Walk the Talk! Exploring (Mis)Alignment of Words and Deeds by Robotic Teammates in a Public Goods Game

*Filipa Correia[1], *Shruti Chandra[1,2], Samuel Mascarenhas[1], Julien Charles-Nicolas[1,3], Justin Gally[1,3], Diana Lopes[1], Fernando P. Santos[4], Francisco C. Santos[1], Francisco S. Melo[1] and Ana Paiva[1]

***Abstract*— This paper explores how robotic teammates can enhance and promote cooperation in collaborative settings. It presents a user study in which participants engaged with two fully autonomous robotic partners to play a game together, named "For The Record", a variation of a public goods game. The game is played for a total of five rounds and in each of them, players face a social dilemma: *to cooperate* i.e., contributing towards the team's goal while compromising individual benefits, or *to defect* i.e., favouring individual benefits over the team's goal. Each participant collaborates with two robotic partners that adopt opposite strategies to play the game: one of them is an unconditional cooperator (*the pro-social robot*), and the other is an unconditional defector (*the selfish robot*). In a between-subjects design, we manipulated which of the two robots criticizes behaviours, which consists of condemning participants when they opt to defect, and it represents either an alignment or a misalignment of words and deeds by the robot. Two main findings should be highlighted (1) the misalignment of words and deeds may affect the level of discomfort perceived on a robotic partner; (2) the perception a human has of a robotic partner that criticizes him is not damaged as long as the robot displays an alignment of words and deeds.**

I. INTRODUCTION

Cooperation is vital in human societies, especially in collective endeavors where groups of individuals need to collaborate to ensure common benefits [1]. In a future, where robots are seen as our collaborative partners in mixed teams of humans and robots [2], [3], [4], [5], an immediate question can, therefore, be raised: *how can we create robotic teammates that enhance and promote cooperation?* In humans, the degree of cooperation depends on people's preferences and beliefs [6]. Despite individuals' differences, sustaining high levels of cooperation is, in general, a challenge: People often face situations where their interest is at odds with the group goals. In these situations, people are said to face social dilemmas and cooperation – together with the collective success – is threatened by selfish attitudes [7]. A paradigmatic social dilemma that neatly captures the conflict between the individual and collective interest is the so-called Public Goods Game. In its simplest version, individuals decide how much to contribute to a common pot that is later multiplied by some factor: the result is then equality divided between everyone in the group, irrespective of who has contributed. Clearly, the social optimum is attained when everyone contributes, thereby availing the multiplicative factor. The individual optimum is however to free-ride, avoiding contributing while expecting that others do so.

Within the vision of creating autonomous agents that help humans to cooperate more and to be more pro-social [5], robotic agents can exploit their physicality by employing complex social mechanisms. The social behaviours of robots can not only foster engagement and strengthen the social ties with humans [8], but also shape the social dynamics of a team in a positive manner [9]. Previous studies have shown that the capacity for a robot to display social behaviors such as verbal and non-verbal feedback, gestures, gaze are all able to influence people's behaviors, choices, and decisions during the interaction [10], [11]. For example, verbal feedback given by a social robot can increase human-robot team performance in a collaborative task [12] and can actively repair violations among team members and help in resolving conflicts [13].

A possible mechanism that can increase cooperation in groups is punishment [14], [15]. Typically, punishment is assumed to directly impact participants material gains, which is also a function of the collaborative task at stake. Here, we pursue a "soft" form of punishment by focusing on the social aspects of the interaction that are not directly tied to the task itself. In particular, this paper explores the display of *verbal criticism* by a robotic teammate towards non-cooperator human partners. Furthermore, we are interested in analysing such behaviour in two particular situations: one where the robot acts according to its own criticism and another where the robot hypocritically adopts the strategy it criticises. Hypocritical criticisms is also connected with a the long-standing question in economics and biology of the emergence reliable and honest signaling [16]. Here we aim at exploring if the (mis)alignment of words and deeds has a role on the behavioural responses of humans, i.e., their cooperative and altruistic behaviours, as well as their perceptions of such robotic partners.

To address those questions, we conducted a user study in which participants collaborated with two robotic partners to play a variation of a public goods game, named "For The Record". The robots adopt opposing strategies to play the game and while one of them is unconditionally altruistic and pro-social, the other unconditionally takes the selfish action. We manipulated an alignment and misalignment of words and deeds by having either the pro-social robot or the selfish robot displaying a criticism when the human acts selfishly. The results of our user study reveal important considerations for the design of collaborative interaction on robotic partners.

*These two authors contributed equally.

[1]INESC-ID & Instituto Superior Técnico, Lisbon University, Portugal

[2]École Polytechnique Fédérale de Lausanne, Switzerland

[3]INSA-Lyon, Université de Lyon, France

[4]Princeton University, Department of Ecology and Evolutionary Biology, NJ, USA.

II. RELATED WORK

Although many HRI studies focus on one-to-one human-robot interaction, recently researchers have started taking interest to explore group dynamics in human-robot collaborative teams. In such teams, there are more than two team members and may consist of different combinations of human-robot participants. Some of the studies are based on investigating people's perceptions and preferences of robotic partners. For instance, Chang et al. studied the interaction between groups of people and robots to examine the effect of group size on people's perception towards the robotic partners [17]. They showed people playing in groups behave more competitively towards the robot than individual human players. Similarly, Fraune et al. investigated people's perceptions towards in-group vs. outgroup teams, each consisting of two humans and two robots [18]. The results suggest that people favored the in-group over the outgroup and human over robots. Correia et al. [19] studied the preference of humans in choosing robot partners in the context of a multi-party game finding that people with a higher level of competitiveness tend to prefer a competitive robot and their preferences are also based on the outcome.

Other studies focused on investigating how the robot's behavior positively shapes the social dynamics of teams. For example, Strohkorb et al. investigated the capability of a social robot to shape trust within a team. In their scenario, the team consisted of one robotic teammate with three human teammates, with the robot showing either vulnerable or neutral statements during the interaction. Their results showed that when the robot provided vulnerable statements, the people displayed a higher level of engagement and performed actions such as consoling, explaining, laughing to reduce the amount of tension [20]. In the same line, Jung et al. studied the effect of a robot intervention on affect, perception of conflict, perception of teammates' contributions and overall team performance in a problem-solving task. The authors found that the robot's repair interventions could aid team functioning by regulating conflicts among teammates [13]. In the study [11], the authors investigated the effect of robot's varied gender on its persuasive behavior on people in the context of providing donations. The results indicated that men preferred more the female robot in terms of trust, engagement and providing donations while the women showed little preference. In our study, among several behaviors of a robot, we explored robot's verbal criticism as *punishment* towards the human.

Researchers in the field of social sciences have investigated the act of *(mis)alignment of words and deeds* and, in fact, exploited it to motivate behavioral changes among people. This act of misalignment of words and deeds can also be referred as 'hypocrisy' which is basically an inconsistency in a person's attitude and behavior [21]. Furthermore, it has been said that 'criticism of others in terms of hypocrisy is one of the moral forms censure in the contemporary world' [21]. Stone et al. evidenced that the act of hypocrisy induces motivation to behavioral change and pro-social behaviors in college student to adopt the use of condoms to prevent AIDS [22]. In our research, the motivation behind the use of robots' hypocrisy i.e., misalignment of their words and deeds, is to examine the change in participants' behaviours and their perceptions towards the robotic partner.

Understanding cooperation in the context of social dilemmas, and accordingly devise incentive mechanisms that prevent free-riding, is a fundamental scientific challenge [23], [14]. Before the interaction takes place, commitments can be arranged such that participants pledge to a certain level of contribution [24]. As the interaction occurs, non-verbal interaction can be used to enhance cooperation. In this context, Kurzban performed an experiment exploring the effect of participants social psychophysical cues, while playing public goods game [23]. The participants were asked to use eye gaze, touch, virtual chat and tap rhythms. All the cues were found to increase contributions in male-participants but not in female participants. The fact that embodied signals can be used to trigger cooperation is particularly relevant for our study, where we use robotic players instead of, e.g., virtual characters or alternative communication devices between participants. Another powerful mechanism to sustain cooperation in rooted in reciprocity, as evidenced in several experiments [25]. Reciprocity can be effective as individuals may use their experience in previous encounters, as well as information about current opponents, to decide between cooperating or defecting in particular groups [26]. On top of reciprocity, the particular arrangement of social networks can, by itself, prompt the emergence of cooperation in public goods games [27].

Costly punishment constitutes an additional mechanism that was extensively explored [28]. In this case, individuals pay a cost to reduce the gains of a defector. On the one hand, costly punishment provides a clear and effective disincentive for defection. This mechanism has some caveats, nonetheless. First of all, there is an infamous second-order free-riding problem: If it is required that individuals pay a cost to punish, it may be attractive to free-ride and refrain from punishing – even if free-riders (defectors) cause strong negative emotions among cooperators [14], which may, *per se*, provide a psychological incentive to punish. Additionally, punishment may be inefficient by requiring extra resources to be spent [29]. A softer alternative to punishment is simply information spreading (which may lead to indirect reciprocity [30]). In fact, people reveal themselves sensitive to reputations and shaming, which can be used to motivate contributions in the context of public goods games [31]. In [32], Jacquet et al. found, through behavioral experiments, that inducing shame and honour in a public goods game led to approximately 50 percent higher contributions (compared with a baseline control scenario, without group exposure). People seem, this way, sensitive to being exposed by others as a result of their selfishness. But are they sensitive to criticism from any member of the group? Individuals that are regarded as hypocritical are negatively judged [33]. Will then people react negatively to criticism from defectors, in the context of public goods games – more so if criticism

comes from a hypocritical robotic partner?

III. USER STUDY

Our user study explored the effects of criticisms by robotic partners in collaborative settings. In particular, if the alignment of words and deeds had a role on the behavioural responses of humans as well as their perceptions of such robotic partners. We aimed at investigating the following research questions:

- Do people perceive differently a robot criticiser that displays an alignment of words and deeds compared to a robot that does not behave according to its own criticism?
- Do people cooperate differently with the team when they are criticised by a robot that displays an alignment of words and deeds compared to when they are criticised by a robot that does not behave according to its own criticism?

A. SCENARIO

To address these questions we used the *For The Record*, a collaborative game proposed in [34]. This collective risk dilemma is a variant of public goods games, where the uncertain variables (which are digital dice in our interactive digital interface) allow us to manipulate the outcome of each game. In order to create an engaging and playful activity, the game also contains a musical metaphor in which "the band of three musicians needs to collect as many successful albums as possible without bankrupting". The most relevant aspect of this game is the fact that each round constitutes a social dilemma, in which each player has to choose between *to cooperate* or *to defect*. By cooperating, the player is compromising individual gains in favour of the team, which can be seen as a pro-social or an altruistic decision. By defecting, the player is compromising the team's goal over his individual gains, which can be seen as a selfish or a greedy decision.

The game is composed by several rounds, each corresponding to the publication of an album. Albums can succeed or fail according to its value being above or below the market threshold, respectively, and if the band accumulates three failed albums they immediately lose the game. The value of an album is the sum of contributions of all players, which are individually determined by rolling dice of 6 faces. The number of dice each player can roll is set by his skill level on the instrument. In the end of a round, a successful album provides individual profit to each player, which once again is determined by rolling dice of 6 faces. However in this case, the level each player has on a different skill, the marketing skill, determines the number of dice each player can roll. Finally, the decision each player faces when starting a new round is which skill should they upgrade by 1 point. By upgrading the instrument skill, a player increases the probability of collective success for the band. On the other hand, by upgrading the marketing skill, a player increases the probability of receiving higher individual profit when a band's album succeeds.

This class of Public Goods Games, where agents face a dilemma in which payoffs are uncertain and collective success requires a minimum number of cooperators, has been associated to many collective action problems, notably group hunting and climate change negotiations [35].

B. EXPERIMENTAL DESIGN & HYPOTHESES

In our user study, each participant partnered with two robots to play the game. We used a mixed design with two independent variables: the strategy adopted by each robot as the within-subjects variable, and which one of the two robots displayed the criticism as the between-subjects variable. Therefore, each participant interacted with both robotic partners that are identified as follows, according to their strategy:

- **Pro-social Robot**: This robot unconditionally chooses to cooperate in every round;
- **Selfish Robot**: This robot unconditionally chooses to defect in every round;

Regarding the expression of criticism, each participant was randomly assigned to one of the two experimental conditions:

- **Pro-social Critic (PC)**: The pro-social robot expresses a criticism whenever the participant chooses to defect. In this condition there is an alignment of words and deeds as the pro-social robot never chooses to defect, which is the action it is condemning;
- **Selfish Critic (SC)**: The selfish robot expresses a criticism whenever the participant chooses to defect. In this condition there is a misalignment of words and deeds as the selfish robot always chooses to defect, which is the action it is condemning.

The utterances to express the verbal criticism were the same in both conditions: "Really? Are you going to play like that?", "If you play like this, our team will never win.", and "You could help more our team...".

The following hypotheses identified our expectations regarding the previously mentioned research questions:

H1: People will perceive more negatively a selfish robot when it criticises, compared to a pro-social robot when it makes the same criticism;

H2: People will cooperate more when criticised by a pro-social robot than when criticised by a selfish robot.

The motivation behind H1 comes from the social sciences, where the (mis)alignment of words and deeds, also refereed as hypocrisy, revealed negative effects on peoples' perceptions [21], [33]. On the other hand, the rationale behind our H2 comes from the game theory field, where cooperating punishers can increase cooperation levels [36].

C. MATERIALS & APPARATUS

In the study, we used a laptop, a touchscreen, a video camera, and two EMYS robots [37], as shown in Fig. 1. The two autonomous robots were developed using the SERA ecosystem [38] and their decision making acts according to emotional mechanisms provided by the FAtiMA toolkit [39].

Fig. 1. Interaction with the robots during the user study.

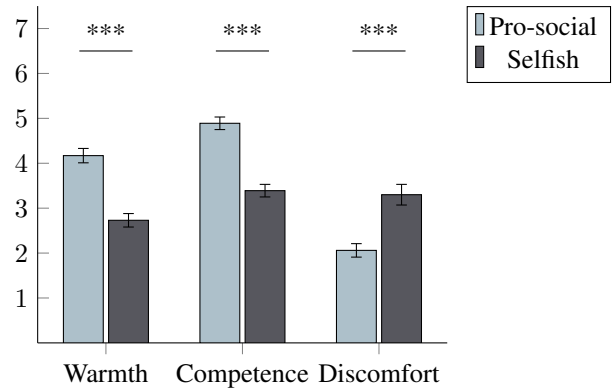

Fig. 2. Main effect of the strategy on the social attributes of warmth, competence and discomfort.

D. PROCEDURE

During the briefing of the study, participants were informed of the procedure and they were asked permission to video-record the experiment. Both subjects that did and did not allow for being recorded signed the consent form accordingly and could participate in either case because the main data source was the final questionnaire.

1) Training (10-15 minutes): The researcher would play a training game with the participants (without the robots) to make them familiar regarding all the rules, features and other game information. The results of the rounds were scripted for the participant to win the training game according to the following sequence, from the 1^{st} to the 5^{th} round, <losing, winning, losing, wining, winning>.

2) Interaction with the Robots (5 minutes): The researcher would leave the room and the participant would play another game with the two robots. Before leaving, the researcher would emphasise that the participant should pay attention to both robots concerning their names and behavior as they would be asked a few questions about them later. According to our previous findings, using the same scenario, negative results exacerbate people's perceptions [34]. Thus, the results of the rounds were manipulated for the participant to lose the main game according to the following sequence, from the 1^{st} to the 5^{th} round, <winning, losing, wining, losing, losing>.

3) Questionnaire (10 minutes): The researcher would ask the participant to fill the questionnaire. In the end, the researcher would thank the participant for his/her participation.

E. MEASURES

To assess H1, regarding the perceptions that participants had towards each robot, we used RoSAS [40] with its three dimensions of warmth, competence and discomfort, which presented good reliability scores. Regarding H2, we used the objective number of times (out of 4) that each participant chose to cooperate.

F. SAMPLE

We collected a sample of 50 participants in the campus of a major technological institute. However, subjects that did not defect during the game were not exposed to the experimental manipulation and, therefore, were removed from the data analysis. Consequently, the resultant sample had 46 participants (22 in the PC condition and 24 in the SC). There were 20 females and their age ranged between 18 and 49 ($M_{age} = 24.04, SD = 5.62$). The participation in this study was voluntary and there was no material incentive.

IV. RESULTS

A. PERCEPTION OF THE ROBOTS

We analysed the impact of the criticism on the participant's perception of each robot by using a Mixed Analysis of Variance (ANOVA). The within-subjects factor is the strategy adopted by each robot (i.e., pro-social and selfish), which is present in both experimental conditions as participants form a team with the two robotic partners. The between-subjects factor is the experimental condition in which one of the robots expressed the criticism (i.e., PC or SC).

1) Warmth: We found a statistically significant main effect of the strategy that each robot adopted on the perception of warmth (Fig. 2, $F(1, 44) = 45.67, p < 0.001$). The pro-social robot was rated with higher levels of warmth ($M = 4.17, SD = 1.11$) compared to the selfish robot ($M = 2.73, SD = 1.04$). However, we did not find a statistically significant main effect of which robot expressed the criticism ($F(1, 44) = 0.495, p = 0.485$). In addition, we did not find significant interaction effect between the two independent variables ($F(1, 44) = 3.50, p = 0.068$).

In other words, the attribution of warmth to each robot seems to have been affected by the strategy it adopted (pro-social or selfish), regardless of being or not the critic.

2) Competence: Regarding the perception of competence, we found a statistically significant main effect of the strategy adopted by each robot (Fig. 2, $F(1, 44) = 53.33, p < 0.001$). Participants rated the pro-social robot as more competent ($M = 4.89, SD = 0.96$) compared to the selfish robot ($M = 3.39, SD = 0.96$). Nevertheless, we did not find a statistically significant difference between the two conditions ($F(1, 44) = 0.06, p = 0.806$). Similarly, there was no significant interaction effect between the two independent variables ($F(1, 44) = 0.194, p = 0.661$).

We can say that the competence attributed to each robot seems to have been affected by its game strategy (pro-social or selfish), regardless of which one has shown the criticism.

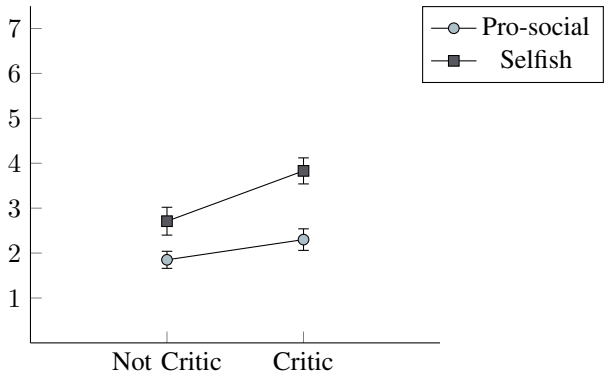

Fig. 3. Interaction effect between the strategy of the robots to play the game and the portrayal of criticism on the attribution of discomfort ($p = 0.001$).

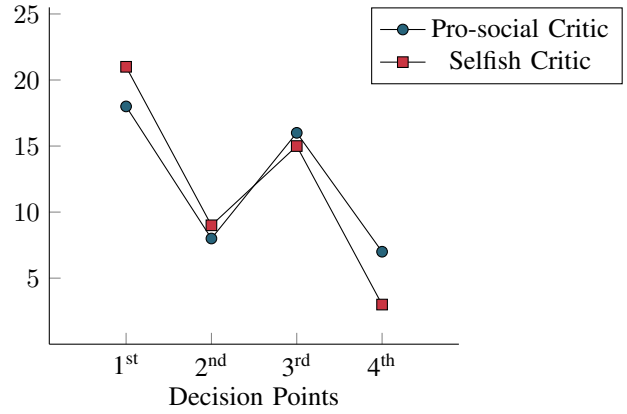

Fig. 4. Number of participants that cooperated in each round per experimental condition.

3) Discomfort: Regarding the social attribute of discomfort, we found a statistically significant main effect of the strategy used in the game (Fig. 2, $F(1, 44) = 27.66, p < 0.001$. The robot that adopted a selfish strategy was rated with higher levels of discomfort ($M = 3.30, SD = 1.53$) compared to the robot that adopted a pro-social strategy ($M = 2.06, SD = 1.05$). Although there was no statistically significant main effect between the conditions (which robot expressed the criticism) ($F(1, 44) = 1.31, p = 0.259$), we found a significant interaction effect between the two independent variables (Fig.3, $F(1, 44) = 11.99, p = 0.001$).

These results show that the attribution of discomfort seems to have been affected by the robot's strategy, pro-social or selfish. Further, the interaction effect also points to different variations according to which robot expressed the criticism.

To understand this interaction, we compared the perception of discomfort attributed to each robot between the two conditions. The discomfort attributed to the pro-social robot was not significantly different when it was the critic compared to when it was not ($U = 195.0, p = 0.127$). However, the discomfort attributed to selfish robot was significantly different when it was the critic ($M = 3.83, SD = 1.42$) compared to when it was not ($U = 151.5, p = 0.013; M = 2.71, SD = 1.46$).

The results show the portrayal of criticisms increased the discomfort attributed to the selfish robot, but it did not affect the discomfort attributed to the pro-social robot.

B. PRO-SOCIAL BEHAVIOUR

We compared the total amount of times participants adopted the pro-social strategy between conditions, i.e., participants that were criticised by the pro-social robot compared to participants that were criticised by the selfish robot. This difference was not statistically significant ($U = 221.5, p = 0.318$).

In order to understand what might have influenced participants to choose between cooperating and defecting, we analysed carefully the strategies of participants. Considering their 4 decision points, there were 11 participants that cooperated only once, 19 that cooperated twice, and 16 that cooperated 3 times. However, when looking at decisions in each round (see Fig.4), there seems to be a salient trend of sequentially choosing *cooperate, defect, cooperate, defect*. Interestingly, the scripted sequence for the result of the rounds in the experiment also contains an alternating pattern of <winning, losing, wining, losing, losing>. As the first decision point occurs in the beginning of the second round (and the result of the last round occurs after the last decision point), the matching pattern suggests people were usually more prone to cooperate after losing in the previous round, and were usually more prone to defect after winning the previous round.

Additionally, we analysed the first decision participants took immediately after being criticised by the robotic partner, i.e., after they have defected for the first time. In this analysis, we excluded 8 participants that defected for the first time in the last round. In the PC condition, there were 13 participants (out of 18) that cooperated, while in the SC condition, only 9 (out of 20) chose to cooperate. Although there was no significant association between this decision and the condition ($\chi(1) = 2.88, p = 0.09$), the trend suggests an effect of the manipulation in the direction of our hypothesis.

Finally, we run a correlation analysis between the number of times people cooperated and the perceptions they had of each robot in terms of their social attributes. We found a significant weak negative correlation with the discomfort attributed to the selfish robot ($r = -0.294, p = 0.047, N = 46$). In other words, as the discomfort attributed to the selfish robot increased, the number of times people cooperated decreased. However, no similar relation was found with the discomfort perceived on the pro-social robot as the the correlation was non-significant ($r = -0.173, p = 0.249, N = 46$).

We have also found a significant weak positive correlation with the competence attributed to pro-social robot ($r = 0.295, p = 0.047, N = 46$). As the level of competence attributed to the pro-social robot increases, the number of times people cooperated also increased.

V. DISCUSSION

In **H1**, we have hypothesised that people would perceive more negatively a robot that gives criticism while displaying a misalignment of its words and deeds, compared to a robot

that also criticises but displays an alignment of its words and deeds. Among the three dimensions of the RoSAS, which was the scale used to assess participants' perceptions towards the robots, only the social attribute of discomfort supported our hypothesis. We found an interaction effect between the strategy of the robot and whether or not it gave the criticism. Specifically, the discomfort attributed to the selfish robot significantly increased when it was the critic, while the discomfort attributed to the pro-social robot was not significantly different between conditions. This result suggests that indeed a misalignment of words and deeds negatively affects the discomfort attributed to the robot.

Nonetheless, **H1** entails an additional expectation that the alignment of words and deeds would legitimate a pro-social robot to express criticisms. This part of our hypothesis was fully supported by the fact that the perceptions of the three social attributes towards the pro-social robot were not significantly different between having or not expressed criticisms. This finding constitutes an important contribution of our experimental study by holding this idea that a robotic partner, that displays an alignment of words and deeds, can condemn people's selfish actions without compromising its perception. Further investigation is needed to ascertain if these results would hold when the other robotic partner adopts different strategies.

We consider our hypothesis was only partially validated as the social attributes of warmth and competence were not negatively affected on the robotic partner that was not behaving according to its criticism. The main effects of the strategy on the attributions of warmth and competence suggest these traits are strongly influenced by degree of cooperation a robotic partner adopts in a collaborative setting, regardless of whether or not it displays criticisms.

Additionally, we would like to emphasise that participants rated the selfish robot with significantly lower warmth, lower competence and higher discomfort compared to the pro-social robot. Generally, these results suggest people have negatively perceived a robotic partner that compromises the team's goal in favor of its individual gains, which is consistent with previous findings [14], [34].

In **H2**, we have hypothesised that people would adopt a more cooperative strategy when they are criticised by a robot that displays an alignment of words and deeds compared to when they are criticised by a robot that does not behave according to its own criticism. The number of times participants cooperated was not significantly different between the two experimental conditions. Although our current results cannot support our hypothesis, we found two slight trends that suggest the manipulation might still have affected participants' decisions (i.e., the average cooperation rate per condition and the decision immediately after being criticised). Another important consideration is the matching patterns found between the result of the round and the trend on the following decision of participants, which suggests the result of the previous round has also affected their decisions and constitutes a limitation of the scenario. Overall, this result also suggests that informal and costless criticisms — in the sense that it does not directly influence the payoff of both the offender and the punisher — imposed by "honest" artificial entities, may influence human decisions by potentially highlighting a social norm that is not followed.

Finally, we would like to mention two additional aspects that might have affected the perception of the criticisms and its consequent effect on the cooperation rate. The first one is the fact that when the pro-social robot was the criticiser, it only criticised the actions of the participant without condemning the other robot that was an actual free-rider. Secondly, participants may have thought that the presence of a free-rider on the team (the selfish robot) was not worth it for them to change their strategy.

VI. CONCLUSIONS

As the capabilities of robots evolve, they will be required to collaborate with humans in several distinct domains (e.g., education or health). Therefore, it is crucial to explore new ways of embracing these collaborations. Moreover, the affordances of physical robots open a wider variety of social mechanisms that can be adopted to foster collaboration and cooperation. Our paper sheds some light on the expression of criticisms by social robots in collaborative interactions as a verbal mechanism to promote cooperation. In particular, it explores the (mis)alignment of words and deeds by comparing a condition where the robot expresses a criticism and behaves accordingly with another condition where the criticism is expressed by a robot that performs the condemned action.

Overall, our results provide new and insightful findings for the HRI field. Although we did not find evidence that the (mis)alignment of words and deeds influenced the cooperative behaviour of people, important considerations can be drawn from its effects on the perception of robotic partners. When analysing our results in a more general perspective, two major points should be highlighted: (1) the misalignment of words and deeds may affect the level of discomfort perceived on a robotic partner; (2) the perception a human has of a robotic partner that criticises him is not damaged as long as the robot displays an alignment of words and deeds.

As future work, we would like to extend the current study with the two robots displaying congruent strategies. Another interesting avenue is the exploration of the robot's physicality, either by replicating the experiment with different embodiments or by manipulating the physical presence of these autonomous agents. Moreover, increasing the number of rounds in the game can also elicit participants to defect later in the game, opening the possibility of having more complex community enforcing mechanisms, such as costly sanctions and reputations, all with exciting implications if hybrid populations, comprising humans and machines, are considered.

ACKNOWLEDGMENTS

Supported by FCT Portugal with the references FCT-UID/CEC/50021/2019, PTDC/MAT/STA/3358/2014, PTDC/EEI-SII/5081/2014, and PTDC/EEISII/7174/2014.

Filipa Correia acknowledges her FCT grant the with reference SFRH/BD/118031/2016. Fernando P. Santos acknowledges support from the James S. McDonnell Foundation.

REFERENCES

[1] M. Nowak and R. Highfield, *Supercooperators: Altruism, evolution, and why we need each other to succeed*. Simon and Schuster, 2011.

[2] H. A. Yanco and J. Drury, "Classifying human-robot interaction: an updated taxonomy," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, vol. 3. IEEE, 2004, pp. 2841–2846.

[3] G. Hoffman and C. Breazeal, "Collaboration in human-robot teams," in *AIAA 1st Intelligent Systems Technical Conference*, 2004, p. 6434.

[4] V. Groom and C. Nass, "Can robots be teammates?: Benchmarks in human–robot teams," *Interaction Studies*, vol. 8, no. 3, pp. 483–500, 2007.

[5] A. Paiva, F. P. Santos, and F. C. Santos, "Engineering pro-sociality with autonomous agents," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] U. Fischbacher and S. Gachter, "Social preferences, beliefs, and the dynamics of free riding in public goods experiments," *American economic review*, vol. 100, no. 1, pp. 541–56, 2010.

[7] H.-Y. Liang, H.-A. Shih, and Y.-H. Chiang, "Team diversity and team helping behavior: The mediating roles of team cooperation and team cohesion," *European Management Journal*, vol. 33, no. 1, pp. 48–59, 2015.

[8] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[9] H. Tennent, S. Shen, and M. Jung, "Micbot: A peripheral robotic object to shape conversational dynamics and team performance," in *14th Int. Conf. on Human-Robot Interaction*. IEEE, 2019, pp. 133–142.

[10] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," in *Proc. 1999 International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients*, vol. 2. IEEE, 1999, pp. 858–863.

[11] M. Siegel, C. Breazeal, and M. I. Norton, "Persuasive robotics: The influence of robot gender on human behavior," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 2563–2568.

[12] A. St Clair and M. Mataric, "How robot verbal feedback can improve team performance in human-robot task collaborations," in *10th Int. Conf. on Human-Robot Interaction*. ACM, 2015, pp. 213–220.

[13] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: the case of repairing violations," in *10th Int. Conf. on Human-Robot Interaction*. ACM, 2015, pp. 229–236.

[14] E. Fehr and S. Gachter, "Cooperation and punishment in public goods experiments," *American Economic Review*, vol. 90, no. 4, pp. 980–994, 2000.

[15] S. Gächter, E. Renner, and M. Sefton, "The long-run benefits of punishment," *Science*, vol. 322, no. 5907, pp. 1510–1510, 2008.

[16] B. Skyrms, *Signals: Evolution, learning, and information*. Oxford University Press, 2010.

[17] W.-L. Chang, J. P. White, J. Park, A. Holm, and S. Šabanović, "The effect of group size on people's attitudes and cooperative behaviors toward robots in interactive gameplay," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 845–850.

[18] M. R. Fraune, S. Šabanović, and E. R. Smith, "Teammates first: Favoring ingroup robots over outgroup humans," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 1432–1437.

[19] F. Correia, S. Petisca, P. Alves-Oliveira, T. Ribeiro, F. S. Melo, and A. Paiva, "Groups of humans and robots: Understanding membership preferences and team formation." in *Robotics: Science and Systems*, 2017.

[20] S. Strohkorb Sebo, M. Traeger, M. Jung, and B. Scassellati, "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams," in *13th Int. Conf. on Human-Robot Interaction*. ACM, 2018, pp. 178–186.

[21] R. J. Wallace, "Hypocrisy, moral address, and the equal standing of persons," *Philosophy & Public Affairs*, vol. 38, no. 4, pp. 307–341, 2010.

[22] J. Stone and N. C. Fernandez, "To practice what we preach: The use of hypocrisy and cognitive dissonance to motivate behavior change," *Social and Personality Psychology Compass*, vol. 2, no. 2, pp. 1024–1051, 2008.

[23] R. Kurzban, "The social psychophysics of cooperation: Nonverbal communication in a public goods game," *Journal of Nonverbal Behavior*, vol. 25, no. 4, pp. 241–259, 2001.

[24] L. M. Pereira, T. Lenaerts, *et al.*, "Evolution of commitment and level of participation in public goods games," *Autonomous Agents and Multi-Agent Systems*, vol. 31, no. 3, pp. 561–583, 2017.

[25] H. Gintis, S. Bowles, R. Boyd, and E. Fehr, "Explaining altruistic behavior in humans," *Evolution and human Behavior*, vol. 24, no. 3, pp. 153–172, 2003.

[26] F. P. Santos, S. Mascarenhas, F. Santos, F. Correia, S. Gomes, and A. Paiva, "Outcome-based partner selection in collective risk dilemmas," in *in Proc. of AAMAS 2019*, 2019.

[27] F. C. Santos, M. D. Santos, and J. M. Pacheco, "Social diversity promotes the emergence of cooperation in public goods games," *Nature*, vol. 454, no. 7201, p. 213, 2008.

[28] E. Fehr and S. Gächter, "Altruistic punishment in humans," *Nature*, vol. 415, no. 6868, p. 137, 2002.

[29] A. Dreber, D. G. Rand, D. Fudenberg, and M. A. Nowak, "Winners don't punish," *Nature*, vol. 452, no. 7185, p. 348, 2008.

[30] F. P. Santos, J. M. Pacheco, and F. C. Santos, "Evolution of cooperation under indirect reciprocity and arbitrary exploration rates," *Scientific reports*, vol. 6, p. 37517, 2016.

[31] E. Yoeli, M. Hoffman, D. G. Rand, and M. A. Nowak, "Powering up with indirect reciprocity in a large-scale field experiment," *Proc. of the National Academy of Sciences*, vol. 110, no. Supplement 2, pp. 10 424–10 429, 2013.

[32] J. Jacquet, C. Hauert, A. Traulsen, and M. Milinski, "Shame and honour drive cooperation," *Biology Letters*, vol. 7, no. 6, pp. 899–901, 2011.

[33] J. J. Jordan, R. Sommers, P. Bloom, and D. G. Rand, "Why do we hate hypocrites? evidence for a theory of false signaling," *Psychological science*, vol. 28, no. 3, pp. 356–368, 2017.

[34] F. Correia, S. F. Mascarenhas, S. Gomes, P. Arriaga, I. Leite, R. Prada, F. S. Melo, and A. Paiva, "Exploring prosociality in human-robot teams," in *14th Int. Conf. on Human-Robot Interaction*. IEEE, 2019, pp. 143–151.

[35] F. C. Santos and J. M. Pacheco, "Risk of collective failure provides an escape from the tragedy of the commons," *Proc. Natl Acad. Sci. USA*, vol. 108, no. 26, pp. 10 421–10 425, 2011.

[36] D. Helbing, A. Szolnoki, M. Perc, and G. Szabó, "Punish, but not too hard: how costly punishment spreads in the spatial public goods game," *New Journal of Physics*, vol. 12, no. 8, p. 083005, 2010.

[37] J. Kędzierski, R. Muszyński, C. Zoll, A. Oleksy, and M. Frontkiewicz, "Emys—emotive head of a social robot," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 237–249, 2013.

[38] T. Ribeiro, A. Pereira, E. Di Tullio, and A. Paiva, "The sera ecosystem: Socially expressive robotics architecture for autonomous human-robot interaction," in *2016 AAAI Spring Symposium Series*, 2016.

[39] J. Dias, S. Mascarenhas, and A. Paiva, "Fatima modular: Towards an agent architecture with a generic appraisal framework," in *Emotion modeling*. Springer, 2014, pp. 44–56.

[40] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas): Development and validation," in *12th Int. Conf. on Human-Robot Interaction*. ACM, 2017, pp. 254–262.