Psaltis, Kiriakos Stefanidis, Spyridon Thermos, Kosmas Dimitropoulos, Evaggelia Dimaraki, and Petros Daras. 2016. Path of Trust: A Prosocial Co-op Game for Building up Trustworthiness and Teamwork. In *Games and Learning Alliance*, Alessandro De Gloria and Remco Veltkamp (Eds.). Springer International Publishing, Cham, 80–89.

[2] Daniel Balliet. 2010. Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution* 54, 1 (2010), 39–57.

[3] C Daniel Batson and Adam A Powell. 2003. Altruism and prosocial behavior. (2003).

[4] Jordi Brandts, Christina Rott, and Carles Solà. 2016. Not just like starting over-Leadership and revivification of cooperation in groups. *Experimental economics* 19, 4 (2016), 792–818.

[5] Tim Bungum, Mindy Meacham, and Nicole Truax. 2007. The Effects of Signage and the Physical Environment on Stair Usage. *Journal of Physical Activity and Health* 4, 3 (2007), 237 – 244. https://doi.org/10.1123/jpah.4.3.237

[6] François Cochard, Julie Le Gallo, Nikolaos Georgantzis, and Jean-Christian Tisserand. 2021. Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game. *Journal of Behavioral and Experimental Economics* 90 (2021), 101613. https://doi.org/10.1016/j.socec.2020.101613

[7] Morton Deutsch and Harold B Gerard. 1955. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology* 51, 3 (1955), 629.

[8] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer.

[9] Nancy Eisenberg, Tracy L. Spinrad, and Ariel Knafo-Noam. 2015. *Prosocial Development*. John Wiley & Sons, Ltd, Chapter 15, 1–47.

[10] Sara Encarnação, Fernando P Santos, Francisco C Santos, Vered Blass, Jorge M Pacheco, and Juval Portugali. 2016. Paradigm shifts and the interplay between state, business and civil sectors. *Royal Society open science* 3, 12 (2016), 160753.

[11] Fei Fang, Peter Stone, and Milind Tambe. 2015. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *Twenty-fourth international joint conference on artificial intelligence*.

[12] Lenka Fiala and Sigrid Suetens. 2017. Transparency and cooperation in repeated dilemma games: a meta study. *Experimental economics* 20, 4 (2017), 755–771.

[13] Alexis Garapin, Laurent Muller, and Bilel Rahali. 2015. Does trust mean giving and not risking? Experimental evidence from the trust game. *Revue D Economie Politique* 125 (2015), 701–716. https://api.semanticscholar.org/CorpusID:55286334

[14] Sherry Huybers, Ron Van Houten, and J. E. Louis Malenfant. 2004. REDUCING CONFLICTS BETWEEN MOTOR VEHICLES AND PEDESTRIANS: THE SEPARATE AND COMBINED EFFECTS OF PAVEMENT MARKINGS AND A SIGN PROMPT. *Journal of Applied Behavior Analysis* 37, 4 (2004), 445–456. https://doi.org/10.1901/jaba.2004.37-445 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1901/jaba.2004.37-445

[15] Beth Karlin, Joanne F Zinger, and Rebecca Ford. 2015. The effects of feedback on energy conservation: A meta-analysis. *Psychological bulletin* 141, 6 (2015), 1205.

[16] Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin* 119, 2 (1996), 254.

[17] Ann-Kathrin Koessler, Juan Felipe Ortiz-Riomalo, Mathias Janke, and Stefanie Engel. 2021. Structuring Communication Effectively—The Causal Effects of Communication Elements on Cooperation in Social Dilemmas. *Environmental and Resource Economics* 79, 4 (2021), 683–712.

[18] Jan Krátkỳ, John J McGraw, Dimitris Xygalatas, Panagiotis Mitkidis, and Paul Reddish. 2016. It depends who is watching you: 3-D agent cues increase fairness. *PloS one* 11, 2 (2016), e0148845.

[19] Joung-Hun Lee, Karl Sigmund, Ulf Dieckmann, and Yoh Iwasa. 2015. Games of corruption: How to suppress illegal logging. *Journal of Theoretical Biology* 367 (2015), 1–13.

[20] Simon A Levin. 2014. Public goods in relation to competition, cooperation, and spite. *Proceedings of the National Academy of Sciences* 111, Supplement 3 (2014), 10838–10845.

[21] Lijia Lin, Robert K Atkinson, Robert M Christopherson, Stacey S Joseph, and Caroline J Harrison. 2013. Animated agents and learning: Does the type of verbal feedback they provide matter? *Computers & Education* 67 (2013), 239–249.

[22] Yiling Lin, Magda Osman, and Richard Ashcroft. 2017. Nudge: concept, effectiveness, and ethics. *Basic and Applied Social Psychology* 39, 6 (2017), 293–306.

[23] Julia Meis and Yoshihisa Kashima. 2017. Signage as a tool for behavioral change: Direct and indirect routes to understanding the meaning of a sign. *PloS one* 12, 8 (2017), e0182975.

[24] Ramona Merhej, Fernando P Santos, Francisco S Melo, and Francisco C Santos. 2021. Cooperation between independent reinforcement learners under wealth inequality and collective risks. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 898–906.

[25] Cees Midden and Jaap Ham. 2009. Using negative and positive social feedback from a robotic agent to save energy. In *Proceedings of the 4th international conference on persuasive technology*. 1–6.

[26] Raquel Oliveira, Patrícia Arriaga, Fernando P Santos, Samuel Mascarenhas, and Ana Paiva. 2021. Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior* 114 (2021), 106547.

[27] Ana Paiva, Fernando Santos, and Francisco Santos. 2018. Engineering pro-sociality with autonomous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[28] Elizabeth Pennisi. 2005. How did cooperative behavior evolve? *Science* 309, 5731 (2005), 93–93.

[29] Sofia Petisca, Ana Paiva, and Francisco Esteves. 2020. The effect of a robotic agent on dishonest behavior. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) *(IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 46, 6 pages. https://doi.org/10.1145/3383652.3423953

[30] François Rebaudo, Carlos Carpio, Verónica Crespo-Pérez, Mario Herrera, María Mayer de Scurrah, Raúl Carlos Canto, Ana Gabriela Montañez, Alejandro Bonifacio, Milan Mamani, Raúl Saravia, and Olivier Dangles. 2014. *Agent-Based Models and Integrated Pest Management Diffusion in Small Scale Farmer Communities*. Springer Netherlands, Dordrecht, 367–383. https://doi.org/10.1007/978-94-007-7802-3_15

[31] T Reeves, H Ohtsuki, and S Fukui. 2017. Asymmetric public goods game cooperation through pest control. *Journal of Theoretical Biology* 435 (2017), 238–247.

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[33] James Schaffer, John O'Donovan, Laura Marusich, Michael Yu, Cleotilde Gonzalez, and Tobias Höllerer. 2018. A study of dynamic information display and decision-making in abstract trust games. *International Journal of Human-Computer Studies* 113 (2018), 1–14. https://doi.org/10.1016/j.ijhcs.2018.01.002

[34] Huiying SHI, Jie TANG, and Pingping LIU. 2022. Instability of the watching eyes effect and perceived norms: A new perspective. *Advances in Psychological Science* 30, 12 (2022), 2718.

[35] Hirokazu Shirado and Nicholas A Christakis. 2020. Network engineering using autonomous agents increases cooperation in human groups. *Iscience* 23, 9 (2020), 101438.

[36] Andreas Spahn. 2012. And lead us (not) into persuasion…? Persuasive technology and the ethics of communication. *Science and engineering ethics* 18 (2012), 633–650.

[37] Silvia Tulli, Filipa Correia, Samuel Mascarenhas, Samuel Gomes, Francisco S Melo, and Ana Paiva. 2019. Effects of agents' transparency on teamwork. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer, 22–37.

[38] van den Hoven and Nijssen. 2015. *Handbook of ethics, values, and technological design*. Springer Netherlands.

[39] Carol M. Werner, Paul H. White, Sari Byerly, and Robert Stoll. 2009. Signs that encourage internalized recycling: Clinical validation, weak messages and "creative elaboration". *Journal of Environmental Psychology* 29, 2 (2009), 193–202. https://doi.org/10.1016/j.jenvp.2009.02.003