Journal of Behavioral and Experimental Economics 104 (2023) 102011

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

If it looks like a human and speaks like a human … Communication and cooperation in strategic Human–Robot interactions

Mario A. Maggioni *, Domenico Rossignoli

CSCC, DISEIS and HuRoLab, Università Cattolica del Sacro Cuore, Largo Gemelli,1, Milano, 20123, Italy

ARTICLE INFO

Keywords:
Prisoner's Dilemma
Communication
Human–Robot interaction
Behavioral experiment

ABSTRACT

This paper presents the results of a behavioral experiment conducted between February 2020 and March 2021 on a sample of about 500 university students that were randomly matched with either a human or a humanoid robot partner to play an iterated Prisoner's Dilemma, to test whether their choices were influenced by the nature and behavior of their partner. The results show that subjects are more likely to cooperate with human rather than with robotic partners; that they are more likely to cooperate after receiving a verbal reaction following a sub-optimal social outcome; and that the effect of the verbal reaction is not dependent on the nature of the partner. Our findings provide new evidence on the effects of verbal communication in strategic frameworks that involves humanoid robotic partners. The results are robust to: the exclusion of students of Economics-related subjects, the inclusion of a set of psychological and behavioral controls, the subjects' perception on robots' behavior, and gender biases in human–human interactions.

1. Introduction

In recent years, the use of robots has expanded from research labs and manufacturing plants to the service sector offices and, most importantly, facilities – such as hospitals, care centers, schools, and even private homes – where their main task is to interact with "fragile" human beings (elderly, children, sick), paving the way for these technologies to play an increasingly important role in daily life interactions with humans. People have consequently raised their expectations regarding robots: More than one-third of respondents in a recent survey by Ipsos say that robots "will look like, think like and speak like humans" in the near future.[1] It has therefore become increasingly crucial that robots act in a reliable and transparent way to be perceived as trustworthy by their human partners in terms of their actions, capabilities and motivations (Felzmann, Villaronga, Lutz, & Tamò-Larrieux, 2019; Zörner et al., 2021).

Although the use of robots in experimental economics can be traced back at least to Walker, Cox, and Smith (1987) and Andreoni and Miller (1993), the economic literature on Human–Robot Interactions (HRI) is still limited, while it is expected to become more significant as more and more people will begin to interact with artificial intelligence in their daily lives.

To analyze and improve the understanding of Human–Robot Interactions (henceforth: HRI), we devised a randomized experiment in which human subjects were randomly matched to either a human or an anthropomorphic robot partner (NAO, produced by Softbank Robotics) and asked to perform an iterated Prisoner's Dilemma (PD). In each of the two experimental conditions, after the first round of the game (played online), and prior to a second round being proposed (in the lab), half of the subjects were randomly assigned to a treatment in which the partner provides a *Verbal Reaction* (henceforth, VR) if a sub-optimal social outcome had occurred in the first round of the PD.

The aim of our experiment is threefold: first, to investigate whether the subjects' behavior depends on the (human or robotic) nature of their partner; second, to analyze whether a VR – which implicitly refers to cooperation as a socially desirable strategy – influences the subjects' subsequent choices; third, and most importantly, to check whether the effect of the VR depends on the (human or robotic) nature of the partner.[2]

Our main result shows that facing a robot partner, when no VR is performed, decreases the subjects' cooperation rate between 16 and 22

* Correspondence to: Università Cattolica del Sacro Cuore, Largo Gemelli, 1 20123 Milano, Italy.
E-mail address: mario.maggioni@unicatt.it (M.A. Maggioni).

[1] Based on 15,700 online interviews in the period October 23rd – November 6th 2020, and across 31 countries, see https://www.ipsos.com/en/global-predictions-2021.

[2] In a sense, our third research question can be thought of as a sort of modified Turing Test. In the original three-person "imitation game" (Turing, 1950) an 'interrogator' chats with two respondents, located in separate rooms, asking questions to detect which of the two is a machine. If the 'interrogator' cannot reliably distinguish machines from human beings, the machine is said to have passed the test.

https://doi.org/10.1016/j.socec.2023.102011
Received 11 August 2022; Received in revised form 22 February 2023; Accepted 15 March 2023
Available online 24 March 2023

percentage points, depending on model specifications, while being exposed to a VR increases the subjects' cooperation rate in the next round by 20 to 25 percentage points, depending on model specifications. Interestingly, the difference in the effect for subjects facing human vs robot partners is not statistically significant. In other words, despite the fact that the VRs do not affect the players' payoffs, as long as the partner subtly evokes cooperation – either in the form of an apology, a reprimand or disappointment – the subject's subsequent choice is nudged towards cooperation, irrespective of whether their partner is a fellow human being or an anthropomorphic robot.

This paper contributes to the existing literature by providing new evidence on the effectiveness of communication in affecting decision-making and extending this result to HRI. In the paper, we show that the effect of a VR on cooperation is positive, significant, and independent of the human vs. artificial nature of the agent, thus contributing to an area in the field of economic interactions between human and non-human subjects that, while still relatively under-researched in the literature, is likely to become more relevant in the near future.

2. Related literature

In the economic literature, the decision to trust another agent to behave as a trustworthy partner in a transaction, or to cooperate in a social dilemma[3] is generally seen as inconsistent with the pursuit of individual self-interest. For this reason, trustful decisions are usually explained either in terms of repeated interactions within a finite uncertain time horizon, or assuming: agents endowed with "self-regarding preferences", an appropriate time discount rate, a certain degree of uncertainty over the type of opponent, a "warm glow effect" (Andreoni, 1990), or "gift-giving" behavior (Akerlof, 1982). In the case of agents endowed with "other-regarding preferences", they tend to be explained by referring to concepts such as equity and fairness (Fehr & Schmidt, 1999; Rabin, 1993).

Relatively few papers explicitly consider the real relational dimension of agents' interactions in social dilemmas. In those papers, cooperative and trustful behaviors are explained as being motivated by an acknowledgment of the other party's attitudes and intentions. Similarly, the literature on so-called psychological games addresses the role of subjects' intentions, by making payoffs belief-dependent (DeAngelo & McCannon, 2020; Dufwenberg & Kirchsteiger, 2004; Geanakoplos, Pearce, & Stacchetti, 1989; Rabin, 1993).

Communication, and even cheap talk (Farrell, 1995; Farrell & Rabin, 1996), have been shown to be crucial in affecting the outcome of social dilemmas. In particular, face-to-face communication has been shown to both promote and sustain cooperation between subjects even in strategic settings such as social dilemmas (Bicchieri, 2002; Braver & Wilson, 1986; Ostrom, 2000; Ostrom & Walker, 1997). Charness and Dufwenberg (2006) and Miettinen and Suetens (2008) showed that communication may ignite guilt aversion and therefore increase cooperation rates in social dilemmas. Similar evidence has been provided by Ben-Ner, Putterman, and Ren (2011), even when communication is not binding, and by Charness, Feri, Meléndez-Jiménez, and Sutter (2021), which explores the role of communication in network games.

Overall, the empirical literature has shown that verbal communication increases trust and cooperation which in turn are pivotal in fostering positive social interactions. Sally (1995) published the first meta-analysis of this stream of the literature and concluded that communication exerts the strongest effect, relative to other variables known to influence cooperation, such as group size, the magnitude of the reward for choosing not to cooperate, and group identity. Subsequently, Balliet (2010) addressed the same issue through an improved meta-analysis, adopting mediation-analysis techniques, and confirmed that communication has a strong positive effect on cooperation within social dilemmas.

Finally, support for the importance of attitudes, intentions and verbal and non-verbal cues in communication emerges from behavioral economics, as well as psychology and neuroeconomics experiments, where players show different behaviors and neurological activation when playing incentivized tasks and games with human counterparts as opposed to *automata* - ranging from PCs, to robots with varying degrees of humanization, to different forms of A.I. - despite facing identical material payoffs (Andreoni & Miller, 1993; Bicchieri & Lev-On, 2007; Chugunova & Sele, 2020; Cominelli et al., 2021; Crandall et al., 2018; Ishowo-Oloko et al., 2019; Kiesler, Sproull, & Waters, 1996; Klockmann, von Schenk, Villeval, et al., 2021; Krach et al., 2008; McCabe, Houser, Ryan, Smith, & Trouard, 2001; de Melo, Carnevale, & Gratch, 2011; Miwa & Terai, 2012; Nouri & Traum, 2013; Paeng, Wu, & Boerkoel, 2016; Pelikan & Broth, 2016; Rilling et al., 2002; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; Tahir, Dauwels, Thalmann, & Thalmann, 2018; Terada & Takeuchi, 2017; Wu, Paeng, Linder, Valdesolo, & Boerkoel, 2016; Zörner et al., 2021). In a pioneering work matching human with *automata* Andreoni and Miller (1993) assigned subjects to two specific computer partners playing an imitative tit-for-tat strategy with different probabilities. Their findings indicate that in the early rounds of a repeated Prisoner's Dilemma, subjects who were paired with the algorithm programmed to adopt the imitative strategy with high probability exhibited a higher cooperation rate.

More recently Cominelli et al. (2021) show that subjects receiving a promise from a humanoid robot have greater trust in their partner, especially when they perceive the robot as very similar to a human being, while this does not happen when the partner is a computer box.

Academic research has accompanied the technological evolution in robotics, although by developing two sub-fields: Social Robotics (SR) and Human–Robot Interactions (HRI). The first (see, among others, Fong, Nourbakhsh, & Dautenhahn, 2003; Sung, Grinter, & Christensen, 2010) is devoted to the design and development of mechanical objects that are able to communicate, both verbally and non-verbally, and to act as "artificial subjects" and "social partners" towards human beings; the second (see, among others, Dumouchel & Damiano, 2017; Gaggioli et al., 2021) is devoted to the analysis of the interactions between these "artificial subjects" and human beings.

Notwithstanding the pioneering work by Walker et al. (1987) who introduced computer algorithms to imitate three different risk-averse bidders, the use of artificial agents as a substitute for human participants in strategic interactions has not been widely adopted in the experimental and behavioral economics literature.

In a series of studies, Hsieh, Chaudhury, and Cross (2020), Hsieh and Cross (2022) investigated human–robot cooperation in the context of Prisoner's Dilemma games, focusing on the impact of incentive structure and emotional displays by robots on people's willingness to cooperate with a robot opponent. In these studies, subjects were facing a non-humanoid social robot (Cozmo), exhibiting a strong reciprocal tendency, which surpasses the influence of the reward value of their decisions (Hsieh et al., 2020). Further, in Hsieh and Cross (2022) the same personal factors predicting cooperation in human–human settings were also important in shaping human–robot interactions.

Although social robots like Cozmo were important at earlier stages of HRI studies, more recently, empirical research in HRI has implemented experimental frameworks in which human subjects have been partnered with humanoid robots in social dilemmas (see, among others, DeSteno et al., 2012; Krach et al., 2008; de Melo et al., 2011; de Melo & Terada, 2020; Paeng et al., 2016; Zörner et al., 2021). This process has almost seamlessly led to the question of whether communication also promotes cooperation in these new settings where people interact with robotic agents.

Among studies involving humanoid robots, the use of NAO has become increasingly popular due to the robot's features and capabilities, which make it appropriate for experimental (especially clinical)

[3] Commonly, operationalized within a game-theoretical framework through a Centipede Game, a "lost letter" experiment, a Trust Game, or a Prisoner's Dilemma. See, among others, Dasgupta (1988), Kreps (1990) Yezer, Goldfarb, and Poppen (1996) and Skeath (1999).

research. Robaczewski, Bouchard, Bouchard, and Gaboury (2020) document 70 experimental studies in which subjects are involved in a HRI with NAO. Of these, 26 studies are specifically designed to assess social interactions between humans and robots, and only 5 of them specifically focus on communication as the main moderating feature. In particular Sandoval, Brandstetter, Obaid, and Bartneck (2016) ask subjects to play a Prisoner's Dilemma – and an Ultimatum Game – but their paper does not focus on the possible effects of partners' verbal messages on subjects' behavior.[4]

Our paper explicitly deals with the interaction between humanoid robots and humans,[5] focusing on the possible differential effect of the partner's communication in determining the behavior of human subjects within a series of strategic interactions.

3. Research design

Our research design addresses three specific questions that relate to the way humans interact with humanoid robots in a repeated Prisoner's Dilemma.

RQ1. **Do subjects' cooperation rates differ according to partner types?** Following recent developments in empirical research in economics, psychology, and social robotics, we aim to understand whether subjects interacting with a robot partner – as opposed to a human – display a different cooperation rate.

RQ2. **Does a VR affect the subjects' cooperation rate?** Empirical evidence in the literature on social dilemmas shows that communication tends to promote cooperation. We aim to investigate whether subjects are more likely to cooperate once the partner has activated a VR, after observing a sub-optimal outcome in the previous round of the game.

RQ3. **Are subjects' reactions to a VR dependent on the partner's nature (Human vs Robot)?** This question originates directly from the former two. Since our experiment is designed as a two × two matrix (see Table 1 below) we might expect subjects' decisions to be affected by either the nature of the partner or exposure to a VR (or both). In other words, we want to explore whether the potential effect of VR is characterized by heterogeneity in the nature of the partner.

To address our RQs, we devised a two×two experimental design, as summarized in Table 1: first, we randomly assigned subjects to either a Human or Robot partner in the interactive situation (the Prisoner's Dilemma); second, we randomly administered a stimulus (treatment) to a fraction of our subjects in both the Human and Robot experimental conditions. The treatment consisted of a "Verbal Reaction" (VR) that the partner delivered after observing a sub-optimal outcome of the interaction.[6] Different stimuli were administered depending on the observed outcomes in the Prisoner's Dilemma,[7] as summarized in Table 3.

Table 1
Experimental design.

		Treatment group	
		No VR	VR
Experimental condition	**Human**	Baseline	Reaction
	Robot	Robot	Interaction

To investigate how subjects behave when faced with a robotic rather than a human partner, we could not rely on an entirely anonymized series of interactions, since the purpose of the investigation was focused on the possible effects that may arise from interactions with partners of different types. At the same time, we also needed to exclude that a specific type of player (with either a pro-robot or a pro-human cooperation bias) was systematically selected into a specific treatment condition.

For this reason, we designed an experimental procedure with two distinct phases, one online (Phase 1) and one in-person at our lab (Phase 2). In Phase 1, we administered to the subjects (university students) an online questionnaire where they were asked to play an incentivized task against an unknown (human or artificial) anonymous partner and were told that previous research has shown that the cooperation rate of Italian people is about 50%. This was the only information they were given on their partner's behavior. Before making their choice, the subjects were administered a test to assess their comprehension of the task.[8]

At the end of the online questionnaire, subjects were asked whether they wanted to come to the University Lab, and proceed to Phase 2. In Phase 2, as summarized in Fig. 1, they: (i) met the partner; (ii) discovered the Human or Robot type of the partner; (iii) learned the result of the interaction; (iv) were randomly assigned to the treatment group (VR or No VR); (v) chose – if asked – to have other interactions with the partner; (vi) played another round of the game if that was the case; (vii) were rewarded; (viii) chose whether to donate 1 euro to an NGO of their choice; (ix) signed a confidentiality agreement and exited the lab.

To ensure that the effect of communication in the Robot and Human conditions were comparable, we enrolled a number of Ph.D. students to act as "confederate agents".[9] They played according to the same random algorithm, which ruled the choices of the robot partner and their verbal reactions were the same as those performed by the robot. To ensure they performed the VRs at the right moment, and only when required by the randomized treatment assignment, we devised a visual cue that they were trained to identify.[10] Ph.D. students were paid an hourly salary for their involvement in the experiment in accordance with university regulations and Italian fiscal rules. In addition, they were told that all of the money corresponding to the payoffs they obtained in the games would be collected and divided fairly among all Ph.D. students involved in the research, weighing each individual share by the number of games he/she played. Please note that none of this

[4] In Sandoval et al. (2016) human partners were asked to be neutral, to interact as little as possible with subjects, and to avoid conversations so as to behave in a similar way to robots. For this reason, they were instructed not to talk but only to nod at subjects in return for their greetings at the beginning of the experiment. Furthermore, an experimenter (called the "referee") was always present during the interactions in the lab room, thus likely introducing a strong bias in subjects' behavior.

[5] For a survey of research contributions on Human–Robot Interactions see also Gaggioli et al. (2021).

[6] Attanasi, García-Gallego, Georgantzís, and Montesano (2013), in a potentially infinitely repeated PD, used a different strategy and obtain implicit dialogic interactions between players through the communication of proposals and counterproposal. In our design, verbal reactions are exogenously determined by the experimenter.

[7] No VR is activated when the aggregate CC Pareto optimal outcome is obtained, i.e when both subject and partner cooperate since the aim of the VR is to move the interaction towards the social optimum.

[8] The test consisted in a simulated version of the interaction: subjects were presented the same instructions as when they would actually take part in the experiment and were asked the subject what the outcome would be in case of a hypothetical choice: half of the subjects were given the CC outcome, while the other half were given the NN outcome. We then verified that this test did not affect the subsequent choices of the subjects: the average cooperation rates in the first round of the game are 0.665 in the NN and 0.682 in the CC version of the tests. This difference is not statistically different from 0 ($p = 0.729$), therefore we can exclude that the test's instruction exerts any priming effect on subjects' choices.

[9] According to the APA Dictionary of Psychology, a confederate agent in an experimental situation is defined as "an aide of the experimenter who poses as a subject but whose behavior is rehearsed prior to the experiment.", see https://dictionary.apa.org/confederate.

[10] The visual cues are displayed as ITEM 6 in Appendix A.

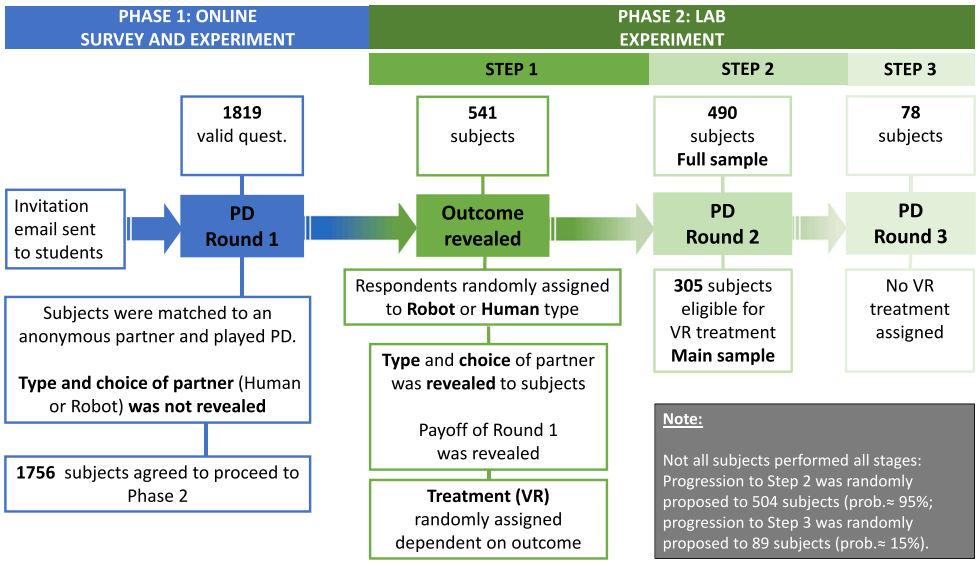

Fig. 1. Flow chart of the experiment. Blue and green boxes and frames refer respectively to Phase 1 (Online) and Phase 2 (Lab).

information was known to the subjects, and thus could not influence their beliefs or choices.[11]

Our experimental design explicitly aims at comparing Human–Human with Human–Robot Interactions. Therefore, verbal reactions were designed for this purpose. We are aware that alternative designs could have explored further mechanisms, such as a voice-over vs. a partner's VR to test the mere reminder effect of the social desirability of cooperation. However, including more treatment arms in an already complex two×two design would have not been feasible under our financial and time constraints. Nonetheless, we believe that further research could expand our protocol (or a subset of it) to address this specific question.

3.1. Experimental procedures

The design of our experiment is summarized by the flowchart in Fig. 1.

We sent invitation emails to all freshmen and sophomore undergraduate students of *Università Cattolica del Sacro Cuore*, enrolled and attending lectures at the Milan campus. The email included a link to an online survey consisting of Phase 1 of the experiment. In addition to a set of psychological, attitudinal, and socio-demographic questions, the survey included the first round of a repeated Prisoner's Dilemma (as in Kreps, Milgrom, Roberts, & Wilson, 1982). At the beginning of each round of the Prisoner's Dilemma (both online and in the lab phases), each player is assigned 3 euros – as initial endowment – and can choose between two alternative actions: if he/she chooses "Cooperate", the entire sum is transferred to the partner who thus receives twice the initial sum (i.e. 6 euros). If he/she chooses "Not Cooperate", he/she keeps the original sum (see Fig. 2).

Table 2
Experiment's monetary payoff matrix.

		Partner's choice	
		Cooperate (C)	Not Cooperate (N)
Subject's choice	**Cooperate (C)**	(6,6)	(0,9)
	Not Cooperate (N)	(9,0)	(3,3)

In each round, players choose simultaneously, after confirming that they understood the task. At the end of each round, players are reminded of their own and their opponents' choices and are shown the resulting payoffs. The total payoff is the undiscounted sum of the round payoffs (plus the one-off show-up fee).[12]

The game proposed to subjects in this experiment consists of a maximum of 3 rounds of a two-person, two-strategy, game (as summarized in Table 2).[13]

Students who agreed to take part in the incentivized game were informed that they had been randomly paired with an unknown anonymous (human or artificial) partner who was playing the same game and they were informed of the Italian average cooperation rate in similar experimental situations.[14] If they agreed to participate, they were asked to make their choice (either "Cooperate", or "Not cooperate"). The outcome of the experimental session would only be revealed to subjects' who agreed to attend and take part (in person) in the second phase of the experiment at the university Lab. Only at that point would

[11] We are aware that, as an alternative strategy, we could have matched each participant to a subject truly playing his/her preferred strategy, but, in this way, it would not have been possible to: (i) control for the occurrence of a verbal reaction (VR), which requires human partners express a VR similar to the one given by robots in order to compare the effects; (ii) make sure that the rate of cooperation among partners was the same in the Human and in the Robot groups, recalling that the strategy of the robot was fixed at a probability of cooperation equal to 50%. For these reasons, we believe that the strategy we chose was the preferable way to address the trade-off between having partners always play their preferred strategy and the possibility of comparing choices after a VR, which is the main goal of this paper.

[12] It seems reasonable to assume that intertemporal discount rate (IDR) is not relevant for our experimental setting given that: (i) when subjects play their first round they are unaware of the delay between the first and second rounds; (ii) this delay is short (on average about 7 days); (iii) all subsequent rounds are played immediately one after the other in the experiment room. However, to address potential biases driven by this issue, in Appendix C we also control for a measure of IDR.

[13] The subjects were not aware of the number of iterations, also some probabilistic "noise" in the number of iterations was added to avoid possible information spillover across subjects.

[14] The partners' cooperation rate was set at 50%, to be consistent with the average cooperation rate observed in 8 experimental studies, which are published in 5 papers, involving Italian subjects, namely Ciardo, Ricciardelli, Lugli, Rubichi, and Iani (2015), Gallucci and Perugini (2000), Meier, Pierce, Vaccaro, and Cara (2016), Pepitone et al. (1967, 1970). Furthermore, the weighted average of the cooperation rate in these studies is 49.23%. Data retrieved from https://app.cooperationdatabank.org/.

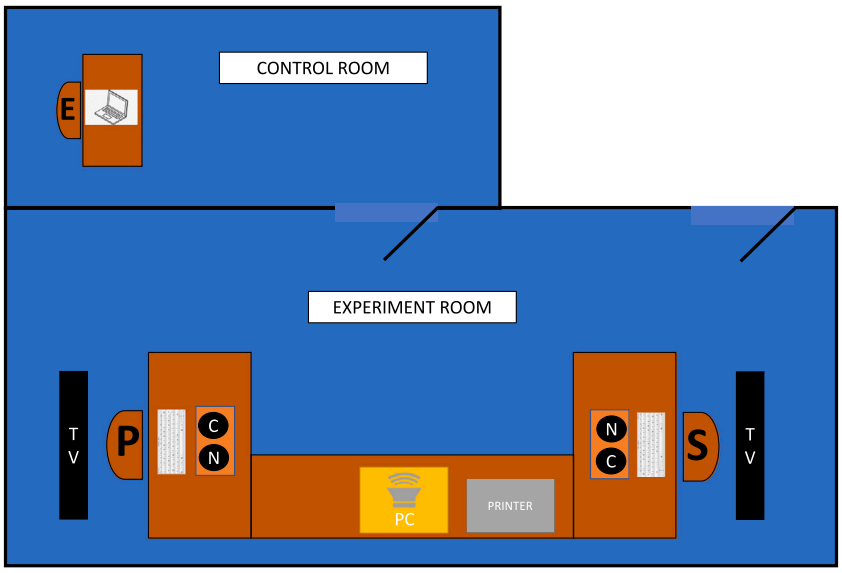

Fig. 2. Experiment room set. S = subject; P = partner; E = experimenter.

respondents discover whether the partner they had faced in the online interaction was a human or a robot. The outcome of the interaction would also be revealed and they might receive an invitation to play another round of the game (or more). Respondents were informed that they would receive the payoff and show-up fee only after having participated in Phase 2 of the experiment.

To proceed to Phase 2, students had to schedule an appointment through an online third-party application. Phase 2, the proper Lab Experiment, consisted of (a maximum of) three sequential rounds. Subjects were unaware of the length of the iteration series. In Step 1, the subject was taken to the lab room by the experimenter, where he/she met and faced his/her partner, whether Human or Robot,[15] with whom he/she was told he/she had played the 1st round of the PD online in Phase 1. When the subject entered the experiment room, in the Robot condition, the robot was already there; whereas, in the Human condition, the human partner entered the room through a different door (leading to a hidden control room) at the same time as the subject.

Once the subject entered the lab room and sat down in front of his/her partner (either robot or human), the experimenter told the subjects that all instructions would be displayed on the TV screen in front of them and read aloud by a voice-over device. The experimenter informed subjects that, as stated in the invitation letter and in the online questionnaire, they would keep all the money earned in the experimental sessions (Phases 1 and 2). If the subject was facing the robot, the experimenter explained that the money earned by the robot would be kept in a special account used for both robot maintenance and hardware and software upgrading.[16] Finally, the experimenter told subjects that he would wait outside the lab room and that he could be called by knocking at the entrance door, in case of need.[17]

Then the experimenter left the room, and the partner (either a Ph.D. student or a humanoid robot) greeted the subject and verbally introduced him/her/itself. In all sessions, the Ph.D. students introduced themselves by their first name, greeted the subjects and asked the subject's first name; while the robot, introduced itself by stating the following text in all experimental sessions: "Hi, I am ToM, a humanoid robot developed by Softbank Robotics able to perform complex interactions with human beings. What's your name?". Finally, the TV screen would reveal the outcome of the Prisoner's Dilemma played online in Phase 1.

Based on the game outcome, a random algorithm determined whether to activate the VR with a 50% probability, triggering the partner – whether human or robot – to deliver the appropriate verbal stimulus in the treated group as shown in Table 3.

Once the outcome of the first round of the game has been revealed, that he/she played online, the subject was asked whether he/she would like to continue to play another round of the game, according to a random algorithm with a probability of 95%. If the subject accepted, a second round was then implemented, without any further treatment assignment. We checked that the probability of both the proposal and acceptance to proceed to Stage 2 were not conditional on the outcome observed in Step 1.[18]

After the results of the second round were revealed, subjects were asked (according to a random algorithm with a 15% probability) whether they wished to continue with a further round of the game. If the subject agreed to play again, a third round of the PD was implemented, again without any treatment assigned. Step 3 of the experiment was of no interest to the analysis. As already explained in the main text, we devised the possibility of playing 3 rounds of the PD

[15] The Robot was NAO, a humanoid robot produced by Softbank Robotics (see Figure D5 in Appendix D for a picture, Gelin, 2019 and Robaczewski et al., 2020 for references); human partners were Ph.D. students recruited and trained for this task. We checked that the choice of subjects in the sample is not affected by being matched to specific individual Ph.D. students. More details in Appendix C.

[16] In a previous work (see Manzi et al., 2021), we administered a survey to a sample of young adults (aged 18–29) investigating their expectations regarding perceived robot needs and desires. From this survey, it emerged that most of the subjects in the sample consider the improvement of hardware and software properties as the main expected need for a robot.

[17] The English translation of all texts are reported in Appendix A, while the original Italian layout is available in the Supplementary Materials.

[18] We analyzed the association between a categorical variable, identifying all 4 possible outcomes observed in Step 1, and (i) a binary variable, identifying whether Step 2 had been proposed to the subject; (ii) a binary variable, identifying whether this proposal had been accepted. In both cases, we are able to exclude any non-random pattern. In case (i), a Chi-sq. test ($Chi = 4.723, df = 3, p = 0.193$) and an ANOVA ($F - stat = 1.58, p = 0.194$) allow inferring that the proposal to continue to Step 2 was independent of the observed outcome of Step 1. Figure D1 in Appendix D summarizes this outcome; in case (ii) the result is even stronger since only 16 subjects out of 506 (3.2%) refused to continue, and any association between acceptance and outcome in Step 1 was excluded by both a Chi-sq. test ($Chi = 1.111, df = 3, p = 0.774$) and an ANOVA ($F - stat = 0.37, p = 0.776$). Figure D2 summarizes this result.

Table 3
Outline of *Verbal Reactions* performed by Partner and description of samples.

Outcome (Subject, Partner)	VR type	Statement	Obs VR	Obs No VR	Obs Total
(C, N)	Apology	"I realize I made a mistake in our online interaction. I meant to press C to cooperate. However I pressed N by mistake. I am really sorry! I will be more careful next time"	109	98	**207**
(N, C)	Reprimand	"I am really upset. If you had chosen to cooperate we would have gained 6 euro each, a reasonable amount! On the contrary you exploited my goodwill and I got nothing"	26	27	**53**
(N, N)	Disappointment	"What a pity! If we had chosen to cooperate we would have gained 6 euros each. Why not cooperate in the next round?"	22	23	**45**
(C, C)	None	None	–	–	**185**

Notes: Observations relate to Step 2 of Phase 2, i.e. the core sample of our analysis. The "Main sample" only includes outcomes in orange-shaded cells; the "Full sample" includes both the outcomes in the orange-shaded and yellow-shaded cells.

only to generate uncertainty about the total number of rounds so as to limit the possible effects of any information sharing across students.

Following the last incentivized interaction, the total amount gained was communicated to subjects through the TV screen. Then, subjects were asked whether they would like to donate 1 euro out of their total payoff to a charity of their choice (as shown in Appendix A), along the line of Eckel and Grossman (1996).[19] We decided to include this task to gather information about individual "types" in terms of generosity, thus allowing for a subsequent heterogeneity analysis. However, we found no significant effects of this variable on the probability of cooperation at any round and therefore omitted any discussion of this specific result from the paper.

Once this last choice was made, the subject had his/her receipt automatically printed and he/she was invited to leave the experiment room and enter the check-out area, where he/she was paid the total amount (show-up fee *plus* payoffs *minus* donation, if this was the case) in cash. Upon leaving the lab the student signed a confidentiality agreement requiring him/her not to share any information about the experiment with fellow students. At the end of the experimental session, we fully disclosed all experimental procedures to all subjects.

We planned to run the experiment from February 13th to March 13th, 2020. However, due to the outbreak of the COVID-19 pandemic, we had to stop all activities in the lab on February 21th, which proved to be our last day of data collection in 2020. In February 2021, universities in Italy opened again for in-person lectures. We seized the opportunity and intended to run another wave of the entire experiment (both Phases 1 and 2) between February 22nd and March 31th.[20] Again, due to the resurgence of the Covid-19 pandemic, we had to stop all activities in the lab on March 4th, complying with the restrictions imposed by the Italian national law.[21]

4. Data and estimation methods

4.1. The sample

We sent a total of 23,552 emails and received 2,205 individual answers and 1,819 valid and completed questionnaires (Phase 1); 1,756 respondents (96.5% of the valid entries) agreed to participate in the Second Phase of the experiment.

Out of a total of 541 subjects showing up to Phase 2,[22] 490 subjects (henceforth "Full sample"), agreed to proceed to Step 2 (see Table D3, bottom panel). Within this subset of subjects, the sample eligible for an analysis of the treatment effect consists of 305 subjects (henceforth "Main sample") who had made their choice in Round 2, having been assigned either to the Treatment group (Verbal Reaction, VR) or to the Control group (No VR) (see Table D3, top panel).[23] An initial inspection shows that subjects in both the Full and Main samples are more likely to cooperate than not to cooperate (see Table D3 in Appendix D).[24] The average payoff obtained by subjects in the Full sample ($n = 490$) is 8.14 euros, whereas in the Main sample ($n = 305$) is 6.66 euros.[25] In addition to this, all subjects also received 4 euros as a show-up fee.

Table 4 shows the summary statistics for both the outcome and control variables by treatment group and experimental condition: the Robot and Human conditions are shown in the top and bottom panels respectively; treatment and control values are shown across columns. As the table shows, all subsamples are well-balanced across treatment groups in terms of control variables (the only minor imbalances relate to gender). A summary of balance tests, estimated using a Logit model, is also displayed in Appendix B, showing that observations are overall

[19] The available alternative charities were *Médecins Sans Frontières* and *Greenpeace*.

[20] Invitations were sent to freshmen and, to avoid duplication to those sophomores who had not opened the invitation e-mail we sent them when they were freshmen in 2020.

[21] The number of subjects taking part in the experiment is very similar in 2020 and 2021. As explained below, we control for the experimental wave in all our models and find that in most cases the dummy variable identifying the experimental wave is not statistically significant.

[22] As explained above, we had to stop our lab experiments due to COVID-19-related restriction rules implemented in Italy, both in February 2020 and March 2021: that is the main reason why only a fraction of all subjects accepting to proceed to Phase 2 then showed up to the lab. We checked that the sample of included subjects is representative of the whole sample of people interviewed online. The balance test's outcome is shown in Figure D4, in the Appendix, and reports only an imbalance for Freshmen that are over-represented in the sample of subjects taking the lab experiment in Phase 2 (61.8% of subjects in Phase 1 are Freshmen while in Phase 2 the share is 78.4%). Probably, since the winter term is usually less demanding for Freshmen, they were able to book earlier slots at the lab. To account for this imbalance, we included a control variable for freshmen in all model specifications, showing that it is not biasing our results.

[23] As illustrated in Table 3 a VR stimulus could only be applied in 3 out of 4 possible outcomes, i.e. excluding the CC outcome.

[24] Note that most of the subjects' cooperative choices, i.e. those yielding an outcome of "CC" in Step 1, although part of the experiment, and rewarded as all other subjects, were not included in the Main sample of our analysis, since they could not be eligible for VR assignment: therefore, by design, subjects in the Main sample are "less cooperative" than in the Full sample.

[25] The difference in the average payoffs for the two samples is easily explained by considering that CC outcome (and the consequent pay-off: 6,6) is not included in the Main sample. Although this outcome is possible – and indeed occurred within the experimental interactions included in the Full sample – it does not lead to a potential VR, and thus it cannot be included in the counterfactual analysis. We, therefore, excluded CC outcomes from the main analysis presented in the paper, although all subjects in the full sample were indeed paid.

Table 4
Summary statistics, by experimental condition.

Partner=Robot								
	VR group			No VR (control group)			T-test	
	Mean	St. Dev.	Obs	Mean	St. Dev	Obs	Diff.	t-stat
Outcome variables								
Cooperate (Online)	0.736	0.443	91	0.650	0.480	80	−0.086	(−1.22)
Cooperate (Lab)	0.630	0.486	81	0.380	0.489	71	−0.249**	(−3.15)
Control variables								
Female	0.769	0.424	91	0.575	0.497	80	−0.194**	(−2.73)
Freshman	0.857	0.352	91	0.800	0.403	80	−0.057	(−0.98)
Economics	0.440	0.499	91	0.450	0.501	80	0.010	(0.14)
Test failed	0.099	0.300	91	0.075	0.265	80	−0.024	(−0.55)
Partner=Human								
	VR group			No VR (control group)			T-test	
	Mean	St. Dev.	Obs	Mean	St. Dev	Obs	Diff.	t-stat
Outcome variables								
Cooperate (Online)	0.640	0.483	86	0.663	0.476	86	0.023	(0.32)
Cooperate (Lab)	0.737	0.443	76	0.558	0.500	77	−0.178*	(−2.34)
Control variables								
Female	0.628	0.486	86	0.802	0.401	86	0.174*	(2.57)
Freshman	0.663	0.476	86	0.791	0.409	86	0.128	(1.89)
Economics	0.302	0.462	86	0.395	0.492	86	0.093	(1.28)
Test failed	0.151	0.360	86	0.186	0.391	86	0.035	(0.61)
Robot vs. Human comparison								
						Obs	Diff.	t-stat
Cooperate (Online): VR						177	−0.10	(1.39)
Cooperate (Online): No VR (control group)						148	−0.01	(0.17)
Cooperate (Lab): VR						157	−0.11	(1.44)
Cooperate (Lab): No VR (control group)						148	−0.18**	(2.19)

Notes: Summary statistics refer to subjects eligible for treatment, i.e. excluding those yielding "CC" as the outcome of the Online choice. All Lab choices refer to Step 2 of the Lab Phase, as described in Fig. 1.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

well balanced across treatment groups (Figure B1) and partner types (Figure B2).[26]

Interestingly, the baseline measure of our outcome variable (i.e. the choice to Cooperate in the online Phase 1) is also not statistically different across treatment groups. Conversely, a t-test shows that subjects assigned to the treatment group (in Phase 2, Step 2, in the lab) are more likely to cooperate than those in the control group, already suggesting a potential effect of the VR on subjects' choices in Step 2.

Finally, the bottom panel of Table 4 provides a summary of the differences in VR effects by subsample, showing that the only statistical difference in the mean values is in the No VR condition, where the cooperation rate is about 18 percentage points lower for subjects facing a Robot than for subjects facing a Human partner. In the next sections, we detail specific tests to assess also the heterogeneity of the effect across partner types.

4.2. Estimation technique

We addressed the aforementioned set of research questions by implementing a Linear Probability Model (LPM) through Ordinary Least Square (OLS),[27] in which the probability that respondent i makes a Cooperative choice in the lab is conditional on a set of control variables and experimental conditions. Formally:

$$Y_i = \begin{cases} 1 & \text{if respondent } i \text{ chooses Cooperate in Step 2 in the Lab} \\ 0 & \text{if respondent } i \text{ chooses Not Cooperate in Step 2 in the Lab} \end{cases}$$

$$Pr(Y_i = 1|Robot_i, VR_i, X_i) = \beta_0 + \delta_1 Robot_i + \delta_2 VR_i + \delta_3 Robot_i \times VR_i + \beta_1 Cooperate(Online)_i + \beta_j X_{j,i} \cdots + \beta_k X_{k,i} \quad (1)$$

where Y is the dependent variable, measured at Step 2 in the Lab: we label this variable Cooperate (Lab); $\beta_0, \beta_1, \beta_j, \ldots, \beta_k$ and δ are the parameters to be estimated, with δ_1, δ_2 and δ_3 being respectively the main coefficients of interest for RQ1, RQ2 and RQ3; $Cooperate(Online)_i$ is the choice made by respondent i in the Online Phase; and $X_j \cdots X_k$ are a set of k experiment-related and control variables, illustrated below.

To increase the precision of the estimates and to account for potential confounding factors, all models are estimated using different specifications. Control variables include experiment-related controls and background characteristics of the subjects. In the former set of controls, we include *Instruction order*, a categorical variable that takes into account which of the potential outcomes of the Prisoner's Dilemma is shown first to the subject during the instructions session of the game, to control for potential priming[28]; *Wave*, to control for the year of the experiment (either 2020 or 2021); and *Experiment day* which relates to the number of days the experiment has been running (e.g. Day 1, 2, 3, ...) when the subject played, to control for potential spillover

[26] In this case, Freshman is statistically significant, and greater in the case of Robot rather than Human partner. In Figure B3 we also provide a balance test for the two experimental waves, 2020 and 2021, showing that only Freshmen are more represented in 2021.

[27] The use of LPM instead of Logit models has been debated in the economics literature, at least since Horrace and Oaxaca (2006). However, Angrist and Pischke postulate that "if the CEF is linear, as it is for a saturated model, regression gives the CEF — even for LPM." (see http://www.mostlyharmlesseconometrics.com/2012/07/probit-better-than-lpm/). We replicated all our LPM results using Logit, finding almost identical results, available upon request.

[28] Subjects received the illustration of all potential outcomes of PD in random order. We control for the outcome that appears first to the subject, to account for the instructions having a potential priming effect. We also replicated our results re-coding this variable to account for the outcome that was shown last. The results remain unchanged.

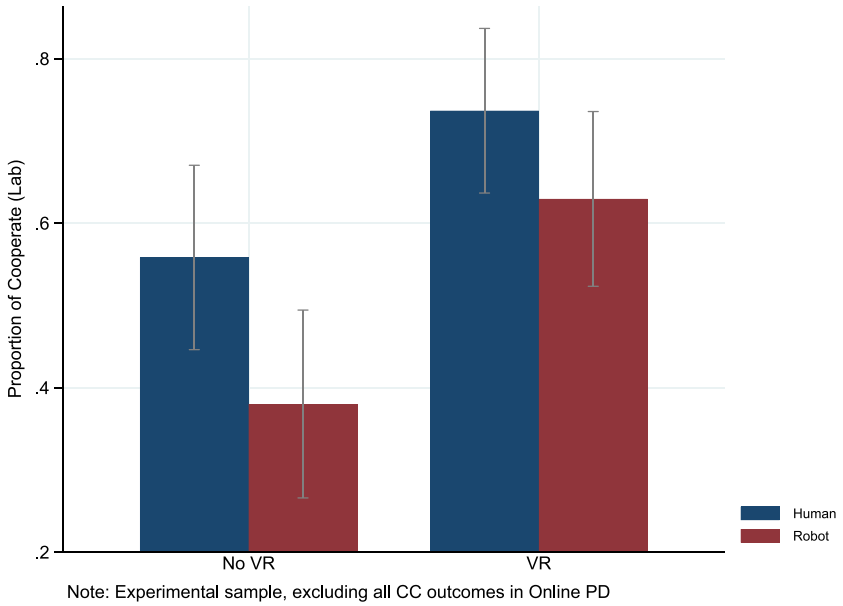

Fig. 3. Cooperate (Lab) by experimental condition and partner type.

effects. In the latter group of controls we include *Female*, to control for subjects' gender; *Freshman*, to control for subjects taking the first year of their BA; *Econ*, to control for subjects enrolled in BA degrees in either Economics, Management, or Finance and Banking; and *Fail test*, to control for subjects failing the pre-game test designed to verify their full understanding of the game's procedures, rules, and payoffs. Moreover, all models are also estimated excluding this latter group of subjects. The description of all variables is summarized in Appendix D, Table D2.

Furthermore, since our experimental design consists of two×two treatment conditions, in Appendix C, Section C.VI, we test the first two hypotheses by estimating separately the effects of *Robot* and *VR* using sample splits.

5. Results

5.1. Cooperation patterns

On average, all subjects eligible for treatment who proceeded to Step 2 (Main sample) chose to Cooperate 58% of the times, as shown in the top panel in Table D3. The breakdown by treatment and experimental condition is illustrated in Fig. 3 and summarized in Appendix D, Table D1.

As Table D1 shows, when no VR is assigned the average cooperation rate in Step 2 is much lower in the Robot condition, by about 17.8 percentage points, and this difference is statistically significant (p=0.030).

When the VR is assigned, subjects still display higher cooperation rates when the partner is Human, but the difference between the Human and Robot conditions is no longer statistically significant (p=0.152). Overall, this descriptive pattern suggests that – in the absence of any verbal reaction to the observed outcome of the online Phase 1 – subjects are more likely to cooperate when facing a Human Partner. However, if a verbal reaction (VR) is delivered by either a Robotic or Human partner, subjects are more likely to cooperate than in the case of no reaction. These patterns are summarized in Fig. 3, in which the bars of the VR group are clearly higher than those of the No VR group for both the Human and Robot conditions.

Our main hypotheses are tested through the estimation of Eq. (1), whose outcome is shown in Table 5.[29] In this table there are three main coefficients of interest: *Robot*, *VR* and *VR*×*Robot*, which we present separately in the next paragraphs.

5.2. The effect of partner type

The estimated coefficient of *Robot* in Table 5 provides evidence that subjects display different behaviors depending on the type of partner in the No VR condition. Being assigned to a Robot partner, which does not verbally react, significantly decreases the cooperation rate, with an estimated probability between 15.2 and 22.1 percentage points, depending on specifications, compared to a human partner who, similarly, does not verbally react.

The middle panel of the table reports also the estimated marginal effects of partner type for subjects assigned to the VR treatment, which is obtained as the sum of the coefficients of *Robot* and *VR*×*Robot*. Also in this case, the negative effect of being assigned to a Robot emerges as a stable pattern, although the size of the effect is a bit smaller and statistically significant only when the choice in the online Phase 1 is also included. Overall, these results show that *ceteris paribus* cooperation is strongly and significantly affected by partner type, especially in the "control" group where no VR is allowed. This result is aligned with the evidence provided by a systematic review of experiments involving computer players performed by March (2019) who finds that player behavior differs for human vs. computer opponents and that subjects generally behave more selfishly and rationally when interacting with computers.

5.3. The effect of VR (verbal reaction)

To evaluate the effect of VR on cooperation rates, we rely on two coefficients in Table 5. The coefficient of *VR* shows that a VR positively affects cooperation rates when subjects face a human partner. In this case, subjects cooperate more, when a VR is introduced, with an increase in probability between 15.8 and 17.9 percentage points. Similarly to the previous Section, the sum of the coefficients of *VR*

[29] Please note that all results are almost identical if control variables are not included.

Table 5
LPM: Main outcome, treatment effect, Main sample, including control variables.

DV: "Cooperate (Lab)"	Benchmark		Including Cooperate (Online)	
	Main sample	Excl. failed tests[a]	Main sample	Excl. failed tests[a]
VR × Robot	0.047	0.095	0.025	0.078
	(0.121)	(0.129)	(0.117)	(0.124)
VR	0.179	0.158	0.178	0.150
	(0.084)**	(0.093)*	(0.083)**	(0.090)*
Robot	−0.152	−0.199	−0.166	−0.221
	(0.088)*	(0.093)**	(0.086)*	(0.092)**
Cooperate (Online)			0.276	0.290
			(0.061)***	(0.065)***
VR × Robot + Robot	−0.105	−0.104	−0.142	−0.144
	(0.083)	(0.087)	(0.079)*	(0.082)*
VR × Robot + VR	0.227	0.253	0.203	0.223
	(0.082)***	(0.085)***	(0.080)**	(0.083)***
Controls	Yes	Yes	Yes	Yes
Inst. order	Yes	Yes	Yes	Yes
Wave	Yes	Yes	Yes	Yes
Exp. day	Yes	Yes	Yes	Yes
Adj. R-sq.	0.05	0.06	0.11	0.13
Obs	305	268	305	268
LL	−200	−174	−190	−164
AIC	441	387	422	367
BIC	515	455	500	439

Notes. LPM model (OLS), dependent variable: choice of Cooperation (Lab) at Step 2, Prisoner's Dilemma. Observations related to outcome "CC" in Phase 1 (Online) are excluded from this sample. Robust standard errors are in parentheses. Control variables include: Female, Freshman, Economics, and Failed test.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
[a]This sub-sample excludes all subjects who failed the test assessing their full comprehension of the instructions.

and *VR×Robot* provides an assessment of the effect of verbal reaction when subjects are assigned to the Robot condition. As the table shows, the effect is always positive and statistically significant, implying an increase in cooperation rates between 20 and 25.1 percentage points. Therefore, we can conclude that communication clearly induces higher cooperation rates, for both subjects facing a human and a robot partner.

All types of VR exert a positive influence on the cooperation rate, as summarized in Figure D3. However, the lack of statistical power, due to the limited sub-sample sizes, hinders a disaggregated analysis. These results provide new evidence of the positive effect of communication in fostering cooperation in strategic decision-making, strengthening and confirming previous similar findings (such as Kollock, 1998; Sally, 1995).

5.4. Heterogeneity of VR effect across partner types

In the previous section, we showed that the effect of VR is strong and significant under both (Human and Robot) experimental conditions. In this section, we show the outcome of the test to assess the heterogeneity of effects across partner types, i.e. whether being faced with a Human or Robot partner affects the way verbal interactions promote cooperative choices. In Table 5, the coefficient of the interaction variable $VR \times Robot$ provides the outcome of this test. As the table shows, the coefficient is never significantly different from 0.

Note that the coefficients of both *Robot* and VR are significantly different from 0, as shown in the previous sections. Furthermore, the marginal effect of VR in the Robot condition, given by the sum of VR×Robot and VR, is also always strongly statistically significant.

Therefore, while the effect of VR is large and significant, this effect is not heterogeneous across partner types. In other words, once the respondent is randomly assigned to the VR treatment group, he/she is more likely to cooperate with his/her partner, irrespective of whether the partner is a Human or a Robot. The lack of heterogeneity provides a striking result: as long as the partner in the PD provides a VR, subjects respond by increasing on average their probability of cooperating.

The results shown in Table 5 may depend on the belief that a robot which is able to implement a sensible contextual verbal interaction[30] is perceived as "more human", thus elicits a behavior that a subject normally reserves for fellow human partners. Alternatively, the VR may produce the effect by simply acting as a soft reminder of the social desirability of cooperation.[31]

As explained in Section 3, our experimental framework is unable to disentangle between these two competing and/or complementary explanations. However, it is worth noting that despite being able to display appropriate gaze and body gesture cues to increase its appearance of "socialness", NAO is still far from being able to reproduce the pitch, accent, and expressiveness of a natural human voice.[32]

Further research is thus needed to disentangle these two possible explanations. A possible strategy would have implied an additional experimental arm in which the VR is provided by the Game Director (the pre-recorded voice-over) commenting on the realization of socially suboptimal results. In this case, the "ethical reminder" would be separated from the partner (and, in particular, from the robot). Nonetheless, our results seem consistent with some recent empirical findings that support the importance of communication in HRI using the same social robots employed by our experiment (NAO). Pelikan and Broth (2016) find that subjects tend to use the same signals as in human–human

[30] Note that we can exclude that the effect is due to the mere fact of NAO being able to talk since it greeted any subject it interacted with, regardless of whether it would perform a VR at a later time.

[31] Similar to the role played by mentioning the Ten Commandments as in Mazar, Amir, and Ariely (2008), or the honor code as in McCabe and Trevino (1993) in stimulating academic honesty.

[32] NAO communicated with the subject via a "*Wizard of Oz* system" controlled by a laptop PC located in the control room. All moves and speech items were coded into the system. The robot followed a pre-programmed protocol where the experimenter did not need to speak or type anything during the interaction and only had to press a button to start the interaction, as in Laban, George, Morrison, and Cross (2021).

interactions, such as adjusting word selection, turn length, and prosody, thus adapting to the perceived limited capacity of the robot. Tahir et al. (2018) asses two modalities to deliver the feedback: audio only and audio combined with gestures and show that when audio and gesture are combined subjects better understand the feedback delivered by the robot.

5.5. Robustness checks

We addressed potential sources of bias in our results through a set of robustness checks, described in detail in Appendix C. Through these ancillary analyses, we are able to show that our main results are robust to the exclusion of *Econ* students from the sample who account for a relevant share of the sample, (around 36%) and are known to be more self-interested than their peers either because of a self-selection process or an indoctrination effect. The coefficients of VR and Robot, shown in Table C1, are almost identical to those shown in Table 5, although a few specifications are on the margin of statistical significance, mostly due to the reduced sample size. The interaction term is never statistically significant, as in all other cases.

We also tested whether some psychological and/or behavioral traits of the subject may have influenced our main result. For this reason, we run another estimation which includes a series of validated psychological scales measuring risk, trust, inter-temporal discount and a "raw" measure of generosity based on Eckel and Grossman (1996). None of the psychological scales nor the Generosity index appear to significantly influence our outcome variable, namely *Cooperate (Lab)*, as displayed in Table C2. Conversely, the main coefficients of interest (VR, *Robot*, and $VR \times Robot$) have similar size and significance as in the main model specifications.

Another check involved the possible effect of the partner's choice in the online phase. Also, this variable, which turns out to be significant and positive in some specifications, does not influence the main effect of our treatment, since, also in this case, all the coefficients of interest (VR, *Robot*, and $VR \times Robot$) have similar size and significance as in the main model specifications. These further results are shown in Tables C3 and C4.

Moreover, we tested whether the previous beliefs about robots[33] might have influenced the subjects' choices when playing against a robot. Also in this case, we find no effect for the previous belief, while all main coefficients of interest keep the same size and similar significance (with only a few minor exceptions), as shown in Table C5.

In a similar way we tested also whether the partner's gender influenced the subjects' choices when playing against a human (confederate) agent. In this case, we observe no effect of the partner's gender in the human subsample, although the significance of VR is weakened, probably due to the limited sample size, as shown in Table C7.

Finally, we also checked for potentially different outcomes when using sample splits, consistently showing that while the effect of VR can be detected both in the Human and Robot partner conditions, the effect of partner type is only present in the case of No VR. These further checks are displayed in Tables C8 to C11.

6. Discussion and conclusions

Human–Robot Interactions are expected to become more and more common in the near future due to the increase in the employment of social anthropomorphic robots in many different occupations and environments. Therefore, it is increasingly important to understand how human subjects behave when facing a social robot in complex strategic interactions and how trust might be developed in these contexts.

In this paper, we devised a randomized experiment in which human subjects are randomly matched to either a human or an anthropomorphic robot partner and are asked to perform a repeated Prisoner's Dilemma to investigate (i) whether subjects behave differently depending on the nature of their partner (Human or Robot); (ii) whether a Verbal Reaction (VR), which implicitly refers to cooperation as a socially desirable strategy, influences the subject's choice in a subsequent round of the game; (iii) whether the effect caused by the VR depends on the nature of the partner (Human or Robot).

We find that facing a robot decreases cooperation rates by about 15 to 22 percentage points on average while being exposed to a VR when facing a Robot makes cooperation between 20 and 25 percentage points more likely at the subsequent round of the PD. Interestingly, and most importantly, the differential effect in cooperation driven by communication is not statistically different in the Human and Robot conditions.

Our results thus suggest that: (i) subjects tend to act more cooperatively with fellow human beings, rather than with robots; (ii) subjects are influenced by a VR towards a more cooperative strategy; and (iii) the effect of a VR is strong enough to make the difference in behavior, based on the partner's type, insignificant.

We show that although people are less likely to cooperate with a robot partner than with a human partner, this difference disappears when a VR is performed. Our result on the VR may depend on the belief that a robot that is able to implement a verbal interaction with the subject appears to be "more human", and deserve a behavior similar to that reserved for fellow human partners, thus supporting the notion of "algorithm aversion" as described by, among others, Burton, Stein, and Jensen (2020) and Chugunova and Sele (2022).

Alternatively, it may consist of a reminder effect of the social desirability of cooperation. Further research is still needed to disentangle the effects of verbal interactions in HRI versus the reminder effect due to the "mere" message delivered.

We are aware of the possible limitation of extending lab experiment results to the "real world", as evidenced for instance by Levitt and List (2007), Rabin (1993) and Levitt and List (2008). Further research may extend our findings to non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) subjects, following the suggestion by Henrich, Heine, and Norenzayan (2010).

Overall, we are convinced that our findings may have interesting implications in a number of cases where robots are used to interact with – especially fragile – human beings (nursing homes, care facilities, hospitals, kinder gardens, etc.) suggesting that an apologizing robot can positively influence the cooperative behavior of a human subject, should a minor contrast occur.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We acknowledge funding from two competitive research funds of Università Cattolica del Sacro Cuore, Italy: D3.2/2018 "Human–Robot confluence" and D3.2/2020 "Behavioral change". This project is a spin-off of the Fetzer Institute's "#3534.00" project granted to M.A. Maggioni on the use of behavioral economics techniques to assess behavioral change. We thank the anonymous referees, two associate editors, together with M. Antonj, M. Colagrossi, E. Colombo, A. Gaggioli, A. Galliera, E. Giacobino, G, Lombardi, C. March, A. Marchetti, P.

[33] We empirically measured these beliefs as a dichotomous variable based on the following question: "When interacting with a human, the robots simply execute a predetermined set of program lines or adapts its behavior to the interaction?".

Natale, F. Perali, G. Riva, G. Sandini, A. Sciutti, L. Stella, A. Tanevska, F. Trombetta, D. Walentek, the participant to the DEF Seminar, Milan May the 27th 2021 and IIT-Contact Seminar, Genua September the 14th, 2021, for useful comments and observations on previous versions of this paper. The usual *caveats* apply. Research assistantship by E. Cerolini, P. Gambacciani, C. Marconi, F. Manzi and P. Zaza is acknowledged. Special mention is due to F. Manzi whose expertise and skills were precious for programming the robot in the lab experiment. We also thank N. Graverini and P. Alberti for technical IT support and C. Codella and L. Sersale for the effective management of students recruitment.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socec.2023.102011.

References

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, *97*(4), 543–569.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464–477.

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, *103*(418), 570–585.

Attanasi, G., García-Gallego, A., Georgantzís, N., & Montesano, A. (2013). An experiment on prisoner's dilemma with confirmed proposals. *Organizational Behavior and Human Decision Processes*, *120*(2), 216–227.

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

Ben-Ner, A., Putterman, L., & Ren, T. (2011). Lavish returns on cheap talk: Two-way communication in trust games. *The Journal of Socio-Economics*, *40*(1), 1–13.

Bicchieri, C. (2002). Covenants without swords: Group identity, norms, and communication in social dilemmas. *Rationality and Society*, *14*(2), 192–228.

Bicchieri, C., & Lev-On, A. (2007). Computer-mediated communication and cooperation in social dilemmas: an experimental analysis. *Politics, Philosophy & Economics*, *6*(2), 139–168.

Braver, S. L., & Wilson, L. (1986). Choices in social dilemmas: Effects of communication within subgroups. *Journal of Conflict Resolution*, *30*(1), 51–62.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*, 220–239.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Charness, G., Feri, F., Meléndez-Jiménez, M. A., & Sutter, M. (2021). An experimental study on the effects of communication, credibility, and clustering in network games. *The Review of Economics and Statistics*, 1–45.

Chugunova, M., & Sele, D. (2020). *We and it: An interdisciplinary review of the experimental evidence on human-machine interaction*: *Max Planck Institute for Innovation & Competition Research Paper*, (20–15).

Chugunova, M., & Sele, D. (2022). An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, Article 101897.

Ciardo, F., Ricciardelli, P., Lugli, L., Rubichi, S., & Iani, C. (2015). Eyes keep watch over you! Competition enhances joint attention in females. *Acta Psychologica*, *160*, 170–177. http://dx.doi.org/10.1016/j.actpsy.2015.07.013.

Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M. A., Greco, A., et al. (2021). Promises and trust in human–robot interaction. *Scientific Reports*, *11*(1), 1–14.

Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., et al. (2018). Cooperating with machines. *Nature Communications*, *9*(1), 1–12.

Dasgupta, P. (1988). Trust as a commodity. In D. G. Gambetta (Ed.), *Trust* (pp. 49–72). New York, NY: Basil Blackwell.

DeAngelo, G., & McCannon, B. C. (2020). Psychological game theory in public choice. *Public Choice*, *182*(1–2), 159–180.

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., et al. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, *23*(12), 1549–1556.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.

Dumouchel, P., & Damiano, L. (2017). *Living with robots*. Cambridge, USA: Harvard University Press.

Eckel, C. C., & Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*(2), 181–191.

Farrell, J. (1995). Talk is cheap. *The American Economic Review*, *85*(2), 186–190.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, *10*(3), 103–118, URL http://www.jstor.org/stable/2138522.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868.

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), Article 2053951719860542.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3–4), 143–166.

Gaggioli, A., Chirico, A., Di Lernia, D., Maggioni, M. A., Malighetti, C., Manzi, F., et al. (2021). Machines like us and people like you: Toward human–robot shared experience. *Cyberpsychology, Behavior, and Social Networking*, *24*(5), 357–361.

Gallucci, M., & Perugini, M. (2000). An experimental test of a game-theoretical model of reciprocity. *Journal of Behavioral Decision Making*, *13*(4), 367–389.

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), 60–79.

Gelin, R. (2019). NAO. In A. Goswami, & P. Vadakkepat (Eds.), *Humanoid robotics: A reference* (pp. 147–168). Dordrecht: Springer Netherlands, http://dx.doi.org/10.1007/978-94-007-6046-2_14.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Horrace, W. C., & Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, *90*(3), 321–327.

Hsieh, T. Y.-T., Chaudhury, B., & Cross, E. S. (2020). Human-robot cooperation in prisoner dilemma games: people behave more reciprocally than prosocially toward robots. In *Companion of the 2020 ACM/IEEE international conference on human-robot interaction* (pp. 257–259).

Hsieh, T.-Y., & Cross, E. S. (2022). People's dispositional cooperative tendencies towards robots are unaffected by robots' negative emotional displays in prisoner's dilemma games. *Cognition and Emotion*, *36*(5), 995–1019.

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

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology*, *70*(1), 47–65.

Klockmann, V., von Schenk, A., Villeval, M. C., et al. (2021). Artificial intelligence, ethics, and intergenerational responsibility. *HAL Open Science*, halshs-03237437.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, *24*(1), 183–214.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS One*, *3*(7), Article e2597.

Kreps, D. M. (1990). *A course on microeconomic theory*. Princeton University Press.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, *27*(2), 245–252.

Laban, G., George, J.-N., Morrison, V., & Cross, E. S. (2021). Tell me more! Assessing interactions with social robots from speech. *Paladyn, Journal of Behavioral Robotics*, *12*(1), 136–159.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, *21*(2), 153–174.

Levitt, S. D., & List, J. A. (2008). Homo economicus evolves. *Science*, *319*(5865), 909–910.

Manzi, F., Sorgente, A., Massaro, D., Villani, D., Di Lernia, D., Malighetti, C., et al. (2021). Emerging adults' expectations about the next generation of robots: Exploring robotic needs through a latent profile analysis. *Cyberpsychology, Behavior, and Social Networking*, *24*(5), 315–323.

March, C. (2019). *The behavioral economics of artificial intelligence: lessons from experiments with computer players*: *154 BERG Working Paper Series*.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835.

McCabe, D. L., & Trevino, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education*, *64*(5), 522–538.

Meier, S., Pierce, L., Vaccaro, A., & Cara, B. L. (2016). Trust and in-group favoritism in a culture of crime. *Journal of Economic Behaviour and Organization*, *132*, 78–92. http://dx.doi.org/10.1016/j.jebo.2016.09.005.

de Melo, C. M., Carnevale, P., & Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. *3*, In *The 10th International conference on autonomous agents and multiagent systems* (pp. 937–944).

de Melo, C. M., & Terada, K. (2020). The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. *Scientific Reports*, *10*(1), 1–8.

Miettinen, T., & Suetens, S. (2008). Communication and guilt in a prisoner's dilemma. *Journal of Conflict Resolution*, *52*(6), 945–960.

Miwa, K., & Terai, H. (2012). Impact of two types of partner, perceived or actual, in human–human and human–agent interaction. *Computers in Human Behavior*, *28*(4), 1286–1297.

Nouri, E., & Traum, D. (2013). A cross-cultural study of playing simple economic games online with humans and virtual humans. In *International conference on human-computer interaction* (pp. 266–275). Springer.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), 137–158.

Ostrom, E., & Walker, J. (1997). Neither markets nor states: Linking transformation process in collective action arenas. In D. C. Mueller (Ed.), *Perspectives on public choice: A handbook* (pp. 35–72). Cambridge: Cambridge University Press.

Paeng, E., Wu, J., & Boerkoel, J. (2016). Human-robot trust and cooperation through a game theoretic framework. *30*, In *Proceedings of the AAAI conference on artificial intelligence* (1), (pp. 4246–4247).

Pelikan, H. R., & Broth, M. (2016). Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4921–4932).

Pepitone, A., Faucheux, C., Moscovici, S., Cesa-bianchi, M., Magistretti, G., Iacono, G., et al. (1967). The role of self-esteem in competitive choice behavior. *International Journal of Psychology*, *2*(3), 147–159. http://dx.doi.org/10.1080/00207596708247212.

Pepitone, A., Maderna, A., Caforicci, E., Tiberi, E., Iacono, G., di Majo, G., et al. (1970). Justice in choice behavior: A cross-cultural analysis. *International Journal of Psychology*, *5*(1), 1–10. http://dx.doi.org/10.1080/00207597008247285.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 1281–1302.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395–405.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage*, *22*(4), 1694–1703.

Robaczewski, A., Bouchard, J., Bouchard, K., & Gaboury, S. (2020). Socially assistive robots: The specific case of the NAO. *International Journal of Social Robotics*, 1–37.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92.

Sandoval, E. B., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics*, *8*(2), 303–317.

Skeath, A. D.-S. (Ed.), (1999). *Games of strategy*. WW Norton & Company.

Sung, J., Grinter, R. E., & Christensen, H. I. (2010). Domestic robot ecology. *International Journal of Social Robotics*, *2*(4), 417–429.

Tahir, Y., Dauwels, J., Thalmann, D., & Thalmann, N. M. (2018). A user study of a humanoid robot as a social mediator for two-person conversations. *International Journal of Social Robotics*, 1–14.

Terada, K., & Takeuchi, C. (2017). Emotional expression in simple line drawings of a robot's face leads to higher offers in the Ultimatum Game. *Frontiers in Psychology*, *8*, 724.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Walker, J., Cox, J., & Smith, V. (1987). Bidding behavior in first-price sealed-bid auctions: Use of computerized Nash competitors. *Economics Letters*, *23*, 239–244.

Wu, J., Paeng, E., Linder, K., Valdesolo, P., & Boerkoel, J. C. (2016). Trust and cooperation in human-robot decision making. In *The 2016 AAAI fall symposium series*: *Technical Report FS 16-01*.

Yezer, A. M., Goldfarb, R. S., & Poppen, P. J. (1996). Does studying economics discourage cooperation? Watch what we do, not what we say or how we play. *Journal of Economic Perspectives*, *10*(1), 177–186.

Zörner, S., Arts, E., Vasiljevic, B., Srivastava, A., Schmalzl, F., Mir, G., et al. (2021). An immersive investment game to study human-robot trust. *Frontiers in Robotics and AI*, *8*, 139.