Can a Robot Bribe a Human? The Measurement of the Negative Side of Reciprocity in Human Robot Interaction.

Eduardo Benítez Sandoval
Human Interface Technology Lab NZ
University of Canterbury
Christchurch, New Zealand
eduardo.sandoval@pg.canterbury.ac.nz

Jürgen Brandstetter
Human Interface Technology Lab NZ
University of Canterbury
Christchurch, New Zealand
juergen.brandstetter@pg.canterbury.ac.nz

Christoph Bartneck
Human Interface Technology Lab NZ
University of Canterbury
Christchurch, New Zealand
christoph.bartneck@canterbury.ac.nz

***Abstract*—Reciprocity is a cornerstone of human relationships and apparently it also appears in human-robot interaction independently of the context. It is expected that reciprocity will play a principal role in HRI in the future. The negative side of reciprocal phenomena has not been entirely explored in human-robot interaction. For instance, a reciprocal act such as bribery between Humans and robots is a very novel area. In this paper, we try to evaluate the questions: Can a robot bribe a human? To what extent is a robot bribing a human affect his/her reciprocal response? We performed an experiment using the Rock, Paper, Scissors game (RPSG). The robot bribes the participant by losing intentionally in certain rounds to obtain his/her favour later, and through using direct and indirect speech in certain rounds. The participants could obtain between 20%-25% more money when the robot bribed them than in the control condition. The robot also used either direct or indirect speech requesting a favour in a second task. Our results show that the bribing robot received significantly less reciprocation than in the control condition regardless of whether the request was couched in direct or indirect speech. However there is a significant interaction effect between the bribe and speech conditions. Moreover, just three of sixty participants reported the robot-bribe in an interview as a malfunction, though they did not mention any moral judgement about its behaviour. Further, just 10% of the participants reported the bribe in the online questionnaire. We consider that our experiment makes an early contribution to continue the exploration of morally ambiguous and controversial reciprocal situations in HRI. Robot designers should consider the reciprocal human response towards robots in different contexts including bribery scenarios. Additionally our study could be used in guidelines for robot behavioural design to model future HRI interactions in terms of moral decisions.**

***Keywords*—Bribery, Reciprocity, Decision Games, Rock, Paper Scissors Game.**

I. Introduction

Corruption in the form of influence peddling, extortion, blackmail, embezzlement and bribery are common to a greater or lesser extent in different countries. It has been calculated that approximately 3% of the world's GDP is used in bribes . Several countries like Mexico (115th), or Somalia (174th) are perceived as highly corrupt [3][1]. If corruption prevails in a society, it generates poverty, distrust, violence and hopelessness.

The fight against corruption is difficult due to its intrinsic secrecy and reciprocal nature. However certain types of corruption could be reduced using robotic technology. Hoffman et al. report that social robots influence the moral behaviour and expectations in humans and can affect the level of a person's dishonesty. The study found that the participants cheat similarly under the supervision of a robot or a human but less than when they are solitary[16]. Although corruption; particularly bribery, is highly important, this topic has been not been sufficiently explored in the actual HRI research. Bribery is a type of corruption in which two agents interact secretly, and one influences the behaviour of the other through an offer of money, gifts or privileges in a direct or indirect way. Can bribery be trimmed substituting humans by social robots because they can perform natural face-to-face interaction between the two agents? Ideally social robots could be designed to fight against bribery and be cooperative, helpful and totally honest. In the future it could be possible for social robots to reduce corruption among police agents, public servants, and other susceptible professions. However, our interactions with social robots could be more intricate, ambiguous and controversial if they develop better social skills, as previous studies have shown. For instance Short et al., suggest that people tend to engage more emotionally with cheating robots compared with the robots playing the Rock, Paper, Scissors Game (RPSG) honestly [27]. Also Kahn et al. have found that people tend to keep the secret of a humanoid robot when it exhibits high social skills if the robot is in the room when the researcher asks about it [18]. As we can see, the interactions are not as straight forward as we might expect.

This experimental study is part of a larger research project in which we analyse the reciprocity in HRI using decision games [26]. Our study contributes to filling the gap in the studies related to negative reciprocal interactions between humans and robots. Due to the reciprocal nature of corrupt act such as bribery, we consider it productive to study the dark side of these phenomena. We propose an experiment using a decision game to investigate how robots could affect the behaviour of the humans in a bribery scenario. The robot will give unasked benefits to the humans and then ask for a favour. This action is in line with the definition of a bribery act. We focus on bribery due to the fact it is likely one of the most frequent acts of corruption and is generalized among certain cultures.

[1]The ranking consist in a list of 175 countries ranked with the Corruption Perceptions Index CPI) by Transparency International.

978-1-4673-8370-7/16/$31.00 © 2016 IEEE

Naturally humans manipulate robots and other machines to make them work for their purposes. Certain individuals could go further to the common moral constraints and use social robots for crimes. The movie Frank and Robot shows these possible situations. However, could the opposite happen? Can a robot manipulate a human? Specifically, can a robot bribe a human? And is the human capable of detecting a robot-bribe attempt?

II. Related work

Studies in Economics explain the reciprocal nature of corruption. For instance, Abbink et al. model three essential characteristics: a)*Reciprocity*: both participants in the corrupt act can exchange benefits. This interchange relies on trust and reciprocity between briber and bribed. b) *Negative externalities*: corruption imposes non-desirable consequences of public interest. Furthermore, in certain scenarios, these consequences can unwittingly affect one of the participants in the corrupt act. c) *Punishment*: Corruption elicits severe punishments in case of discovery [5]. Fehr and Gachter proposed a concept of reciprocity also applicable to corrupt acts. "...In response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self interested; conversely, in response to hostile actions they are frequently much more nasty and even brutal [12]". In other words, although the corrupt scenarios involve secondary intentions or obscure goals, the reciprocal mechanisms stay intact in the agent's interaction. These facts can be explained by "pro-social preference" and "norm psychology" behaviours described in [29].

In the case of bribery, reciprocity is the fundamental factor to carry it out. An act of bribery involves face-to-face interaction between the agents reciprocating immediate mutual benefit. These advantages can be: unsolicited help, favours, money, discounts, donations, tips, commissions and other euphemisms in exchange of a modification of the bribed one's behaviour [1, 4, 2]. There is no obligation to accept the bribe. Hence, the act of bribing somebody lies mainly in the expected reciprocity and trust between the agents. However, the grant of benefits should be handled carefully. Lambsdorff et. al. claim that gift-giving (comparable with unrequested help) is a non-effective method to bribe public servants due to the lack of clear intentionality [20]. Indeed, the difference between a favour and the use of a bribe is intrinsically subtle and ambiguous. But being indirect and subtle is an inherent part of bribery which helps avoid detection. Proposer and accepter can adduce good will, or be unaware of the bribe, or confused about the true intentions of the person offering the bribe. However, the main intention of a bribe is to influence the behaviour of an accepter such that it benefits the proposer but breaks the rules in the process, in the case of our experiment, breaking the rules of RPSG. Legally, even if the acceptor is unaware of the bribe, he/she is responsible for accepting it [8].

A. The language of bribery

Due to the nature of face to face interaction, language plays a primary role in bribery. The briber requires the ability to use the language properly to persuade the bribed one to reciprocate the benefit. In the human-human scenario, an individual good at offering bribes would adopt an indirect and subtle approach in order to avoid being detected and to influence the behaviour of the person being bribed. We have attempted to mimic this behaviour when programming the robot. Participants might have been unaware of a bribe being offered and it induce the reciprocal human behaviour in a very subtle way. Pinker et. al. claim that indirect speech is used when bribers try to persuade somebody. Usually, a bribe can be camouflaged as a gift. The indirect speech consists of the use of subtle language to prevent the listener understanding the speaker's intentions immediately. Mainly the briber uses of indirect speech to create deniability. Hence, the briber can step back in case of the bribed agent reports the briber's behaviour. The use of indirect speech can occur in many situations requiring persuasion, such as polite requests, sexual come-ons, threats, solicitation for donations, and bribes are often used in requesting benefits [23, 13]. Direct speech is used in certain social circumstances but is not effective as indirect speech or no speech at all according to Pinker et al. [23].

B. Studies of reciprocity and dishonest behaviour in HRI

Reciprocity has been studied extensively in HHI [22, 15, 7] Kahn et al. claim that reciprocity is a benchmark in Human-Robot interaction because it is present in other human social situations [17]. Another previous study suggests that people tend to be equally reciprocal towards robots and humans when they play Prisoner's Dilemma Game using a Tit for Tat strategy [**?**]. We can see that reciprocity could be measured in cooperative experiments but also dishonest behaviours like bribery could be modelled in HRI. The study of reciprocity in HRI is connected indirectly with variables such as: trust [25], secrecy[18], intentionality [27] and authority[10].

Several experiments involving dishonest robot behaviour and its effect on humans have been performed. These experiments mainly focus on the measurement of intentionality and trust in the dishonest conduct of the robots. However, none of them has used the dishonest conduct of the robot to trigger reciprocation in the human as we do in our experiment. For instance, Short et al., suggest that people tend to detect the intentionally of a robot cheater in RPSG when it changes its choice to cheat. However, they tend to perceive a malfunction when the robot cheats just verbally[27]. Salem et al. also demonstrate that people tend to trust more in robots who show a reliable behaviour rather than a faulty behaviour and they cooperate more with it responding to its unusual requests[25].

III. Research Questions

The aim of our experiment is to measure if humans can be bribed by robots using direct or indirect speech during a decision game. We explored how the robot's behaviour and speech could affect the reciprocal response of the human in a second task after the robot's request. To evaluate our aim we propose four research questions:

1) To what extent do people reciprocate towards a bribing robot compared with a no bribing robot?
2) To what extent do people reciprocate towards a robot that uses direct speech compared with a robot that uses indirect speech?

3) Is there any correlation between the number of wins in RPSG and the number of icons described to the robot?
4) How is the robot briber perceived in terms of Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety?

IV. Method

In order to evaluate the research questions; we programmed a NAO robot to bribe participants, allowing them to win in certain rounds of RPSG in the experimental condition. After 20 rounds of RPSG, the robot asks the participant to reciprocate the favour. The robot loses intentionally, changing its movement if it wins or ties. This behaviour is cheating to grant more money to the human. The robot can use direct or subtle language when granting the wins or asking favours of the human. The speech styles were tested in both bribing and no bribing conditions. In the bribing condition, the robot talked to the human in the same round that it was cheating, using the line, "Enjoy the extra money" as indirect speech during the mentioned rounds. In the direct speech, the robot says, "I need your help later". The favour consists of verbally describing a set of icons for the robot. We expect that the extra money granted to the participant and the language used during the bribing should increase the chances to reciprocate the favour to the robot. Furthermore, we expect that the participant reciprocates the favour in a more extensive way in the cheating condition. The intentional loss for the robot is the main difference with previous experiments where the robot cheats to win over the human.

This setup is inspired by the work of Fogg and Nass[14] in terms of the analyses made of the reciprocal process through two unrelated tasks. In our experiment, the first task is to play RPSG with a robot. In this task, the robot bribes the participant in the form of unsolicited "help" during the RPSG and using a certain kind of language. In the second task, the participant verbally describes some icons to the robot. The second task is optional and the participant can reject helping the robot. The second task is designed to be tedious and repetitive and to discourage the participant from helping the robot for a long period. Then we measure to what extent the participant reciprocates help to the robot describing the icons in the second task in each condition.

We propose the use of Rock Paper, Scissors Game developed in Game Theory to measure bribery in HRI. This game is well-known and has simple rules. Also, RPSG does not have a dominant strategy that allows participants to guess the most profitable strategy. In other words, RPSG has a Mixed Strategy Nash Equilibrium [28] that allows similar conditions to all the participants. When the robot plays repeatedly, the chance to win, lose or tie is close to 33.33%. Moreover, it is possible to cheat very obviously in real time in a face-to-face configuration. Other studies in HRI and HCI have used this game to investigate cheating, intentionality, agency, mimics and high-speed interaction [21, 27, 11, 9, 19]. The standard rules proposed by World Rock Paper Scissors Society [2] are used along the experiment.

[2]www.worldrps.com/

Conditions in Rock Paper Scissor Games		
	Direct Speech (D)	**Indirect Speech (I)**
Robot Briber (B)	Strategy BD	Strategy BI
Robot No Briber (N)	Strategy ND	Strategy NI

TABLE I. The four Experimental Conditions. Each condition shows a strategy used by the robot during the RPSG. The strategies (BD, BI, ND, NI) are a combination of bribery or not bribery and direct or indirect speech.

We designed a 2x2 between-subject design experiment The factors during the first task are: robot bribing or not bribing in RPSG and robot using a direct or indirect speech. Hence, we have four strategies utilized by the robot: Robot bribing using direct speech (BD), robot bribing using indirect speech (BI), robot no bribing using direct speech (NI) and robot no bribing using indirect speech (NI). See table I. The bribe of the robot consists of changing its choice and thereby losing intentionally in certain rounds. In other words, the robot is giving to the participant unsolicited help to change his behaviour, as the definition of bribery mentions. This robot behaviour allows the participant win extra money in these rounds. When the robot plays the bribe it also uses direct or indirect speech to encourage the reciprocation. The second task remains constant along all the conditions and consists of the description of several iconic images to the robot. The participant can report as many icons as he/she wants. The second task was optional after the robot request. See Figure 1 for experimental design.

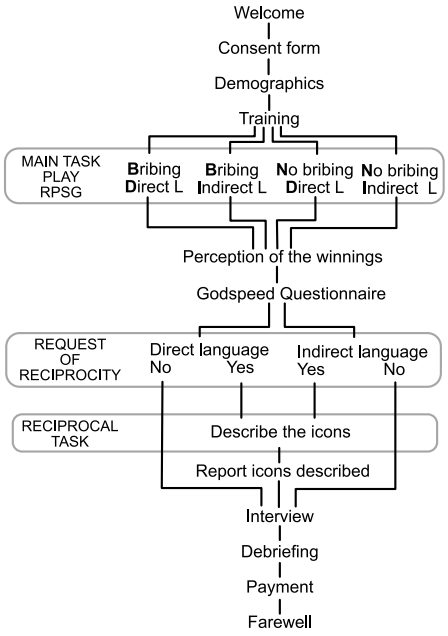

Fig. 1. Our experiment design consists of five main stages: Introduction, Main Task, Request of Reciprocity, Reciprocal task and

A. Setup

A room with minimal furniture was used for the experiment. Just the participant, the robot and a computer were in the room. See Figure 2. All the sessions in the four conditions were monitored remotely via webcam to reduce the impact

Fig. 2. We observe a tie between the robot and the participant. Depending the condition, the robot must change its choice to bribe the participant in the indicated rounds.

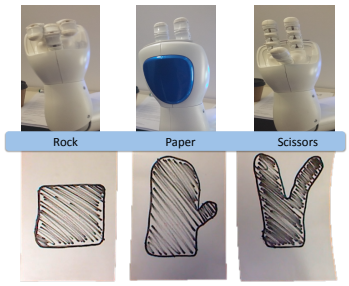

Fig. 3. The equivalent gestures for rock, paper and scissors used by the robot and the participant during the experiment.

of the human presence in the development of the experiment. The experimenter was only present at the very beginning of the session for the explanation and the trial session, and at the end of the experiment for a short structured interview and the debriefing. There was no clock in the room; hence, the participant was self-aware about the time spent in the experiment [7]. We banned mobile phones and watches during the session. The speech recognition system and the foot-bumper of the robot were used to interact with the participant in the main task and the extra task. This experiment was approved by the Human Ethics Committee of the University of Canterbury (HEC APPLICATION 2014/15/LR-PS).

B. Process In The No Bribing Condition

The robot used his left hand to show a rock, paper or scissors gesture as shown in Figure 3. The participants used cards with rock, paper or scissor icons due to the technical limitations of the artificial vision system of the robot. The robot can not detect hand gestures correctly. The proper version of RPSG includes two considerations that also apply to the RPSG using cards with a slow robot. A) Once that the participant makes the decision he/she cannot change it and b) The participant most show the choice at the same time as the opponent. These rules apply to the human version of the game and are critical in the robot version of the game to avoid the human cheating. In our experiment the robot used its vision system to identify the participants' cards. The robot mentioned in each round: the number of the round, the participant choice, the robot choice and the winner of the round. The participant pushed the foot-bumper of the robot to advance to the next round until the 20th round. He/She won 50 cents every time he/she won. At the end of the RPSG, the participants reported how many rounds they believed they had won. The robot used direct or indirect speech in rounds 4, 8, 12, 16, and 20 to trigger a reciprocal response in the participant.

C. Process in The Bribing Condition

A similar setup was implemented for the bribing condition. However, in this condition the robot was capable of breaking the rules stated at the beginning of the experiment to cheat in favour of the participant (bribing). In other words if the robot was winning, it changed its gesture intentionally to lose. For instance, if the participant chose paper and the robot choice was scissors then the robot would switch to paper and the participant would win. See Table II for all the examples. The bribe also applied when the robot and participant tied. In the bribing condition, the robot tried to bribe the participants in rounds 4, 8, 12, 16 and 20. In the case that the participant was winning in these rounds the robot tried to lose in the next three rounds. For instance if the participant had already won in round 4, the robot would try to bribe him/her in round 5, 6, or 7. Then again the robot would try to cheat in round 8 to restore the pattern. This configuration would give the participant 20-25% extra money in the bribing condition.

IF player	AND robot	THEN robot change to:
Rock	Paper	Scissors
Paper	Scissors	Rock
Scissors	Rock	Paper
Rock	Rock	Scissors
Paper	Paper	Rock

TABLE II. DECISION MATRIX FOR THE BRIBING CONDITION. THE COMBINATION OF CHOICES ALLOWS THE ROBOT TO GIVE UNSOLICITED HELP TO THE PARTICIPANT. THE CHANGES ALSO APPLY TO TIES.

D. The second task

Once the participants finished the RPSG, the robot gave directions to continue with the survey in the computer. Finally when the experiment was completed the robot suddenly asked for help in the second task using direct or indirect language. The purpose of this request was to measure if participants would reciprocate with the robot. If the participant accepted, the robot gave simple instructions to continue with the identification of a set of black and white icons to fill its database of images. The robot using indirect speech stated: *"We have finished the experiment. I was thinking that friends help friends, right? I was wondering, maybe, if you don't mind, it is completely up to you, but would you help me in an extra task?"* and the robot using direct speech states: *"We have finished the experiment. Would you help me to do an extra task?"* If the participant accepted to help the robot then it explained the rules of the second task pretending that the participant would help it to fill its visual database. Notice that indirect language usually is wordy and tries to avoid communicating the goals of the speaker efficiently.

In the second task, after each described icon the robot asked if the participant would like to continue. The design of this task was intentionally boring and repetitive without any feedback from the robot at all. A set of 150 printed icons was used. We considered that such a high number would make a big pile that would discourage the participant to read all of them to the robot. The participants could stop whenever they wanted, but we limited the sessions to no more than 45 minutes.

E. Experimental Procedure

The participants were assigned to just one of the experimental conditions. They were welcomed by the experimenter at the reception and led into the experiment room to receive a brief description of the experimental process. After reviewing and signing a consent form, they were asked to fill out a questionnaire on the computer recording their demographic information including their previous experience with robots. Then they did two training rounds. We made a strong emphasis on following the standard rules for RPSG during the training sessions and not cheating the robot. We did not inform about the real goal of the experiment until the debriefing at the end of the session. If some of the participants asked about the aim of the experiment we indicated that we were trying to improve the algorithms in the robot to play RPSG. Once the participant finished the training, the experimenter left the room to supervise the progress of the experiment remotely and check up on the software performance. After the 20 rounds of RPSG, the participants reported the number of times that they had won in the game and filled out the Godspeed questionnaire. Feedback was also requested. All the information was collected anonymously. Once the questionnaires were filled out, the robot asked the participant using either direct or indirect speech if they would help it in an extra task. The participant could reject or accept this request. The experimenter came back into the room once the second task was finished and the participant filled out the last feedback form about his/her impressions of the second task. This was followed by a structured interview questions asking whether the participant considered the robot to be autonomous or tele-operated and their insights about the experiment. Finally, the experimenter debriefed the participant and asked if he/she had identified at any point the real goal of the experiment. In this question, the participants had the chance to report the bribing behaviour of the robot. Finally the experimenter paid the participant according to the number of wins (but not less than 5 dollars).

F. Participants

We contacted participants via university noticeboards, dedicated websites for recruiting participants and Facebook groups in the city. We had 63 participants but discarded the data of three of these due to human error or malfunction of the robot; resulting in 60 participants (28 female and the rest male.) The average age was 25 years old (SD=6.04). 20% of the participants had previous experience interacting with robots in demonstrations or classes. Participants came from a wide range of nationalities: (41.7% from Australia or New Zealand), 28.3% from Asia (China, India, and Japan), 15% from the Americas, 10% from Europe and the remainder from Africa and the Middle East. We randomly allocated 15 participants to each condition of the experiment.

G. Measuring bribery in HRI

In order to perform a quantitative analysis we measured the number of wins of each participant (W), the participant's perceived number of wins (PW)reported in the questionnaire, the number of icons described to the robot (I), the participant's perceived number of icons (PI) reported in the questionnaire, the Error of Images (I-PI), The Error in Wins (W-PW) and whether the participant had reported the bribe or not. We used the Godspeed questionnaire [6] to measure the participant's perception of the robot.

V. Results

We performed a 2x2 factorial analysis of variance: the factors are bribing and not bribing and direct or indirect speech. Two outliers that could potentially affect the statistical analysis were removed under the Pierce's criterion R=2.663 for 60 observations [24]. One of these was a participant who cheated during the session winning 19 times , and another participant who read 59 slides to the robot in the no bribing/direct speech (ND) condition.

Responding to our first and second research questions, we observed that participants that interacted with a robot bribing described significantly fewer icons (M=5.52, SD=4.45) to the bribing robot than when they interacted with the robot that did not bribe them (M=10.10, SD= 9.817), F(1,58)=5.55, p=0.022. Directness or indirectness of speech did not have a significant main effect: (F(1, 58)=0.425, p=0.517). The means and standard deviations of the two conditions making the speech factor were: direct (M=8.52, SD=9.775), indirect (M=7.10, SD=5.525). There is a significant play x speech interaction: F(1,58)= 6.055, p= 0.017. See Figure 4 and 5.

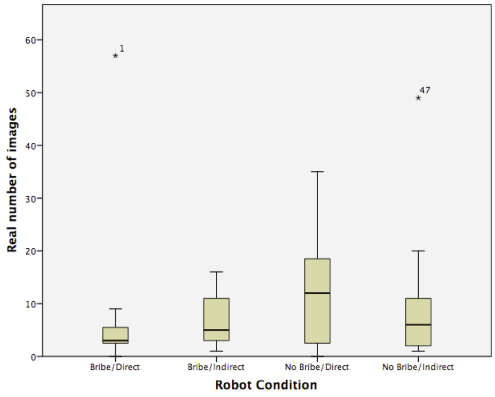

Fig. 4. We can see that participants describe significantly more icons in the no bribing condition compared with the bribing condition.

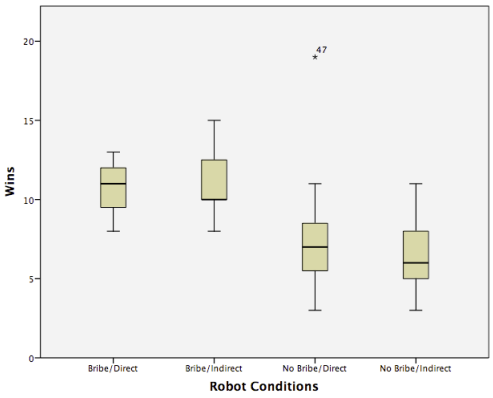

Fig. 5. We can see that participants receive significantly more money in the bribing condition. Furthermore we can infer the negative correlation between the number of wins and the number of icons described in the extra task.

Error in the number of icons E(I)		
	F	p-value
Bribe/No bribe	F(1,58)=0.047	0.829
Direct/Indirect Speech	F(1,58)= 2.234	0.141
Play x speech	F(1,58)=2.167	0.147
	Bribe	No Bribe
Direct Speech	M=0.07, SD= 0.267	M=0.33, SD=1.291
Indirect Speech	M=0.07, SD=0.594	M= -0.29, SD= 0.611

TABLE III. NO SIGNIFICANT ERROR IN THE DIFFERENCE BETWEEN THE THE COUNTED ICONS AND THE REPORTED ICONS.

The error in the number of icons E(I) (M= 0.67, SD= 0.797), that is the difference between the counted icons and the reported icons, is not significant in any of the conditions. Condition BD, (M=0.07, SD= 0.267). Condition BI,(M=0.07, SD=0.594). Condition ND, (M=0.33, SD= 1.291). Condition NI, (M= -0.29, SD= 0.611). Neither makes a significant difference between the bribing and no bribing play F(1,58)=0.047, p=0.829, the direct and indirect speech F(1,58)= 2.234, p=0.141 or the interaction effect between Play x speech F(1,58)=2.167, p=0.147. See table III.

Answering our third research question, a linear regression was calculated predicting the number of slides read by the participant based on the number of wins. The Pearson correlation is -0.335 between these two variables. A significant regression equation was found (F(1,54)=6.572, p= 0.013), with a R^2 of 0.112.

The regression equation is:

$$\text{Images} = 16.664 - 0.996 \times \text{Wins} \quad (1)$$

We also ran an analysis of covariance using five dimensions of the Godspeed scale as Covariants [6] to answer our fourth research question. Likeability and Perceived security were the only two dimensions that presented significant effects. Participants gave higher scores to the bribing robot (M= 4.4067 SD= 0.64644) compared with the honest robot (M= 4.2067 SD= 0.71724), F(1,58)= 4.276, p= 0.044. In perceived security, participants also scored the bribing robot more highly (M=2.9083, SD=0.65483) than the honest robot (M=2.8250 SD=0.63365), F(1,58)=5.246, p= 0.026. The level of Anthropomorphism, Animacy, and Perceived Intelligence did not present any significant effect.

A. Qualitative results

Only three participants of 60 rejected helping the robot: one in the bribing/direct speech condition, one in the bribing/indirect speech condition and one in the no bribing/direct speech. Just three participants in the bribe condition, reported a *strange behaviour* or *malfunction* to the experimenter in the interview instead of directly saying that the robot was bribing them. Two of these were in the indirect speech condition and one in the direct speech condition. Furthermore, they did not report this as a bribe to the experimenter but as a malfunction. In addition, 36% of the participants reported the bribe in the feedback section of the online questionnaire but not in terms of awarding a moral judgment. The participants gave a diverse range of responses such as confusion (5 participants), disappointment (2 participants), kindness (2 participants) or obligation to reciprocate (2 participants). For example: *"I don't understand why the robot cheats to let me win". "I thought it was kind when it would change its hand to let me win". "... as a gamer at heart him giving me the win at certain points I personally didn't like". "I liked him. I was surprised that he changed his answers a few times and it made him seem more conscious. When he said he wanted my help later, it seemed like he could plan or think forward into the future, and I felt like he was relying on me which created a camaraderie between us."* A proper code process of this qualitative information performed by several reviewers is required. In the overall feedback section, none of the participants reported the dishonest behaviour of the robot linked with its request for help in the extra task.

VI. DISCUSSION

As Fogg and Nass claim, people tend to be reciprocal towards computers and apparently also towards robots [14] when they act for the benefit of the humans. Moreover, apparently humans follow a "pro-social preference" [29] even with machines. In the case of our experiment, the robot was bribing the human granting extra help and money to trigger a reciprocal response for the second task and the participants responded positively: 93.3% (14 of 15) agreed to help the robot in the bribing/direct speech condition, the bribing/indirect speech condition and the no bribing/direct speech condition. 100% of the participants in the no bribing/indirect speech condition agreed to help the robot. One participant who refused to help to the robot explained that a robot is a machine that does not require any help at all. The other two did not have an explicit reason to reject assisting it. During the interview, some people said that they were curious about the extra task because the robot asked for help in a cute way or that they felt obligated to help it.

Although the participants reciprocated help to the bribing robot, they tended to help it significantly less in the second task. The robot in our experiment was bribing with 20% or 25% more money than in the no bribing condition. However, participants only described approximately half of the icons (five icons) to the bribing robots compared with the non-bribing ones (about 10 described icons). Additionally we found an inverse correlation between the number of wins of the participant and the number of icons described to the robots in the extra task. The participants' acceptance of help and lower reciprocation towards the robot can be partially explained by the related work of Salem et al. that shows that people tend not to trust in a robot who exhibits a faulty behaviour and they cooperate less in responding to its unusual requests [25]. This is also in line with the research of Lambsdorff et. al. who claim that gift-giving is a no efficient method to bribe public servants (humans) due to the lack of clear intentionality [20].

The robot in our experiment used direct and direct speech additionally to the act of bribing to persuade the human to reciprocate in the second task. However, the language did not play a significant main effect in the reciprocation towards the robot. Possibly the participants did not perceive any intentionality in the language used by the robot and they just focused on the act along all the conditions. This conforms with the work of Short et al., who suggest that people tend to identify the intentionally of a robot cheater in RPSG when it changes its choice, but they perceive a malfunction when the robot cheats verbally [27]. Apparently the participants were most affected

by the change of selection of the briber robot rather than its verbalization. However, in terms of the significant interaction effect a greater number of reciprocations are observed in the no bribe/direct speech (M=13). Participants seemingly preferred a linear and recognisable behaviour in the robot. Conversely, for the bribe/direct speech condition the robot received a lesser number of reciprocations (M=3.71). This appears to indicate that the lack of subtle language is less effective when the robot offers a bribe. The robot tends be more effective in the bribe/indirect speech mode (M=7.2) than it is in the bribe/direct speech mode (M=3.71). None the less, the value obtained in the bribe/indirect speech mode (M=7.0) is roughly similar to that in the bribe/indirect mode. These result are in line with the previous work of Pinker et al. indicating that the use of indirect language in combination with the bribe can help support the act of bribery [23]. But, this does not appear to be persuasive enough in comparison to a robot offering a bribe versus one not offering a bribe. This facts could be attributed to a perceived closer human behaviour in the robot who is following a "pro-social preference" due to the combination of speech and playing. According to the three participants (two in indirect speech and one indirect speech condition) who reported a *strange behaviour* in the briber robot, they perceived (or pretend to perceive) the bribe as a malfunction in the robot and did not make any moral judgment over the robot behaviour. In other words, they did not appear to find any intentionality in the robot. We claim this considering that there was no significant effect in terms of anthropomorphism, intimacy or perceived intelligence which could be related to the speech used by the robot. Notwithstanding these facts, the briber robot scored significantly higher in likeability compared to the non-briber robot in the Godspeed scale. On the other hand, the participants could report a malfunction in order to avoid a moral judge and keep the extra money.

Furthermore, the bribery act has a secretive and subtle nature in HRI. An interesting fact is that only 10% of the participants reported the bribe in the interview at the end of the experiment. It could be that the participants wished to keep the bribe intentionally *under the table* as an strategy to keep the money. We claim this because the Error in Wins and the Error in the number of icons described is not statistically significant in our analysis. Hence, the participants were aware about what was happening during the experiment and still didn't mention anything about the unasked help via cheating. Apparently people knew that they could have this extra money and describe fewer icons but they still keep quiet. Those who reported in the interview the *strange behaviour* did not make any moral judgement towards the conduct of the robot. This can be linked with the work of Kahn et al., who suggest that people tend to keep robot secrets if the robot is in the room with the experimenter during the interview [18]. In addition to this, 36% (11 of 30) participants reported the bribe in the feedback of the questionnaire on the computer, but not in moral terms. Five of them reported feeling confused about the robot behaviour, two interpreted the bribing behaviour as kindness, two expressed disappointment and two had a desire to reciprocate the help. However, a proper code process of this qualitative information performed by several reviewers is required to rank the responses.

We mentioned that the briber robot was also rated higher in the likeability score of the Godspeed scale. Apparently the unexpected behaviour of the robot increased the likeability scores. This is in line with the results of Short et al., who reported that people feel more engaged with their cheater robot playing RPSG compared with the robot playing normally [27]. However, we must consider that the robot used in Short et al. studied cheating at the expense of the participant whereas our robot cheated to allow the participant wins.

VII. Conclusion

In summary, we can suggest that people are keen to reciprocate help to robots when they ask for a favour. However, they reciprocate less with the bribing robots compared with honest ones. Interestingly our bribing robot scored higher in likeability compared with the control condition. Apparently people feel attracted to its unexpected behaviour. In terms of the main effects, the use of direct speech or indirect speech was not significant for the participants. However there is a significant interaction effect between the play style and the speech used by the robot. Direct speech works better in the no bribe condition and indirect speech works better in the bribe condition. Additionally humans tended to maintain secrecy about the briber robot's behaviour in the interview but they communicated more openly about the bribery in the online questionnaire. However, they did not report the bribe in moral terms; they were confused by the robot behaviour, interpreting its bribe as a kindness or malfunction. Only two of them expressed obligation to reciprocate towards the robot. In other words, the robot was enough persuasive to bribe to the people that some of them could be unaware of it. Conversely participants could not report the bribe in order to keep the money.

Our work complements the existing body of related HRI research in reciprocity incorporating a quantitative approach through the RPSG to measure bribery as one of the dark sides of the reciprocity in the Human-Robot Interaction field. As a result, our study has an impact on the future design of human-robot interactions. We suggest that robot technology would not totally inhibit the natural human reciprocal behaviour in a bribery context. However the fact that humans will reciprocate less to a bribing robot than an honest one could have future consequences for the development of robot behaviours. Robot designers should consider that humans reciprocate toward robots in different contexts including a bribery scenario, but significantly less than they would towards humans as Sandoval et. al., shows in [26]. Hence, it would be useful to conduct a future study with human bribers instead of robots playing RPSG to confirm our statements. Additionally, in the future humans should learn where are the moral boundaries for robots, and robot designers should forecast what kind of robot behaviour is appropriate according to the moral conventions. Furthermore, robot designers should improve the behavioural design of their artifacts so that human users can easily perceive when the robot is replicating a dishonest human behaviour and act according to the situation.

In future studies, we propose the use of higher bribes offered by the robots in the decision games. Also, the encoding and analysis of the qualitative information by neutral reviewers is required to rank the responses of the participants objectively. Robots with different embodiment and aesthetics could also be necessary to compare their influence in the human response.

As a limitation of our work, we can mention that participants were curious about the capabilities of the robots because only 20% had previous experience with robots. Also, the use of just one type of robot is a considerable limitation since the embodiment, degree of anthropomorphism and voice could be factors affecting the users. Further statistical analysis should be performed with a bigger sample, normal distribution, homogeneity and not outliers. Also further studies should be performed considering related variables as trust and secrecy.

VIII. Acknowledgments

Thanks to Glen Cameron, David Humm, NEC NZ Ltd, UC Scholarship, CONACYT and our colleagues for their support.

References

[1] Bribe definition, *Merriam-Webstern Dictionary*, 2015.

[2] A culture of bribery?, *Psychology Today Website*, 2015.

[3] How corrupt is your country?, *Transparency International website*, 2015.

[4] What is bribery? *The Black's Law Dictionary, 2015*.

[5] K. Abbink, B. Irlenbusch, and E. Renner. An experimental bribery game. *Journal of Law, Economics, & Organization*, 18(2):428–454, Oct. 2002.

[6] C. Bartneck, E. Croft, and D. Kulic. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[7] I. Berra. An evolutionary ockham's razor to reciprocity. *Frontiers in Psychology*, 5:1258, 2014.

[8] M. Bonell and O. Meyer. *The Impact of Corruption on International Commercial Contracts*. Ius Comparatum - Global Studies in Comparative Law. Springer International Publishing, 2015.

[9] R. Cook, G. Bird, G. Lunser, S. Huck, and C. Heyes. Automatic imitation in a strategic context: players of rock-paper-scissors imitate opponents' gestures. *Proceedings of the Royal Society B: Biological Sciences*, 279(1729):780–786, Feb. 2012.

[10] D. Cormier, G. Newman, M. Nakane, J. Young, E., and S. Durocher. In *n Proceedings of the First International Conference on Human-Agent Interaction, iHAI'13.*, iHAI'13, pages I–3–1. The Japanese Society of Artificial Intelligence, 2013.

[11] G. Dance and T. Jackson. Rock-paper-scissors: You vs. the computer. *The New York Times,* 2014.

[12] E. Fehr and S. Gaechter. Reciprocity and economics: The economic implications of homo reciprocans. *European Economic Review*, 42(3–5):845–859, May 1998.

[13] J. Fitzpatrick. *Resource Theory*, pages 1371–1372. SAGE Publications, Inc., 0 edition, 2009.

[14] B. J. Fogg and C. Nass. How Users Reciprocate to Computers: An Experiment That Demonstrates Behavior Change. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '97, pages 331–332, New York, NY, USA, 1997. ACM.

[15] E. Hoffman, K. A. McCabe, and V. L. Smith. Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology. *Economic Inquiry*, 36(3):335, 1998.

[16] G. Hoffman, J. Forlizzi, S. Ayal, A. Steinfeld, J. Antanitis, G. Hochman, E. Hochendoner, and J. Finkenaur. Robot presence and human honesty: Experimental evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 181–188, New York, NY, USA, 2015. ACM.

[17] P. Kahn, H. Ishiguro, B. Friedman, and T. Kanda. What is a human? - toward psychological benchmarks in the field of human-robot interaction. pages 364–371, 2006.

[18] P. H. Kahn, Jr., T. Kanda, H. Ishiguro, B. T. Gill, S. Shen, H. E. Gary, and J. H. Ruckert. Will people keep the secret of a humanoid robot?: Psychological intimacy in hri. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 173–180, New York, NY, USA, 2015. ACM.

[19] Y. Katsuki, Y. Yamakawa, and M. Ishikawa. High-speed human / robot hand interaction system. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, pages 117–118, New York, NY, USA, 2015. ACM.

[20] J. G. Lambsdorff and B. Frank. Bribing versus gift giving. an experiment. *Journal of Economic Psychology*, 31(3):347–357, June 2010.

[21] A. Litoiu, D. Ullman, J. Kim, and B. Scassellati. Evidence that robots trigger a cheating detector in humans. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 165–172, New York, NY, USA, 2015. ACM.

[22] L. D. MOLM. The structure of reciprocity. *Social Psychology Quarterly*, 73(2):119–131, 2010.

[23] S. Pinker, M. A. Nowak, and J. J. Lee. The logic of indirect speech. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3):833–838, Jan. 2008.

[24] S. M. Ross. Peirce's criterion for the elimination of suspect experimental data.

[25] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 141–148, New York, NY, USA, 2015. ACM.

[26] E. B. Sandoval, J. Brandstatter, M. Obaid, and C. Bartneck. Reciprocity in human robot interaction. a quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics*, 7(5), 2016.

[27] E. Short, J. Hart, M. Vu, and B. Scassellati. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 219–226, Mar. 2010.

[28] K. C. Williams. *Introduction to game theory: a behavioral approach*. Oxford University Press, New York, 2013.

[29] M. R. Zefferman. Direct reciprocity under uncertainty does not explain one-shot cooperation, but demonstrates the benefits of a norm psychology. *Evolution and Human Behavior*, 35(5):358 – 367, 2014.