ERROR: Evaluating tRust weaRing Off in Robots

ALESSANDRA ROSSI, RAFFAELLA ESPOSITO, DAVIDE MAROCCO, and SILVIA ROSSI,

University of Naples Federico II, Italy

In this work, we introduce the Evaluating tRust weaRing Off in Robots (ERROR) project. This ERROR project aims to address the lack of users' trust in robots, starting from the need for more persuading and personalized robotic mechanisms that can favour people's behavioural changes and compliance with instructions in order to improve their health, social and work lives. We aim to investigate how to deploy trustworthy and transparent robot behaviours in these contexts with particular attention to the techniques for mitigating peoples' trust after a loss of trust whether this was intentional (i.e., deceptive behaviours) or unintentional (i.e., erroneous and unexpected behaviours) for a balanced trustworthy and successful long-lasting interaction with people. We present our initial works in this direction, where participants (n. 37) played an assistive game with a deceptive robot endowed with Theory of Mind (ToM). We observed that a deceptive robot was less trusted by the participants, even though not all participants recognised the intentionality of the robot to deceive them.

Additional Key Words and Phrases: Trust, Social Robotics, Deception

1 INTRODUCTION

In the ever-evolving landscape of robotics, a fundamental prerequisite for successful integration of robots into assistive applications is their ability to facilitate behavioural changes and encourage compliance with instructions, especially in medical and health contexts [13]. Beyond these domains, the extent to which individuals follow guidance from robots has broad implications for service robotics and emergency scenarios [18]. Researchers are exploring ways to imbue robots with human-like social cues, personalities, and cognitive capabilities to foster compliance and trust in human-robot interactions [13, 15]. In particular, trust is a critical factor in human-robot interaction, influenced by perceptions of a robot's reliability in performing its functions [16]. Trust is also connected to the willingness to take calculated risks when the outcomes are uncertain [8], and the belief that the robot can assist individuals in achieving their goals in situations characterized by vulnerability and uncertainty [12]. Additionally, trust may be intertwined with the emotional connection between humans and robots [14, 17]. Factors influencing trust include personal differences, task utility, prior experiences, reliability, and the nature of the relationship with the robot [10, 19, 21]. The integration of robots into society, mimicking human behaviour and cognitive capabilities, in contrast, also raises concerns about over-trust in these machines. Both under- and over-trust can hinder technology adoption, necessitating mechanisms for balanced trust [19].

Another technique explored to ensure compliance with robot requests and advice is deception in human-robot interaction. Controlled deception can prevent conflicts, reduce emotional distress, and enhance working relationships in contexts like education, healthcare, and rescue operations [1, 3, 6, 9, 22–24]. At the same time, however, deception is also a topic of philosophical and psychological controversy [2]. Different forms of deception, from white lies to tactical and intentional deception, may influence trust in diverse ways. Deceptive behaviours, whether intentional or unintentional, can lead to misunderstandings, negative attributions, and a loss of confidence in robots, potentially eroding trust and discouraging their use. A mismatch between user expectations and robot behaviour can alter the

This work has been supported by Italian PON R&I 2014-2020 - REACT-EU Azione IV.4 (CUP E65F21002920003, and by the US Air Force Office of Scientific Research under award number FA8655-23-1-7060.

perception of trust, jeopardizing the success of the interaction [11]. To this extent, it is essential to investigate both the mechanisms for recalibrating trust and preventive measures to avoid trust erosion. Ethical considerations surrounding the use of deception for behavioural change in human-robot interactions cannot be overlooked. Transparency, consent, and respect for user autonomy are paramount in responsibly employing deceptive mechanisms [7].

The ERROR project aims to provide a comprehensive assessment of deceptive mechanisms' impact on trust in autonomous robots, emphasizing responsible use and alignment with existing regulations for the safe development of autonomous agents in human-centred environments [7]. To this extent, we started investigating whether the ability of mentalizing - which can be defined as a multi-modal system that allows people to naturally communicate and understand each other by inferring others' intentions, desires, and beliefs [4] - may be used by a robot to foster trust and mitigate the potential issues connected to deception.

2 APPROACH

As a first step, we investigated how people's perception of trust in a robot with the ability to mentalize varies when it has deceptive behaviours. Specifically, we decided to use verbal deception (i.e., lies), instead of non-verbal deception, and the robot's perspective and people's perception of the deception. In this initial work, we decided to focus on the definition of lies as "a false statement made by an individual which knows that the statement is not true" [5], but not to take advantage of the lie. We chose an assistive gaming scenario (i.e., Memory Game[1]), in which the robot does not compete against humans, and it used a Q-learning-based approach to provide advice relying on people's beliefs and their intended game strategies. The robot's ToM was shown to the players by generating move suggestions (e.g., the column or position of the first or the second card) based on the robot's knowledge of the game. The robot keeps track of the cards discovered by the player, the frequency and during which turn they were flipped. This info is used by the robot to generate a suggestion based on the current state of the game, possible beliefs and intentions of the participant. Participants were assigned to one of the two following conditions: 1) without a deceiving behaviour (**ND-ToM** condition) in which the robot provided assistance in the game by generating suggestions; and 2) with a deceiving behaviour (**D-ToM** condition) in which the robot suggested the wrong cards to the player. In order to not create an entirely faulty robot that would have never gained trust [20], the robot provided only 20% of wrong suggestions.

3 RESULTS

We recruited 37 people, aged between 18 and 59 years old (avg. 29, std. 11), and they identified themselves as female (43.3%) and male (56.7%). The majority of participants had no previous or close experience with robots. Participants were distributed as 20 participants in the ND-ToM condition, and 17 participants in the D-ToM condition.

We observed that the increase in wrong suggestions negatively affected people's trust in the robot, as people did not accept the robot's suggestion to choose a card ($\tau_b(37) = -0.483, p < 0.001$). We also asked participants to state whether they relied on the robot and had faith that the robot is able to succeed in performing even in situations in which it is untried. Participants had higher trust in the robot's reliability and capabilities in the ND-ToM condition compared to those in D-ToM condition (respectively, $t(35) = 2.701, p = 0.011$, and $t(35) = 2.071, p = 0.046$).

[1] Open source GIT repository of the game https://github.com/yunkii/animal-memory-game

4 CONCLUSIONS & FUTURE WORKS

Our first step has been to investigate whether and how people's trust in a deceptive robot vary when they share the same awareness of the situational context, and the robot can mentalize people. Our next step is to identify different types of deception that an autonomous embodied agent, such as a robot, may provide while interacting with a human being, and evaluate how these types of deception affect a loss of people's trust in robots based on people's exposure to deceiving behaviours in tasks with different criticality.

REFERENCES

[1] Eytan Adar, Desney Tan, and Jaime Teevan. 2013. Benevolent deception in human computer interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 1863–1872. https://doi.org/10.1145/2470654.2466246

[2] Gabriella Airenti. 2015. The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy. *International Journal of Social Robotics* 7 (02 2015). https://doi.org/10.1007/s12369-014-0263-x

[3] Ronald Arkin, Patrick Ulam, and Alan Wagner. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proc. IEEE* 100 (03 2012), 571–589. https://doi.org/10.1109/JPROC.2011.2173265

[4] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition* 21, 1 (1985), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

[5] Thomas L. Carson. 2006. The Definition of Lying. *Noûs* 40, 2 (2006), 284–306. http://www.jstor.org/stable/3506133

[6] Mark Coeckelbergh. 2018. How to Describe and Evaluate Deception Phenomena: Recasting the Metaphysics, Ethics, and Politics of Icts in Terms of Magic and Performance and Taking a Relational and Narrative Turn. *Ethics and Information Technology* 20, 2 (2018), 71–85.

[7] John Danaher. 2020. Robot Betrayal: A Guide to the Ethics of Robotic Deception. *Ethics and Inf. Technol.* 22, 2 (jun 2020), 117–128. https://doi.org/10.1007/s10676-019-09520-3

[8] Morton Deutsch. 1958. Trust and suspicion. *Journal of Conflict Resolution* 2, 4 (1958), 265–279.

[9] Sanjiv Erat and Uri Gneezy. 2012. White Lies. *Management Science* 58, 4 (2012), 723–733.

[10] Peter Hancock, Deborah Billings, Kristin Schaefer, Jessie Chen, Ewart de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human factors* 53 (10 2011), 517–27. https://doi.org/10.1177/0018720811417254

[11] Markus Https://Orcidorg Kneer. 2021. Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents. *Cognitive Science* 45, 10 (2021), e13032. https://doi.org/10.1111/cogs.13032

[12] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 arXiv:https://doi.org/10.1518/hfes.46.1.50$_3$0392 PMID: 15151155.

[13] Namyeon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. 2017. The Influence of Politeness Behavior on User Compliance with Social Robots in a Healthcare Service Setting. *International Journal of Social Robotics* 9 (11 2017). https://doi.org/10.1007/s12369-017-0420-0

[14] J. David Lewis and Andrew Weigert. 1985. Trust as a Social Reality. *Social Forces* 63, 4 (06 1985), 967–985. https://doi.org/10.1093/sf/63.4.967

[15] G. Maggi, E. Dell'Aquila, I. Cucciniello, and S. Rossi. 2021. "Don't Get Distracted!": The Role of Social Robots' Interaction Style on Users' Cognitive Performance, Acceptance, and Non-Compliant Behavior. *International Journal of Social Robotics* 13 (2021). https://doi.org/10.1007/s12369-020-00702-4

[16] R. C. Mayer, J. H. Davis, and F.D. Schoorman. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review* 20(3) (1995).

[17] Daniel McAllister. 1995. Affect- and Cognition-Based Trust Formations for Interpersonal Cooperation in Organizations. *Academy of Management Journal* 38 (02 1995), 24–59. https://doi.org/10.2307/256727

[18] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. 2017. Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436. https://doi.org/10.1109/THMS.2017.2648849

[19] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *Social Robotics*, Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He (Eds.). Springer International Publishing, Cham, 42–52.

[20] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *Social Robotics*, Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He (Eds.). Springer International Publishing, Cham, 42–52.

[21] Alessandra Rossi, Patrick Holthaus, Kerstin Dautenhahn, Kheng Koay, and Michael Walters. 2018. Getting to know Pepper: Effects of people's awareness of a robot's capabilities on their trust in the robot. 246–252. https://doi.org/10.1145/3284432.3284464

[22] Jaeeun Shim and Ronald Arkin. 2015. The benefits of robot deception in search and rescue: Computational approach for deceptive action selection via case-based reasoning. 1–8. https://doi.org/10.1109/SSRR.2015.7443002

[23] Jaeeun Shim and Ronald C. Arkin. 2013. A Taxonomy of Robot Deception and Its Benefits in HRI. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*. 2328–2335. https://doi.org/10.1109/SMC.2013.398

[24] Henrik Skaug Sætra. 2021. Social robot deception and the culture of trust. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 276–286. https://doi.org/doi:10.1515/pjbr-2021-0021