The Crisis of Trust in AI and Autonomous Systems

Amandus Krantz[1]

[1]*Lund University Cognitive Science*

amandus.krantz@lucs.lu.se

The future is robotic. Already we are seeing how society is changing with self-driving cars and robots at hospitals and schools. The considerable potential of autonomous systems (AS) is highly discussed. What is not highly discussed, and rarely even acknowledged, is the key role trust plays in realizing these benefits and the problems this may cause for human-AS interaction research. There currently exists no common way of defining, testing, or measuring trust. This lack of common foundation, both for trust in general but also in human-AS relations, may at best result in sporadic progress and adoption of these systems and may at worst lead to public disillusionment and abandonment, delaying the potential benefits of AS.

According to Glikson & Woolley (2020) the match between a user's trust in technology and that technology's abilities is a predictor for future use; low trust in capable technology leads to disuse, while high trust in incapable technology leads to frustration which leads to misuse which in turn may lead to dangerous situations. For example, a user with low trust in the capabilities of their robotic vacuum cleaner will be more inclined to vacuum manually, negating the benefits of the robotic vacuum. On the other hand, a user who puts too much trust in their self-driving car's ability to avoid obstacles may feel comfortable enough to sleep at the wheel, potentially causing accidents if the car encounters an obstacle it cannot avoid. Enholm, Papagiannidis, Mikalef, & Krogstie (2021) agrees that for an AI-system to be used at all in a business setting, the user must have some level of trust in it. Marsh (1994), in an early attempt at creating a taxonomy for trust in human-AS interaction, writes that trust should be considered a fundamental part of cooperation and communication.

Given this, it would seem that research on trust, both in general and in human-AS relations, should be a highly prioritized area. Understanding how trust works could reduce resources wasted on disused technology, and minimize the dangers that come with misuse and over-reliance on incapable technology, making trust research beneficial for society. Unfortunately, this is often not the case. Research that focus on how trust works and how to measure it in human-AS relations is pretty limited, and the little focused research that does exist is often plagued by several problems.

The first of these problems is the use of short, single measure, experiments. Trust is not a constant, it changes as the interaction proceeds, it is dynamic (Blomqvist, 1997). A single question about trust at the end of a study only shows what the participant thought at that particular time, but tells you nothing about how the experiment actually impacted the trust (Glikson & Woolley, 2020).

Second is the problem of unclear terminology (Cameron et al., 2021; Jessup, Schneider, Alarcon, Ryan, & Capiola, 2019). Trust in AS is typically presented in terms of performance and reliability; however, there is a second type of trust that is more general, established before one can make a rational judgement about reliability. It is based more on instinct, emotions, and gut feeling (Fiske, Cuddy, & Glick, 2007; Marsh, 1994; McAllister, 1995). Without clear terminology to indicate which type of trust is being measured, researchers run the risk of measuring something they are not intending (Chita-Tegmark, Law, Rabb, & Scheutz, 2021).

Related to the problem of unclear terminology is finally the problem of overly simple, varied, and non-standard methodology (Glikson & Woolley, 2020). There exists no common method of measuring trust (Chita-Tegmark et al., 2021; Gao, Sibirtseva, Castellano, & Kragic, 2019), leading many researchers to make use of vague Likert scale questioning, for example "On a 5 point scale, how much do you trust this robot/person/agent?". These types of questions are problematic since small changes in the scale (e.g., a 7 point scale instead of a 5 point scale) or, as mentioned, the terminology, can make it difficult, if not impossible, to generalize and compare the results with other studies (Chita-Tegmark et al., 2021). Using home-brew methodologies may also cause issues with statistical significance, as shown by Schrum, Johnson, Ghuy, & Gombolay (2020) who discovered that only 3 of the 110 peer-reviewed human-robot interaction papers they examined had properly implemented and analysed their questionnaires.

Some attempts have been made to create methodologies for the measurement of trust (Bartneck, Kulić, Croft, & Zoghbi, 2009; Berg, Dickhaut, & McCabe, 1995; Schaefer, 2016). They have, however, failed to reach any kind of common usage as there seems to exist some doubts about whether they are transferable across field boundaries (Glikson & Woolley, 2020). For example, the investment-style games proposed by Berg, Dickhaut, & McCabe (1995) seems to work well for human-human trust related to investments, but the methodology may not work as well for human-AS interaction as it may require the participant to make unrealistic assumptions about the capabilities of the AS (e.g. its intelligence or level of autonomy) which may impact the reported level of trust.

At the heart of these problems lies the more in-depth problem of defining trust. No commonly accepted definition of trust currently exists, and progress towards creating one is slow. Blomqvist (1997), giving an overview of the many definitions of trust, writes that of the fields investigated, only the field of social psychology has a reasonable definition of trust, while moral philosophy and economics either do not address the topic at all or have created definitions that allow them to ignore it. Yet, the word trust is used in pretty much every field, from AI to political science to law. It is used so much, in so many fields, that one almost has to assume that it is referring to the same concept. However, trust in a human and trust in technology are two very different things, and trust in technology and trust in AS is another one still (Glikson & Woolley, 2020). Distinctions like these are vitally important when developing methodologies and measures for trust studies, as a methodology that works for evaluating trust in humans may be nonsensical when applied to trust in a self-driving car or humanoid robot. Researchers have to keep this in mind not only when transferring measures and methodologies from human-human trust research to human-AS research, but also when dealing with different types of AS. Trust in a self-driving car, for example, may not work the same as trust in a robotic vacuum cleaner, factory robot, or military drone.

Trust, then, should be considered a fundamental part of human-AS interaction, and is likely a requirement for cooperation between humans and AS to even start (Enholm et al., 2021; Marsh, 1994). Yet, our understanding of what it actually is, how it behaves, and its mechanics is very limited, and attempts at increasing this understanding are often hindered by fundamental problems, such as unclear terminology and methodologies that make results difficult to generalize.

A fully unified and universal definition of trust may not be possible, but we must at least attempt, through interdisciplinary efforts, to find a common foundation from which discussion and progress can grow. If we do something about this fundamental problem at this early stage by establishing a solid foundation, well implemented methodologies, and clear terminology, we will have the chance to gain an increased understanding for one of the most fundamental requirements for society, communication, and interaction.

Acknowledgement

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

References

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, *13*(3), 271–286. https://doi.org/10.1016/S0956-5221(97)84644-1

Cameron, D., de Saille, S., Collins, E. C., Aitken, J. M., Cheung, H., Chua, A., … Law, J. (2021). The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in Human Behavior*, *114*. https://doi.org/10.1016/j.chb.2020.106561

Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can You Trust Your Trust Measure? *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 92–100. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3434073.3444677

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2021). Artificial Intelligence and Business Value: A Literature Review. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10186-w

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gao, Y., Sibirtseva, E., Castellano, G., & Kragic, D. (2019). Fast adaptation with meta-reinforcement learning for trust modelling in human-robot interaction. *ArXiv:1908.04087 [Cs]*. Retrieved from http://arxiv.org/abs/1908.04087

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality. Applications and case studies* (pp. 476–489). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_32

Marsh, S. P. (1994). Formalising trust as a computational concept (PhD). University of Sterling.

McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, *38*(1), 24–59. https://doi.org/10.5465/256727

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust intelligence and trust in autonomous systems* (pp. 191–218). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-7668-0_10

Schrum, M. L., Johnson, M., Ghuy, M., & Gombolay, M. C. (2020). Four years in review: Statistical practices of likert scales in human-robot interaction studies. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 43–52. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3371382.3380739