Old Wine in New Bottles: Are Agent-Specific Trustworthiness Measures Necessary?

Connor Esterwood*, Samia Cornelius Bhatti and Lionel P. Robert

University of Michigan School of Information, Ann Arbor, United States of America.

Abstract

A vital aspect of human–robot interaction (HRI) is trustworthiness. Measuring trustworthiness in the context of HRI, however, presents new challenges. One such challenge is the ongoing debate between whether the type of agent examined requires the use of a trustworthiness measure specific to that agent. This paper presents both sides of this debate and argues that there is no compelling reason to consider one of these measures more appropriate to a specific agent over another until evidence suggests otherwise.

Keywords

Trust, Measures, Human–Robot Interaction, Methods and Metrics

1. Introduction

Trust in robots is central to their successful integration into our daily lives [1]. Humans rely on trust to negotiate social interactions with such technologies [2]. Understa the role of trust in social relationships with robots is crucial today, where the deployment of social robots in organizations is becoming increasingly common.

To date, no universally accepted definition of trust exists in the field of human-robot interaction (HRI). One forming consensus, however, is that trust is based on one's positive expectations or attitudes towards a given trustee [3, 4, 5]. The higher order concept of *trustworthiness* often represents attributes that determine how much someone or something should be trusted. In this way, trustworthiness precedes and ultimately determines trust [3, 6, 1].

Measuring the trustworthiness of agents, such as robots, presents new challenges. Previous research often distinguishes between trustworthiness in humans and technology [7, 8]. However, advanced automation is designed to blur these boundaries [2]. Consequently, many studies have adapted measures validated for assessing humans' trustworthiness in robots [1]. These measures, however, were originally designed solely for humans, casting doubt on their reliability when applied to robots.

In response, measures have been developed specifically for advanced automation, such as robots [8, 4]. Yet these measures seem to capture similar, if not the same, attributes found in measures of human trustworthiness. This begs the question of whether such specific measures for robots are needed. No clear empirical evidence suggests that measures for humans or robots are less effective when used on the other. Without such empirical evidence, the debate remains speculative and driven by intuition rather than solid data.

This paper argues that there is no compelling reason to prefer one set of trustworthiness measures (for humans versus automation) over another until evidence suggests otherwise. This debate risks repackaging old ideas in new forms, with significant implications for research practices and theories. If distinct measures are necessary, it becomes crucial to adopt them and assess the extent to which prior research may have been compromised. Conversely, if distinct measures are not needed, researchers can confidently proceed with meta-analyses. From a theoretical perspective, resolving this question will

ALTRUIST, BAILAR, SCRITA, WARN 2024: Workshop on sociAL roboTs for peRsonalized, continUous and adaptIve aSsisTance, Workshop on Behavior Adaptation and Learning for Assistive Robotics, Workshop on Trust, Acceptance and Social Cues in Human-Robot Interaction, and Workshop on Weighing the benefits of Autonomous Robot persoNalisation. August 26, 2024, Pasadena, USA

[*] Corresponding author.

✉ cte@umich.edu (C. Esterwood*)

deepen our understanding of potential philosophical differences. If trustworthiness is similar across agents, it would be insightful to explore why. If trustworthiness differs, understanding those differences is equally important [9].

2. Background

Trustworthiness is the degree to which humans believe a robot is worthy of their trust [10, 3, 1, 11]. Trustworthiness is subdivided into 2 or 3 sub-dimensions. However, the number and nature of these sub-dimensions often vary based on whether one sees a robot as more like a human, more like automation, or unique. As such, measures of trustworthiness in HRI have diverged with the formation of various agent-specific trustworthiness measures. In HRI, the most popular of these trustworthiness measures are Mayer's trust measure [3], Jian et al. trust checklist [8], and the multidimensional measure of trust in HRI (MDMT) [4].

Mayer's trust measure was developed as part of a larger trust model in the psychology literature. This measure considers trustworthiness a composite of ability, benevolence, and integrity. Ability is a trustee's perceived competence and skillfulness [3]. Integrity is the extent to which the trustee is viewed as honest and adherent to principles [12, Pg.2]. Lastly, benevolence captures the trustee's intention to act in the best interest of the trustor [3, Pg.718].

Jian et al.'s trust checklist [8] adopts a different approach by targeting human–automation trust. Specifically, [8] divided trust into human–human trust, human–machine trust, and "trust in general" [8, Pg.12]. Through multiple cluster analyses [8] established a measure comprised of 12 unique items that specifically applied to human–machine trust. This checklist, however did not organize these items into higher order constructs implying a uni-dimensional structure. Subsequent work, however, has argued that the checklist for trust "has questions that closely map to ability, integrity, and benevolence" [13, Pg.4] implying the presence of similar sub-dimensions to that of Mayer's trust measure [3].

The MDMT examines trustworthiness by dividing it into two 3rd order constructs. At the highest level these are moral trust and performance trust [4]. These dimensions of trustworthiness can be further subdivided where performance trust is a composite of reliability and capability while moral trust is a composite of sincerity and ethics [4]. At this level the MDMT also begins to mirror Mayer's trust measure, however, [4] specifically argue that these dimensions are not fully compatible with Mayer's trust model.

The variety of trustworthiness measures in Human-Robot Interaction (HRI) complicates the field's ability to discuss findings and compare results across studies. Each measure is often based on different theoretical perspectives, depending on the type of agent it applies to. Despite this, many scholars treat these diverse measures as interchangeable [14]. On the surface, this might seem misguided, like comparing apples to oranges. However, it may be justified given these measures' lack of substantial agent-specific differences. Below, we outline the arguments for and against each perspective.

2.1. The Argument for Agent Specific Trustworthiness

Extant research posits that trust development in technology differs from trust development in humans [7]. Specifically, trust in humans hinges on factors such as the trustee's ability, integrity, and benevolence [3]. In contrast, trust in technology typically depends on the automated system's performance, purpose, and process [15]. However, as technologies like robots increasingly incorporate human-like attributes, there is a growing effort to develop standardized measures for trust in robots that integrate both human and technical factors [16]. This is evident in measures proposed by [8] and [4], which include integrity, ethics, and benevolence to address the relational aspect of trust in robots. Yet, the research also suggests that some of these factors, for example, "genuine" [4], are deemed to be relevant to robots [16].

The various trust measures and the efforts to integrate social and technical dimensions highlight that perceptions of a technology's human-like attributes play a critical role in shaping trust in human-technology interactions. If perceptions significantly impact trust, choosing between human or technical trust factors in measuring trust can be crucial empirically. For instance, using human-like trust measures for non-human-like technology might confuse respondents, making it difficult for them to provide accurate answers. Conversely, applying system-like trust measures to human-like technology could lead to respondents struggling to relate to the measures. In both scenarios, mismatched trust measures might result in lower path coefficients between trust variables and outcomes than if the appropriate measures were used. Hence, there is a need for nuanced trust measures that account for the perceived humanness of the technological agent to predict trust behaviors more accurately.

The variations in human-like design in technology and the subjective nature of perceived humanness underscore the importance of having agent-specific trust measures. Historically, people have found it easy to categorize technology as an object versus a human, hence using human-like trust measures for technology including robots may not effectively predict trust behavior [17]. However, technology can now display more or less human-like attributes, leading to varying perceptions of human-likeness [18, 17]. For most human-like robots, human-like trust measures may predict behavior. On the other hand, for less human-like robots, a combination of human and technical measures might be more appropriate. This variability highlights that a one-size-fits-all approach to trust measurement is inadequate. Tailoring trust measures to fit the technological agent's specific design and perceived humanness is crucial for accurately predicting trust behaviors.

2.2. The Argument Against Agent Specific Trustworthiness

Generally, the justification for agent-specific trustworthiness measures argues that different agents possess unique characteristics, thus justifying the need for agent-specific measurements reflective of these differences. It is, therefore, reasonable to expect that each of these measures of trustworthiness would possess significantly different items. A closer examination of these items, however, shows that these measures appear more similar than different. For example, of the three most popular measures of trustworthiness used in HRI, each measure generally contains individual items that can be mapped onto one of the three sub-dimensions of trust proposed by [3] as visible in table A.

Given the relative similarities across the trust measures [8, 4, 3], it is, therefore, unclear if these existing measures are as distinct as one might expect. As a result, scholars do not yet know if they should strictly adhere to the use of agent-specific trustworthiness measures or if these measures are truly measuring agent-specific forms of trustworthiness.

If there were support for an agent-independent trustworthiness measure, it would allow for more meaningful comparisons across different studies [14]. In particular, using a standardized measurement instrument would ensure that scholars examine the same thing –i.e., trustworthiness– rather than permutations or variations. This could accelerate research on trustworthiness in HRI and allow for more robust consolidations and replications of existing work.

3. A Call to Action

There has yet to be a robust empirical examination comparing the efficacy of trustworthiness measures across different agents. This prevents us from knowing if the field of HRI should strictly adhere to these discrete measures or if a more generalizable measure can be used across different agent types. This paper's position is that this is a critical gap in the literature. By examining if agent-specific trustworthiness measures are warranted, we can re-examine the current practice of creating and using agent-specific trustworthiness measures.

Re-examining agent-specific trustworthiness measures could enable the use of more generalizable metrics. This would, in turn, facilitate more robust comparisons of empirical results across studies. Additionally, comparing these measures can show the similarities and differences between robots and humans. This approach has the potential to offer new insights into human-machine relationships, not only in terms of trustworthiness but also in aspects like robot acceptance and usage.

References

[1] C. Esterwood, L. P. Robert Jr, Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness, Computers in Human behavior 142 (2023) 107658.

[2] S. You, L. Robert, Trusting and working with robots: A relational demography theory of preference for robotic over human co-workers, You, S. and Robert, LP (2023). Trusting and Working with Robots: A Relational Demography Theory of Preference for Robotic over Human Co-Workers, MIS Quarterly,(conditionally accepted) (2023).

[3] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of management review 20 (1995) 709–734.

[4] B. F. Malle, D. Ullman, A multidimensional conception and measure of human-robot trust, in: Trust in human-robot interaction, Elsevier, 2021, pp. 3–25.

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

[6] L. Robert, S. You, Are you satisfied yet? shared leadership, trust and individual satisfaction in virtual teams, in: Proceedings of the iConference, 2013.

[7] K. A. Hoff, M. Bashir, Trust in automation: Integrating empirical evidence on factors that influence trust, Human factors 57 (2015) 407–434.

[8] J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an empirically determined scale of trust in automated systems, International journal of cognitive ergonomics 4 (2000) 53–71.

[9] J. Kraus, I. Valori, M. Fairhurst, The social bridge: An interdisciplinary view on trust in technology, Computers in Human Behavior 149 (2023).

[10] S. Ye, G. Neville, M. Schrum, M. Gombolay, S. Chernova, A. Howard, Human trust after robot mistakes: Study of the effects of different forms of robot communication, in: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2019, pp. 1–7.

[11] L. P. Robert, A. R. Denis, Y.-T. C. Hung, Individual swift trust and knowledge-based trust in face-to-face and virtual team members, Journal of management information systems 26 (2009) 241–279.

[12] W. Kim, N. Kim, J. B. Lyons, C. S. Nam, Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach, Applied ergonomics 85 (2020) 103056.

[13] S. C. Kohn, E. J. De Visser, E. Wiese, Y.-C. Lee, T. H. Shaw, Measurement of trust in automation: A narrative review and reference guide, Frontiers in psychology 12 (2021) 604977.

[14] T. Law, M. Scheutz, Trust: Recent concepts and evaluations in human-robot interaction, Trust in human-robot interaction (2021) 27–57.

[15] J. Lee, N. Moray, Trust, control strategies and allocation of function in human-machine systems, Ergonomics 35 (1992) 1243–1270.

[16] M. Chita-Tegmark, T. Law, N. Rabb, M. Scheutz, Can you trust your trust measure?, in: Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction, 2021, pp. 92–100.

[17] N. K. Lankton, D. H. McKnight, J. Tripp, Technology, humanness, and trust: Rethinking trust in technology, Journal of the Association for Information Systems 16 (2015) 1.

[18] C. Nass, Y. Moon, Machines and mindlessness: Social responses to computers, Journal of social issues 56 (2000) 81–103.

A. Supplemental Table

Common Sub-Dimension	Mayer's Trust Measure [H-H] Items [3]	Checklist for Trust [H-A] Items [8]	Multidimensional Model of Trust [H-R] Items [4]
Ability	The agent is very capable of performing its job. The agent is known to be successful at the things it tries to do. The agent has much knowledge about the work that needs to be done. I feel very confident about the agent's skills. The agent has specialized capabilities that can increase our performance. The agent is well-qualified to perform the job.	I have confidence in the agent. The agent is dependable. The agent is reliable. I can trust the agent. I am familiar with the agent.	The agent is capable. The agent is skilled. The agent is competent. The agent is meticulous. The agent is reliable. The agent is predictable. The agent is dependable. The agent is consistent.
Benevolence	The agent is very concerned about my welfare. My needs and desires are very important to the agent. The agent would not knowingly do anything to hurt me. The agent really looks out for what is important to me. The agent will go out of its way to help me.	The agent behaves in an underhanded manner.* I am suspicious of the agent's intent, action, or output.* The agent's actions will have a harmful or injurious outcome.*	The agent is benevolent. The agent is kind. The agent is considerate. The agent has goodwill.
Integrity	The agent has a strong sense of justice. I never have to wonder whether the agent will stick to its word. The agent tries hard to be fair in dealings with others. The agent's actions and behaviors are not very consistent. I like the agent's values. Sound principles seem to guide the agent's behavior.	The agent is deceptive.* I am wary of the agent.* The agent provides security. The agent has integrity.	The agent is sincere. The agent is genuine. The agent is candid. The agent is transparent. The agent is ethical. The agent is moral. The agent is principled. The agent has integrity.

Table 1
Comparison of trust measures across common sub-dimensions based on [3]. ∗ indicates reverse coded item.