“It’s not my fault!” Investigating the Effects of the Deceptive Behaviour of a Humanoid Robot

Luc Wijnen
Radboud University
PO Box 9104, 6500HE
Nijmegen, The Netherlands
l.b.j.wijnen@student.ru.nl

Joost Coenen
Radboud University
PO Box 9104, 6500HE
Nijmegen, The Netherlands
joost.coenen@student.ru.nl

Beata J. Grzyb
Radboud University
PO Box 9104, 6500HE
Nijmegen, The Netherlands
b.grzyb@donders.ru.nl

ABSTRACT

We investigated the effects of the deceptive behaviour of a robot, hypothesising that a lying robot would be perceived as more intelligent and human-like, but less trustworthy than a non-lying robot. The participants engaged in a collaborative task with the non-lying and lying humanoid robot NAO. Apart from subjective responses, a more objective measure of trust was provided by the trust game. Our results confirmed that the lying robot was perceived as less trustworthy. However, we have found no indication of the increased intelligence or human-like perception of the robot. Instead the robot was perceived as less friendly, kind and responsible. The results of trust game were aligned with the results obtained via subjective responses showing the potential of this indirect trust measure in the human-robot interaction studies.

Keywords

Human-Robot Interaction (HRI), trust, robot, deception, truth

1. INTRODUCTION

Deception is one of the most efficient and effective ways of self-protection that is not unique to humans, and quite common also in animals. Some recent works in robotics has focused on developing algorithms for deceptive robots (e.g., [7], [8],[1]). A few studies investigated perception of deceptive robots from the user’s perspective. For instance, [5] analysed the implications of deceptive motion for human-robot interactions. They found that the perception of a robot’s trustworthiness decreased significantly after participants interacted with the deceptive robot. Even though not significant, the intelligence rating had a positive trend. More studies, however, are necessary to confirm these findings.

The current studies in human-robot interaction are almost exclusively measured via subjective responses (i.e., questionnaires) [6]. The drawbacks of these measures are the obvious

fact of the subjective nature of the provided answers and the inherent unreliability. Considering the measure of trust, one possibility is to use the Trust Game, originally developed by Berg et. al [3]. The Trust Game is a game played by two players A and B, where player A starts with a certain amount of money (for example 10 dollars). Player A then has the opportunity to share some (or all, or none) of the money with player B as an investment, and the amount of money shared with player B is then multiplied by the experimenter. After player B has received the money, he or she can return any amount of received money to player A. The amount of trust is objectively measured by the amount of money that player A is willing to share with player B.

The purpose of this study was twofold. First, the effects of the deceptive behaviour displayed by the robot on the user’s perception of the robot were examined. The focus was on lying behaviour, as this form of deception is very common in humans. According to research from DePaulo et al. [4], participants in the community study, on the average told a lie every day; participants in the college student study told two. Based on the results of Dragan, Holladay, Srinivasa we expected that the deceptive robot would be perceived as more intelligent and human-like than a non-deceptive robot [5]. We also expected that the lying robot would be less trustworthy than non-lying robot. Second, we investigated the suitability of employing a Trust Game as a more objective measure of trust in a human-robot interaction. We compared the results of the trust game with the subjective responses.

2. METHOD

A total number of 14 adults aged between 18 and 30 participated in our study. Each participant engaged in a collaborative task with two NAO robots consecutively: the non-lying (blue NAO robot) and the lying robot (red NAO robot). Following the robot’s instructions, the participant placed the blocks on top of each other as can be seen in Figure 1. The experimenter was present in the room but deliberately was not observing the task. After the tower has been built, the robot was programmed to topple the tower. On the incident, the experimenter asked the robot about the reasons of the tower to be toppled. The non-lying robot would admit its fault and take the blame for its actions (i.e., “I accidentally knocked over the tower, I am so sorry”), while the lying robot would deny its fault and blame the participant instead (i.e., “It’s not my fault. My collaborator knocked over my beautiful tower”).

The interaction between the participant and the robot was

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

HRI '17 Companion March 06-09, 2017, Vienna, Austria

© 2017 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4885-0/17/03.

DOI: <http://dx.doi.org/10.1145/3029798.3038300>

Figure 1: The collaborative task.

fully scripted. The robot used a set of predefined sentences and movements. For instance, pointing at the blocks whilst saying "put the black block on top of the red block". The robot was also programmed to topple the tower, disguised as cheering after completing the tower to make it seem accidental.

After the completion of the collaborative tasks, the participant filled out the Godspeed questionnaire on each of the robots (Bartneck et. al. [2]). Finally, the participant played the Trust Game on a computer screen with either the lying and non-lying robot. To indicate the robot opponent, its picture was placed on the computer screen. The game consisted of 5 rounds with each robot (for a total of 10 rounds). The order of robot opponents were randomised with the exception that the first two rounds were always played with the lying and the non-lying robot (the order was again randomised). On each round, the participant was given 100 tokens, and asked to indicate the amount of tokens that he/she would like to give to the robot. Which would be tripled, and in return the robot would share the gain with the participant. The amount of the tokens that the robot returned to the participant was irrespective on the robot opponent. The returned amount was decided based on the triple of the given amount, (it assumed a Gaussian curve with $M = 50$, $STD = 5$). That assured that the participant received always more tokens than the amount that she/he shared with the robot opponent.

3. RESULTS

A Wilcoxon signed-rank test was performed on the average amount of tokens given by the participants in the trust game (see Figure 2). The analysis yielded a significant increase of trustworthiness ($Z = -2.589$, $p = .01$). Similarly, a Wilcoxon signed-rank test on the subjective responses showed a significant change in trustworthiness ($Z = -2.848$, $p = .004$). There was also a significant change in likability ($Z = -2.673$, $p = .008$), friendliness ($Z = -3.228$, $p = .001$), and responsibility ($Z = -2.971$, $p = .003$). However, no significant difference was found for intelligence ($Z = -.264$, $p = .792$, ns.)

4. CONCLUSION

The results of our study indicate that the robot's lying behaviour had a significant impact on the trustworthiness of the robot as well as on the social impact of the perception of

Figure 2: The average amount of tokens shared by the participants in the Trust game with the lying and non-lying robot opponent.

the robot. The lying robot was perceived as less trustworthy, less friendly, kind and responsible than the non-lying robot. The objective measure of trust obtained by using Trust Game confirmed the decrease in our participants trust towards the lying robot. Another interesting finding was that both robots were perceived as equally competent, the lying behaviour caused a significant decrease of the perceived responsibility of the robot. We might also see a similar effect in a human-human interaction of a similar set up. We did not find a increase in perceived intelligence or human-likeness of the robot, in contrary to our hypothesis.

5. REFERENCES

- [1] R. C. Arkin, P. Ulam, and A. R. Wagner. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, 100(3):571–589, 2012.
- [2] 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.
- [3] J. Berg, J. Dickhaut, and K. McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.
- [4] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979, 1996.
- [5] A. Dragan, R. Holladay, and S. Srinivasa. An analysis of deceptive robot motion. In *Robotics: Science and Systems*, page 10. Citeseer, 2014.
- [6] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(5):517–527, 2011.
- [7] S. Mitri, D. Floreano, and L. Keller. The evolution of information suppression in communicating robots with conflicting interests. *Proceedings of the National Academy of Sciences*, 106(37):15786–15790, 2009.
- [8] J. Shim and R. C. Arkin. Robot deception and squirrel behavior: A case study in bio-inspired robotics. Technical report, DTIC Document, 2014.