Trust in deceptive robots

Raffaella Esposito[1,*,†], Alessandra Rossi[1,†], Michela Ponticorvo[1,†] and Silvia Rossi[1,†]

[1]University of Naples Federico II, Via Claudio 21, Naples, Italy

Abstract

In this work, we explore the dynamics of trust and deception in human-robot interaction (HRI). Trust is crucial for successful interactions between humans and robots, especially when users are required to rely on robotic systems for assistive tasks. While often viewed negatively, certain forms of deception, such as prosocial deception which could take form of white lies or emotional expressions, can paradoxically enhance trust. Our aims are to investigate how deceptive practices affect trust, to set a starting point or further research that delves into the development, deterioration, and possible restoration of trust following deceptive interactions in robotics. In particular, here we start by drawing on findings from existing literature and experimental work in human-robot interaction.

Keywords

Trust, Social Robotics, Human-Robot Interaction, Deception

1. Introduction

The effectiveness of a robot in an assistive capacity depends on whether individuals follow the guidance or instructions it offers [1]. Different methodologies exist for favoring the human compliance with the robot's suggestions. One prerequisite is trust, as usage of automation by humans is strongly influenced by their trust in the robot [2]. At the same time, deception is also considered a necessity in situations like rehabilitation and rescue operations, preventing emotional distress and possible conflicts, and in these cases it is considered prosocial [3, 4]. However, deception is an ethical dilemma, since it can be harmful to human's trust in the robot and the human-robot interaction [4]. Although a loss of trust may be a consequence of certain types of deception, actually trust can be obtained through deception, especially where interpersonal relationships and emotional support are involved [5]. Thus, it is crucial to understand the effects of deception on trust in robots, and how trust can be established and restored in cases in which a robot's deception has caused a decrease in trust. This might be the case of lies or intentional errors [6]. In this contribution, we present traditional concepts of trust in HRI, and which are the effects of the interaction with a deceptive robot on people's trust.

2. The concept of trust in human-human relationships

Trust is a complex feeling, and several definitions exist. For example, Schilke's [7] defines trust as the willingness of one party (the trustor) to expose themselves to vulnerability in relation to another party (the trustee). In doing so, the trustor takes a risk, believing that the trustee will act in ways that are beneficial to the trustor, despite these actions being outside the trustor's control.

The BDT (Basis-Domain-Target) framework [8] specifies that trust comprises three bases: reliability, that is the consistent fulfilment of commitments; emotional trust as the assurance that one will not cause emotional harm to another; honesty, involving straightforwardness and truthfulness in one's

MULTITRUST workshop at the third International Conference on Hybrid Human-Artificial Intelligence - HHAI '24, June 10–14, 2024, Malmo, Sweden

*Corresponding author.

†These authors contributed equally.

raffaella.esposito3@unina.it (R. Esposito); alessandra.rossi@unina.it (A. Rossi); michela.ponticorvo@unina.it (M. Ponticorvo); silvia.rossi@unina.it (S. Rossi)

0009-0002-6582-0002 (R. Esposito); 0000-0003-1362-8799 (A. Rossi); 0000-0003-2451-9539 (M. Ponticorvo); 0000-0002-3379-1756 (S. Rossi)

actions and statements. Individuals can rely on others to demonstrate the three bases of trust from a cognitive, affective and behavioral point of view, and the target dimensions of trust are familiarity and specificity.

Jones defines trust as properly an attitude [9] which is characterized by optimism about another's goodwill and competence, combined with the expectation that the trusted person will be positively motivated by the act of being trusted. Trust is more than just a rational assessment: it involves a positive emotional tilt toward the trusted, that leads to a favorable interpretation of actions that might otherwise be ambiguous or even suspect. By trusting others, people are more likely to engage in cooperative behaviors, share resources, and support one another in mutual goals.

3. Towards trusting human-robot relationships

Lee and See [10] defined trust in automation as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. Wagner et al. also highlight the dimension of risk in trusting someone [11].

The three bases of trust [8] could be also applied to HRI. Indeed, reliability in robots could translate to consistent performance across diverse and potentially challenging environments. Secondly, robots expressing concern about human's emotional safety could help in establishing trust towards robots. However, robots displaying emotions are considered deceptive, as they show characteristics that they don't actually possess [12]. Finally, honesty in robots could involve straightforwardness about their functionalities and limitations.

The emotional component of trust [9] could be also included in evaluating trust towards robots, as the attribution to the robot of goodwill, competence, and motivation to act in a trustworthy manner.

4. Deception's role in human-robot trust dynamics

Existing research highlights that deception can be particularly detrimental to trust. However, some forms of deception can increase trust as firstly shown by Levine and Schweitzer [5]. In fact, it is fundamental to distinguish between the different types of deception. We can categorize deception into two main types: hiding the truth or showing the false [13].

In the context of robotic deception, Shim and Arkin developed a taxonomy consisting of eight types of robotic deception. This taxonomy is based on three dimensions: the deceived party (human or non-human), the deception's goal (self-oriented or other-oriented), and the method of deception (physical or mental) [3], acknowledging the existence of a kind of robotic deception specifically intended to benefit to the user.

In HRI, trust in robots can diminish following exposure to deceptive motion paths [14, 15] and after discovering a robot's deceit in a driving simulation [16]. The robot's inability of taking ownership of its faults and attempt of blaming someone else (i.e., a human) has been showed to decrease notably people's trust in the robot [17]. Moreover, a robot that gives incorrect suggestions during a memory card game prevents people to accept future suggestions, and negatively affect their perception of robot's reliability and the faith that the robot is able to succeed [18].

In the context of a multiplayer game (whack-a-mole), in which the robot subtly balanced wins among participants, trust was affected by the order in which participants experienced the robot's modes of announcing wins (honest and balancing): trust increased when participants experienced the honest mode after the deceptive mode and decreased when the sequence was reversed [19].

While considering deception in HRI, two theories about deception are of particular interest. These are: Saetra's belief that displaying emotions, which it is also referred to as anthropomorphism, is a type of deception [4]; and Danaher's distinction of deception in external, superficial and hidden state deception [12].

5. Conclusions & Future Works

The aim of our research is to deepen the understanding of how trust in robots is built, eroded and recovered after a loss, and the role of deception in these dynamics. To this extent, we started by expanding upon the studies conducted so far.

By establishing the different types of deception, we intend to explore the role of prosocial deception in trust establishment and maintenance. Further steps will include an investigation of the use of human social cues and empathetic responses for maintaining trust and facilitating its recovery after breaches caused by deceptive behaviors. Indeed, the ability of a robot to show emotions can garner higher goodwill ratings compared to the robot merely expressing preferences [20]. Moreover, we aim at using a robot endowed with Theory of Mind and ability to use human mechanisms, such as by providing justifications that refer to mental states for a robot's actions, to mitigate the loss of moral trust in a robot, even amidst disagreement with its moral decisions [21].

Research shows that a robot's ability to express emotions effectively is crucial for engaging, trustworthy, and productive interactions with humans. Emotionally expressive robots are perceived as autonomous social agents, which enhances familiarity and comfort. Positive emotional expressions from robots can also boost human performance by creating a supportive and motivating environment [22].

References

[1] P. Robinette, A. M. Howard, A. R. Wagner, Effect of robot performance on human–robot trust in time-critical situations, IEEE Transactions on Human-Machine Systems 47 (2017) 425–436. doi:10.1109/THMS.2017.2648849.

[2] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human Factors 39 (1997) 230–253. doi:10.1518/001872097778543886.

[3] J. Shim, R. C. Arkin, A taxonomy of robot deception and its benefits in hri, in: 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 2328–2335. doi:10.1109/SMC.2013.398.

[4] H. S. Sætra, Social robot deception and the culture of trust, Paladyn, Journal of Behavioral Robotics 12 (2021) 276–286. URL: https://doi.org/10.1515/pjbr-2021-0021. doi:doi:10.1515/pjbr-2021-0021.

[5] E. E. Levine, M. E. Schweitzer, Prosocial lies: When deception breeds trust, Organizational Behavior and Human Decision Processes 126 (2015) 88–106. URL: https://www.sciencedirect.com/science/article/pii/S0749597814000983. doi:https://doi.org/10.1016/j.obhdp.2014.10.007.

[6] S. Daronnat, L. Azzopardi, M. Halvey, Impact of agents' errors on performance, reliance and trust in human-agent collaboration, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 64, SAGE Publications Sage CA: Los Angeles, CA, 2020, pp. 405–409.

[7] Schilke, Trust in social relations, Annual Review of Sociology 47 (2021) 239–259. doi:10.1146/annurev-soc-082120-082850.

[8] K. Rotenberg, The conceptualization of interpersonal trust: A basis, domain, and target framework, Interpersonal Trust During Childhood and Adolescence (2010) 8–27. doi:10.1017/CBO9780511750946.002.

[9] K. Jones, Trust as an Affective Attitude, 2005, pp. 253–279. doi:10.1007/978-0-230-20409-6_11.

[10] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (2004) 50–80.

[11] A. R. Wagner, R. C. Arkin, Recognizing situations that demand trust, in: 2011 RO-MAN, IEEE, 2011, pp. 7–14.

[12] J. Danaher, Robot betrayal: A guide to the ethics of robotic deception, Ethics and Inf. Technol. 22 (2020) 117–128. URL: https://doi.org/10.1007/s10676-019-09520-3. doi:10.1007/s10676-019-09520-3.

[13] J. Bell, B. Whaley, Cheating and deception, 1991.

[14] A. Dragan, R. Holladay, S. Srinivasa, Deceptive robot motion: synthesis, analysis and experiments, AUTONOMOUS ROBOTS 39 (2015) 331–345. doi:10.1007/s10514-015-9458-8, 10th Conference on Robotics - Science and Systems (RSS), Univ Calif, Berkeley, CA, JUN, 2014.

[15] A. Ayub, A. Morales, A. Banerjee, Using markov decision process to model deception for robotic and interactive game applications, in: 2021 IEEE INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS (ICCE), International Conference on Consumer Electronics, IEEE, 2021. doi:10.1109/ICCE50685.2021.9427633, iEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, JAN 10-12, 2021.

[16] K. Rogers, R. J. A. Webber, A. Howard, Lying about lying: Examining trust repair strategies after robot deception in a high-stakes hri scenario, in: COMPANION OF THE ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION, HRI 2023, IEEE; Assoc Comp Machinery; Honda Res Inst Japan; Toyota Res Inst; Amazon; Furhat Robot; LuxAI; Diligent Robot; Navel Robot; Google; PAL Robot; Sci Robot; IEEE Robot & Automat Soc; SIGCHI; ACM SIGCAI, 2023, pp. 706–710. doi:10.1145/3568294.3580178, 18th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), Stockholm, SWEDEN, MAR 13-16, 2023.

[17] L. Wijnen, J. Coenen, B. J. Grzyb, "it's not my fault!" investigating the effects of the deceptive behaviour of a humanoid robot, in: COMPANION OF THE 2017 ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION (HRI'17), ACM IEEE International Conference on Human-Robot Interaction, Assoc Comp Machinery; IEEE; ACM SIGCHI; ACM SIGAI; IEEE Robot & Automat Soc; AAAI; HFES, 2017, pp. 321–322. doi:10.1145/3029798.3038300, 12th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), Vienna, AUSTRIA, MAR 06-09, 2017.

[18] A. Rossi, S. Rossi, Evaluating people's perception of trust of a deceptive robot with theory of mind in an assistive gaming scenario, in: 2023 32ND IEEE INTERNATIONAL CONFERENCE ON ROBOT AND HUMAN INTERACTIVE COMMUNICATION, RO-MAN, IEEE RO-MAN, IEEE, 2023, pp. 1375–1380. doi:10.1109/RO-MAN57019.2023.10309647, 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Busan, SOUTH KOREA, AUG 28-31, 2023.

[19] M. Vazquez, A. May, A. Steinfeld, W.-H. Chen, A deceptive robot referee in a multiplayer gaming environment, in: Proceedings of International Conference on Collaboration Technologies and Systems (CTS '11), 2011, pp. 204 – 211.

[20] K. Winkle, P. Caleb-Solly, U. Leonards, A. Turton, P. Bremner, Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots, in: 2021 16TH ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION, HRI, IEEE; Assoc Comp Machinery; ACM SIGCHI; ACM SIGAI; IEEE Robot & Automat Soc, 2021, pp. 101–109. doi:10.1145/3434073.3444666, 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boulder, CO, MAR 09-11, 2021.

[21] A. Rosero, "using justifications to mitigate loss in human trust when robots perform norm - violating and deceptive behaviors, in: COMPANION OF THE ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION, HRI 2023, IEEE; Assoc Comp Machinery; Honda Res Inst Japan; Toyota Res Inst; Amazon; Furhat Robot; LuxAI; Diligent Robot; Navel Robot; Google; PAL Robot; Sci Robot; IEEE Robot & Automat Soc; SIGCHI; ACM SIGCAI, 2023, pp. 766–768. doi:10.1145/3568294.3579979, 18th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), Stockholm, SWEDEN, MAR 13-16, 2023.

[22] R. Stock-Homburg, Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research, International Journal of Social Robotics 14 (2021). doi:10.1007/s12369-021-00778-6.