Computational Trust in Robotics: Preliminary Investigations and Evidence

Francesco Semeraro[1,*,†], Marta Romeo[2,†], Angelo Cangelosi[1] and Samuele Vinanzi[3,†]

[1]Manchester Centre for Robotics and AI, The University of Manchester, Manchester, UK
[2]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK
[3]Department of Computing, Sheffield Hallam University, Sheffield, UK

Abstract

Trust plays a crucial role in the design of human-robot interaction. Most of the current research focuses on the human factors that affect trust towards the robot, while only few works mathematically model the trust a robot can have towards the user and viceversa. In this work, we term this line of research as "Computational Trust" and provide empirical evidence of this trend through preliminary results from an ongoing systematic review.

Keywords

Computational Trust, Artificial Trust, Natural Trust, Human-Robot Interaction, Human-Robot Collaboration, Social Robotics, Artificial Intelligence

1. Introduction and background

Trust is essential in shaping human-human relationships [1]. As autonomous agents are starting to enter our everyday environments, we see an increasing number of researchers in Human-Robot Interaction (HRI) directing their efforts towards understanding how this factor influences the relationships between humans and intelligent machines [2]. For example, in a collaborative setting between humans and robots, establishing a trust relationship enables the user to delegate a portion of the shared task to the robot [3, 4, 5]. As a result, the user can concentrate on their own responsibilities within the task, thus enhancing the overall outcome [6]. Trust in HRI has been defined as the "attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" [7], and it is mainly studied from the point of view of the human. In fact, most efforts have been directed towards understanding what robotic characteristics facilitate or hinder human partners' trust [8, 9], or what strategies lead to trust repair once the latter is lost [10]. Research on trust and trustworthiness in automation has become even more critical with the integration of Artificial Intelligence (AI) in the decision-making processes of autonomous agents.

Within this research panorama, we are witnessing an increasing interest in trying to mathematically model human trust towards robots in an attempt to understand and exploit the human partner's internal state [2]. We term this modeling attempt as "Natural Trust". The focus of the latter is understanding the human partner's internal model of the robot to facilitate the establishment of trust. In a dual fashion, few mathematical models address the trust that an artificial agent could have towards the user it is interacting with, named "Artificial Trust" [11, 12]. Both these forms of trust are rarely utilized during a human-robot interaction to alter the behavioral policy of the artificial agent.

MultiTTrust: 3rd Workshop on Multidisciplinary Perspectives on Human-AI Team Trust, June 11, 2024, Malmö, Sweden
*Corresponding author.
†These authors contributed equally.
francesco.semeraro@manchester.ac.uk (F. Semeraro); m.romeo@hw.ac.uk (M. Romeo); angelo.cangelosi@manchester.ac.uk (A. Cangelosi); s.vinanzi@shu.ac.uk (S. Vinanzi)
https://research.manchester.ac.uk/en/persons/francesco.semeraro (F. Semeraro); https://www.edinburgh-robotics.org/academics/marta-romeo (M. Romeo); https://research.manchester.ac.uk/en/persons/angelo.cangelosi (A. Cangelosi); https://www.shu.ac.uk/about-us/our-people/staff-profiles/samuele-vinanzi (S. Vinanzi)
0000-0002-8812-0968 (F. Semeraro); 0000-0003-4438-0255 (M. Romeo); 0000-0002-4709-2243 (A. Cangelosi); 0000-0003-0241-9983 (S. Vinanzi)

From these considerations, a gap in the current state-of-the-art emerges: robots are often considered as passive receptacles of trust, and not as active social agents that makes use of such trust to improve their own behaviour. Starting from the consideration that trust is a bidirectional relationship needed to successfully complete collaborative tasks [13], we are interested in investigating when and how it is possible for robots to model the trustworthiness of their human partners [14] and/or the trust put by the users in them, to allow them to exploit this knowledge and enhance their interactions with the users. Attempts at regularizing the design of models of trust for robots towards their users or other agents are still scarce, yet of great importance. Research in the current literature [15, 14] has shown that robots that possess cognitive mechanisms to identify and anticipate mistakes or deceptions in their human partner's strategy (in other words, evaluating their trustworthiness) can increase the success rate of joint collaborative tasks.

In an attempt to address this knowledge gap, we propose the term "Computational Trust" (CT) to refer to the mathematical models that can be used by a robot or an artificial agent to perform trust evaluations on other agents. This term incorporates both cases of Artificial Trust and Natural Trust. We discuss initial results from a systematic review we are currently finalizing. This work aims to be the first systematic dive into this new research domain.

2. Discussion and preliminary results

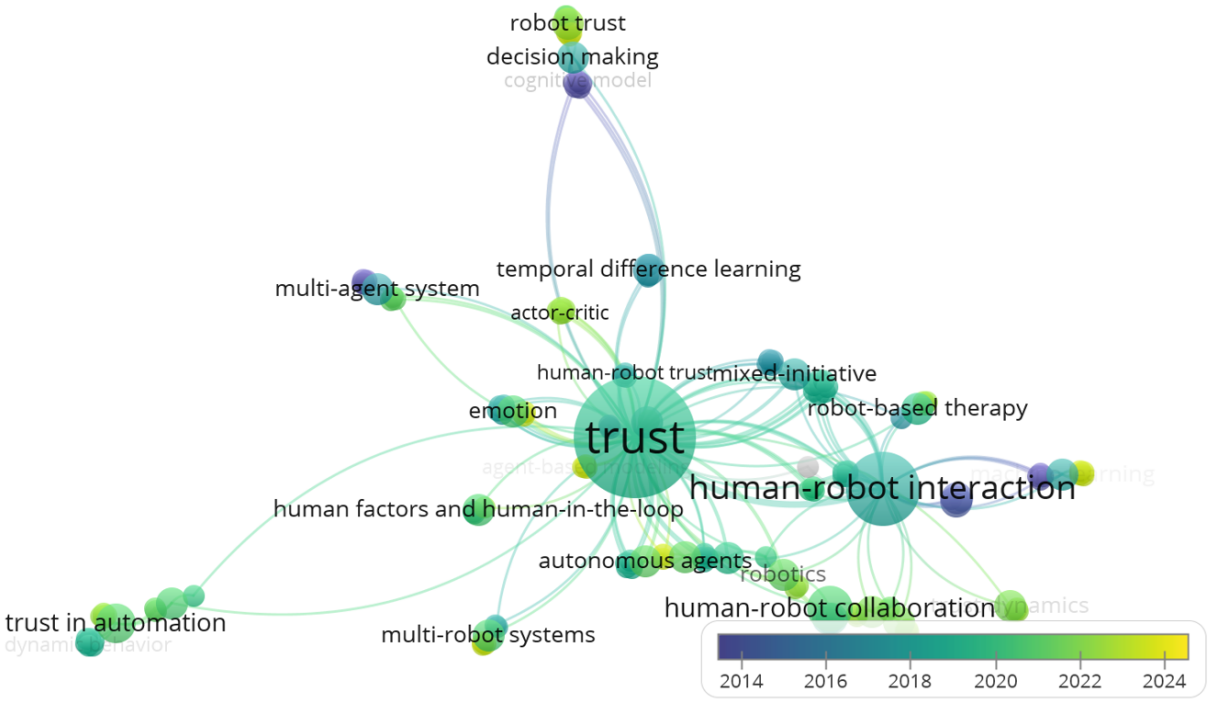

Figure 1: Keyword network of the selected set of papers (generated through VOSviewer [16]).

From the analysis of the state-of-the-art in trust in HRI detailed in Section 1, we have identified a gap in the current research landscape. Specifically, many works presented in the literature focus on the human side of robotic trust, i.e., the trust that humans place in robots. As we have seen, this is an important issue to consider during the worldwide effort to integrate social robots into our daily lives and environments. Despite this, we argue that this does not draw a complete picture of trust relationships between humanity and automation. In fact, there is evidence that trust is a bidirectional relationship and that the dynamics of mutual trust should not be ignored [13].

To address this gap in the current knowledge, we have performed a systematic review of the current

literature, searching for scientific publications that describe computational models of CT in HRI. Our investigation across three major scientific databases (IEEE Xplore, Scopus, and Web of Science) has led to the selection of 101 papers. By analysing them, we have generated a map of the co-occurrence of their keywords, reported in Figure 1. In its upper branch, the keyword "*robot trust*" appears, which is very closely linked to CT. Not only it appears in very recent publications, but it derives from topics such as "*decision making*" and "*cognitive model*". This is evidence of a recent trend in the literature to embed trust within robotic agents. However, this term is ambiguous, as it is mainly used to depict the classical perspective of human trust in robots. To avoid any confusion, we should instead refer to the more objective term of CT, which unambiguously refers to any way of mathematically modeling trust estimates in HRI. Furthermore, the keyword "*human-robot collaboration*" is closely linked to "*trust*".

Finally, it is worth noting that, since we looked for CT models, it emerges that these models are increasingly being used to modulate the behaviour of robots during collaborative tasks with humans. All this evidence underscores the importance of pursuing standards in designing CT models.

Acknowledgments

Francesco Semeraro's work was supported by the UKRI DTP CASE-conversion "Human-Robot Collaboration for Flexible Manufacturing" (Ref. 2480772), sponsored by UKRI Engineering and Physical Sciences Research Council and BAE Systems plc.

Marta Romeo's work was supported by the UKRI Node on Trust (Ref. EP/V026682/1, https://trust.tas.ac.uk).

Angelo Cangelosi's work was partially supported by the Horizon projects PRIMI, MUSAE and the ERC Advanced eTALK (funded by UKRI) and the UKRI Trustworthy Autonomous Systems Node on Trust (Ref. EP/V026682/1).

Samuele Vinanzi's work was partially supported by Sheffield Hallam University's Early Career Research and Innovation Fellowship. This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USA. Funder award Ref. FA9550-19-1-7002. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

References

[1] D. M. Rousseau, S. B. Sitkin, R. S. Burt, C. F. Camerer, Not so different after all: A cross-discipline view of trust, Academy of management review 23 (1998) 393–404.

[2] Y. Wang, F. Li, H. Zheng, L. Jiang, M. F. Mahani, Z. Liao, Human trust in robots: A survey on trust models and their controls/robotics applications, IEEE Open Journal of Control Systems 3 (2023) 58–86.

[3] F. Semeraro, J. Carberry, J. Leadbetter, A. Cangelosi, Good things come in threes: The impact of robot responsiveness on workload and trust in multi-user human-robot collaboration, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024.

[4] F. Semeraro, J. Carberry, A. Cangelosi, Simpler rather than challenging: Design of non-dyadic human-robot collaboration to mediate human-human concurrent tasks, in: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2023, p. 2541–2543.

[5] F. Semeraro, J. Carberry, A. Cangelosi, Towards multi-user activity recognition through facilitated training data and deep learning for human-robot collaboration applications, in: 2023 International Joint Conference on Neural Networks (IJCNN), 2023, pp. 01–09. doi:10.1109/IJCNN54540.2023.10191782.

[6] M. G. Collins, I. Juvina, K. A. Gluck, Cognitive model of trust dynamics predicts human behavior

within and between two games of strategic interaction with computerized confederate agents, Frontiers in Psychology 7 (2016) 1–17.
[7] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80. doi:10.1518/hfes.46.1.50/_30392.
[8] P. A. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, J. L. Szalma, Evolving trust in robots: Specification through sequential and comparative meta-analyses, Human Factors 63 (2021) 1196–1229.
[9] M. Romeo, I. Torre, S. L. Maguer, A. Cangelosi, I. Leite, Putting robots in context: Challenging the influence of voice and empathic behaviour on trust, in: 32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN, 2023.
[10] S. S. Sebo, P. Krishnamurthi, B. Scassellati, "I don't believe you": Investigating the effects of robot trust violation and repair, in: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Association for Computing Machinery, 2019, pp. 57–65. doi:10.1109/HRI.2019.8673169.
[11] H. Azevedo-Sa, X. J. Yang, L. P. Robert, D. M. Tilbury, A unified bi-directional model for natural and artificial trust in human-robot collaboration, Ieee Robotics and Automation Letters 6 (2021) 5913–5920. doi:10.1109/lra.2021.3088082.
[12] C. C. Jorge, M. L. Tielman, C. M. Jonker, Artificial trust as a tool in human-ai teams, in: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2022, pp. 1155–1157.
[13] J. Zonca, A. Sciutti, Does human-robot trust need reciprocity?, RO-MAN 2021 Workshop on Robot Behavior Adaptation to Human Social Norms (TSAR), 2021.
[14] S. Vinanzi, M. Patacchiola, A. Chella, A. Cangelosi, Would a robot trust you? developmental robotics model of trust and theory of mind, Philosophical Transactions of the Royal Society B 374 (2019) 20180032.
[15] S. Vinanzi, A. Cangelosi, C. Goerick, The collaborative mind: intention reading and trust in human-robot interaction, Iscience 24 (2021).
[16] N. Van Eck, L. Waltman, Software survey: Vosviewer, a computer program for bibliometric mapping, scientometrics 84 (2010) 523–538.