Tool or team-mate?

Tool or team-mate?

Extending our understanding of the specific requirements for trust in collaborative robotics

Melanie McGrath*

Data61, CSIRO, melanie.mcgrath@data61.csiro.au

Andreas Duenser

Data61, CSIRO, andreas.duenser@data61.csiro.au

Cécile paris

Data61, CSIRO, cecile.paris@data61.csiro.au

Technology is an integral part of almost every aspect of our modern lives, but mounting evidence suggests we get the best results from technological systems when humans work collaboratively with machines. Collaborative Intelligence is an approach based on creating sustained, meaningful, and productive partnerships between human and artificial intelligence. Essential to creating and maintaining these collaborative relationships are user trust and acceptance. We highlight limitations of existing trust models for understanding trust in the context of collaborative human-technology relationships and call for the development of a specific framework of antecedents, processes, and outcomes of trust in collaborative intelligence.

CCS CONCEPTS • **Human-centered computing - Human computer interaction (HCI) - Interaction paradigms - Collaborative interaction**

Additional Keywords and Phrases: Trust, Collaborative Intelligence, Collaborative Robotics, Artificial Intelligence, Collaboration

* Corresponding author

1 INTRODUCTION

After decades of research developing automated and autonomous systems with the capacity to replace human function, interest is growing in how machines may instead be deployed alongside humans in ways that maximise the strengths of each. Collaborative intelligence (CINTEL) combines human intelligence, adaptability, creativity and values with narrower but powerful machine intelligence to develop collaborative human-technology systems [1].

Irrespective of the robustness of a system, it is likely that it will in some instances fall short of expectations. Therefore, trust in technology must be appropriately calibrated to actual system performance [2]. Over-trust may lead to users depending upon the system even when it is faulty or performing outside its parameters. Under-trust may lead to disuse of the system, resulting in excessive user cognitive load [3] and diminished system performance [4]. Thus, facilitating appropriate levels of trust in emerging collaborative systems, such as collaborative robotics, is central to maximising performance and safety of human-technology teams [5, 6]. A wide array of models has been developed to account for trust in automation and autonomous systems [7]. However, given the unique requirements of collaboration, as well as blending and leveraging of human and machine capabilities, we anticipate the dynamics of trust formation and outcomes may diverge, at least to some extent, from established models.

In this paper we review different models of trust in automation, trust in artificial intelligence, as well as interpersonal trust in human teams, and how these may relate to trust formation in collaborative intelligence systems. Most of these models have certain limitations when it comes to considering dynamic trust processes in collaborative human-technology teams. We highlight some of these limitations with the intention of opening a discussion of the need to develop a more specific framework for trust in CINTEL with respect to trust antecedents, processes and outcomes.

2 COLLABORATIVE INTELLIGENCE

While we often imagine an either-or choice between human and technology, for most tasks, evidence suggests humans and machines work better together [8, 9]. Fully realised, CINTEL applications, including collaborative robotics, involve a human and a machine collaboration with a shared understanding of an objective and the collaborative context, working interdependently over a sustained period to meet that objective. To meet this shared objective, both parties have the capacity and opportunity to share knowledge and respond adaptively.

Collaborative robotics requires new ways for humans and machine agents to interact beyond a mental model that sees robots as merely tools or involves a simple division of labour. This might require re-imagining tasks and workflows, identifying when it is best for humans and robots to collaborate and how this collaboration should take place. Effective functioning of human-robot teams will also rely on sophisticated communication capabilities between humans and machines. Robotic systems will not only need to explain their processing in a way that human team members can engage with, but also respond to communication from the human considering the history of what has been communicated and achieved to that point. Whether between humans or between humans and machines, collaboration requires shared understanding of the problem to be solved and the situation. This includes being able to interpret the current status of a collaborative partner, as well as the ability on the part of the robot to construct and share an abstraction of its actions and the data it is dealing with, so that the human can understand what is happening, provide feedback, or take over if required, without being cognitively overwhelmed. These points of difference between human-AI collaboration and traditional applications of automation and AI have implications for the formation and outcomes of human-machine trust.

3 TRUST IN TECHNOLOGY

Trust is a complex, multi-faceted construct investigated in many contexts, with a multitude of methods, and linked to an array of significant outcomes. The scientific literature is replete with conceptions of trust; however, the most established and widely used tend to converge on a set of key features; namely, (a) an expectation or belief that (b) a specific subject will (c) perform future actions with the intention of producing (d) positive outcomes for the trustor in (e) situations characterised by risk and vulnerability [10].

Trust in technology has many aspects. It includes trust that a system is behaving as expected; that it provides accurate outputs and its reasoning is sound; that the data employed is appropriate and accurate, without bias or with a bias that is well understood; that the system protects data and resists attacks (i.e., the system is secure); that the humans will remain physically safe (e.g., robotic arms moving in close proximity to humans, or ensuring safe operations in a chemical laboratory) and also psychologically safe when relying on the outputs of the system. Perceptions of risk associated with these outcomes influence trust in technology and its uptake [11].

3.1 Trust in automation and autonomous systems

A comprehensive model of factors that influence trust in automation technology has been presented by [5]. This model attributes trust formation to three high level factors: dispositional trust, situational trust, and learned trust. Factors contributing to dispositional trust include individual differences, such as culture, age, or a person's tendency to trust technology in general. Situational trust responds to external factors such as system complexity, type of system or task, as well as internal factors such as user expertise and mood. Learned trust is based on experience with similar systems (initial learned) and actual experience of the system (dynamic learned).

The consequences of various human factors for use and acceptance of automated systems have also been explored in depth. Frameworks such as the Technology Acceptance Model [12] and the Unified Theory of Acceptance and Use of Technology [13] have been successful in a wide range of contexts in explaining when users will be more or less willing to engage with new technology. For example, Ghazizadeh et al. [14, 15] found that trust was a major determinant of intention to use an onboard vehicle monitoring system, suggesting that the acceptance model can be augmented by considering trust in technology. Such models provide a strong basis for considering the antecedents and consequences of trust in human-robot teams, and human-AI teams more broadly; yet, given the collaborative nature of such systems, we may also expect to find convergences with models of trust between humans.

3.2 Human-human (interpersonal) trust

Much research on trust in technology has taken inspiration from human-human trust research. Although definitions of interpersonal trust vary, most incorporate two broad principles reflecting willingness to accept risk or vulnerability, combined with positive expectations of the trustee's capabilities and intentions [16, 17, 18]. Factors contributing to formation of trust between humans can be categorised into trustor characteristics, trustee characteristics, shared characteristics, communication processes, and situational or external characteristics [14]. Trustor characteristics have parallels with Hoff and Bashir's [5] concept of dispositional trust. Propensity to trust [19] and attachment style [20] are examples of personal attributes shown to predispose an individual to trust another. Trustee characteristics encompass perceptions of the ability, integrity, and benevolence of the trustee [21], and correspond to factors that contribute to learned trust in technology models. Communication processes that contribute to trust between humans may have particular relevance for collaborative robotics. Smoothness or ease of communication has been shown to have a positive influence on trust [22], as has legitimate and easy-to-understand explanation of decisions and actions [23].

4 LIMITATIONS OF CURRENT MODELS FOR UNDERSTANDING TRUST IN COLLABORATIVE ROBOTICS

Many existing models of trust in technology share limitations, especially when it comes to extending our understanding of trust dynamics in collaborative human-technology teams. One such limitation is the relative neglect of the contribution of individual differences to trust formation and maintenance. While most models of trust in technology recognise the role of trustor characteristics [e.g., 4, 5], in practice these individual differences are often given negligible weight or are treated as noise in empirical research [24]. This failure to account for individual differences in human-machine trust is likely to have strong implications in the context of collaborative robotics given evidence of the importance of trust to effective human teaming [25, 26].

Models of trust in technology also tend not to allow for variation in antecedent factors between trust trustees. In their meta-analyses, Hancock and colleagues [27] and Kaplan and colleagues [28] demonstrated that the factors contributing to trust development differ when referring exclusively to robots or AI. Preliminary investigation from [28] further found that antecedents varied in significance across classes of AI technology such as algorithms, automated vehicles and chatbots. Similarly, Glikson and Woolley [29] discuss various forms of AI representation (robot, virtual, embedded) and different levels of machine intelligence or capabilities as important antecedents to the development of trust in AI. This is consistent with other evidence that the factors influencing trust development are dependent on the nature of the trust referent. Biermann et al [30], for example, found that anthropomorphic design features had a different influence on reported trust in robots in a care versus production task context. It is reasonable, therefore, to anticipate that, in the context of human-robot collaboration, factors shown to be relevant to trust in human teams (e.g., communication capability, team tenure, and task parameters) may take on greater prominence.

Finally, although many models recognise the role of feedback and calibration, in practice there has been little exploration of the temporal dynamics of trust formation and maintenance, especially in the context of longitudinal teaming [31]. The temporal trajectories of trust formation have been shown to differ for interpersonal trust and trust in technology [32]. Trust between humans tends to start at a lower level and build with time and interaction [33]. In contrast, people tend to have high initial expectations of the trustworthiness of an automated system, which have proven to be highly sensitive to violation with significant consequences for ongoing trust [34]. It remains an open question whether collaborative interaction with robot technologies will follow this latter trajectory or trend towards the progression associated with interpersonal forms of trust.

To inform the design of trustworthy collaborative robotic systems and reduce potential barriers for system adoption, we are developing a broad framework of trust in CINTEL that should overcome limitations such as these and addresses the specific implications of collaborative human-technology teams. Such a framework will need to integrate a complex array of inputs, outcomes, and processes. It must support the context-dependency of antecedent factors and reflect the potential for their differential influence on trust over the lifespan of the human-technology team. In capturing the trajectory of trust development, maintenance, and effects over time, such a framework will also facilitate the meaningful embedding of trust calibration processes and outcomes.

5 CONCLUSION

The formation, maintenance, loss, or repair of trust in collaborative human-technology and robot teams is multi-faceted. To develop our understanding of which factors contribute to such trust, we can look at insights from research on trust in automation, autonomous systems, artificial intelligence, and trust in human-human teams. However, models developed in these contexts do not comprehensively incorporate factors that allow us to fully understand how trust develops in

collaborative human-technology teams. The models often do not properly consider the importance of individual differences to trust formation and maintenance, the importance of the nature and type of trust referent, and the dynamic nature of trust formation over time.

We encourage the community to join the discussion about these and other factors that may be particularly important when investigating trust in human-technology teams. With a better understanding of the limitations of current models of trust in technology, we will be in a better position to design and develop a framework of trust in collaborative human-technology teams and collaborative robotics in particular. In the development of and empirical validation of this trust in CINTEL framework we employ an interdisciplinary approach. Drawing on input from a variety of disciplines, such a framework can serve as a starting point for considering in more detail which components are important for trust to develop and maintain in different contexts. The management literature is rich in theoretical approaches to trust in teaming that have the potential to powerfully supplement the existing knowledge in the robotics and human computer interaction domains, while the research from psychology and other social sciences can deepen our understanding of the complexity of human cognitive processes and behaviour when collaborating with CINTEL systems.

PARTICPANT PROFILES

Melanie McGrath is a social psychological researcher in the CINTEL Future Science Platform of CSIRO's Data61. Within the FSP she engages in interdisciplinary research investigating the nature and function of human trust in collaborative intelligence systems. Previously she explored questions of harm and ethics in a range of applied contexts. Melanie is interested in finding practical and impactful pathways for insights from the social sciences to inform the development, deployment and evaluation of novel technologies.

Andreas Duenser is a research scientist at CSIRO's Data61. His work lies at the confluence of humans and technology and aims at gaining better understanding of human-machine and human-data interaction with a focus on trust in socio-technical systems such as collaborative intelligence.

Cécile Paris is a Chief Research Scientist and the Director of the CINTEL Future Science Platform at CSIRO. She is an expert in Natural Language Processing, Intelligent User Interfaces and User Modelling. Her current work aims to create teams that maximise the benefits of both human and machine intelligence in a wide range of domains including human-robot teams, cybersecurity control systems, and marine surveillance.

REFERENCES

[1] C. Paris and A. Reeson. Nov. 30, 2021. "What's the secret to making sure AI doesn't steal your job? Work with it, not against it," *The Conversation*.

[2] B. M. Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 2007, doi: 10.1080/00140139408964957.

[3] K. Gupta, R. Hajika, Y. S. Pai, A. Duenser, M. Lochner, and M. Billinghurst. 2019. In AI we trust: Investigating the relationship between biosignals, trust and cognitive load in VR. In *25th ACM Symposium on Virtual Reality Software and Technology (VRST '19*), November 12-15, 2019, Parramatta, NSW< Australia. ACM, New York, NY, USA, 10 pages. doi: 10.1145/3359996.3364276.

[4] J. D. Lee and K. A. See, 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, vol. 46, no. 1, pp. 50–80. doi: 10.1518/hfes.46.1.50_30392.

[5] K. A. Hoff and M. Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, vol. 57, no. 3, pp. 407–434. doi: 10.1177/0018720814547570.

[6] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Human-Computer Interact.*, vol. 5, no. CSCW1, pp. 1–21, Apr. 2021, doi: 10.1145/3449287.

[7] B. French, D. Andreas, and H. Andrew. Trust in Automation: A literature review. Sandy Bay, 2018. Accessed: May 10, 2021. [Online]. Available: https://publications.csiro.au/publications/publication/PIcsiro:EP184082/SQtrust in automation/RP1/RS25/RORECENT/STsearch-by-keyword/LISEA/RI4/RT5.

[8] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu. 2022. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *The Lancet Digital Health*, vol. 4, No. 7. e507-e519. doi: 10.1016/S2589-7500(22)00070-X

[9] P. Phillips, A. Yates, Y. Hu, C. Hahn, E. Noyes, K. Jackson, J. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J. Chen, C. Castillo, R. Chellappa, D. White and A. O'Toole. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. In *Proceedings of the National Academy of Sciences*, vol. 115, No. 24, 6171-6176. doi: 10.1073/pnas.1721355115

[10] S. Castaldo, K. Premazzi, F. Zerbini. 2010. The meaning(s) of trust: A content analysis of the diverse conceptualizations of trust in scholarly research on business relationships. *Journal of Business Ethics*, vol. 96, no. 4. doi: 10.1007/s10551-010-0491-4

[11] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *FAccT 2021 - Proc. 2021 ACM Conf. Fairness, Accountability, Transpar*., pp. 624–635, Mar. 2021, doi: 10.1145/3442188.3445923.

[12] F. D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q. Manag. Inf. Syst*., vol. 13, no. 3, pp. 319–339. doi: 10.2307/249008.

[13] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS Q. Manag. Inf. Syst*., vol. 27, no. 3, pp. 425–478. doi: 10.2307/30036540.

[14] M. Ghazizadeh, J. D. Lee, and L. N. Boyle. 2012. Extending the Technology Acceptance Model to assess automation. *Cogn. Technol. Work*, vol. 14, no. 1, pp. 39–49. doi: 10.1007/S10111-011-0194-3/FIGURES/3.

[15] M. Ghazizadeh, Y. Peng, J. D. Lee, and L. N. Boyle. 2012. Augmenting the Technology Acceptance Model with trust: Commercial drivers' attitudes towards monitoring and feedback. In *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting - 2012* pp. 2286–2290. doi: 10.1177/1071181312561481.

[16] C. Ashley Fulmer and M. J. Gelfand. 2012. At what level (and in whom) we trust: Trust across multiple organizational levels. *J. Manage*., vol. 38, no. 4, pp. 1167–1230/ doi: 10.1177/0149206312439327.

[17] R. C. Mayer, J. H. Davis, and F. D. Schoorman. 1995. An Integrative Model Of Organizational Trust. *Acad. Manag. Rev*., vol. 20, no. 3. doi: 10.5465/amr.1995.9508080335.

[18] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer. 1998. Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, vol. 23, no. 3, pp. 393-404.

[19] K. T. Dirks and D. L. Ferrin. 2002. Trust in leadership: Meta-analytic findings and implications for research and practice. *J. Appl. Psychol*., vol. 87, no. 4, pp. 611–628. doi: 10.1037/0021-9010.87.4.611.

[20] B. L. Simmons, J. Gooty, D. L. Nelson, and L. M. Little. 2009. Secure attachment: Implications for hope, trust, burnout, and performance. *Journal of Organizational Behavior*, vol. 30, no. 2. doi: 10.1002/job.585.

[21] J. A. Colquitt, B. A. Scott, and J. A. LePine. 2007. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol*., vol. 92, no. 4. doi: 10.1037/0021-9010.92.4.909.

[22] A. F. Cameron and J. Webster. 2011. Relational outcomes of multicommunicating: Integrating incivility and social exchange perspectives. *Organ. Sci*., vol. 22, no. 3. doi: 10.1287/orsc.1100.0540.

[23] S. Sonenshein, M. Herzenstein, and U. M. Dholakia. 2011. How accounts shape lending decisions through fostering perceived trustworthiness. *Organ. Behav. Hum. Decis. Process*., vol. 115, no. 1. doi: 10.1016/j.obhdp.2010.11.009.

[24] K. Schaefer, B. Perelman, G. Gremillion, A. Marathe and J. Metcalfe. 2021. A roadmap for developing team trust metrics for human-autonomy teams. *Trust in human-robot interaction* (pp. 261-300). Academic Press. doi: 10.1016/B978-0-12-819472-0.00012-5.

[25] A. Costa, C. Fulmer and N. Anderson. 2018. Trust in work teams: An integrative review, multilevel model, and future directions. *Journal of Organizational Behavior*, vol. 39, no. 2, pp. 169-184. doi: 10.1002/job.2213.

[26] B. De Jong, K. Dirks and N. Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology*, vol. 101, no. 8, pp. 1134-1150. doi: 10.1037/apl0000110.

[27] P. Hancock, D. Billings, K. Schaefer, J. Chen, E. de Visser and R. Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, vol. 53, no. 5., pp. 517-527. doi: 10.1177/0018720811417254.

[28] A. Kaplan, T. Kessler, C. Brill, and P. Hancock. 2021. Trust in artificial intelligence. Meta-analytic findings. Human Factors. doi: 10.1177/00187208211013988

[29] E. Glikson and A. Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, vol. 14, no. 2. doi: 10.5465/annals.2018.0057.

[30] H. Biermann, P. Brauner and M. Ziefle. 2020. How context and design shape human-robot trust and attributions. *Paladyn (Warsaw)*, vol. 12, no. 1, pp. 74-86. doi: 10.1515/pjbr-2021-0008.

[31] E. de Visser, M. Peeters, M. Jung, S. Kohn, T. Sahw, R. Pak and M. Neerincx. 2019. Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, vol. 12, no. 2, pp. 459-478. doi: 10.1007/s12369-019-00596-x.

[32] G. Alarcon, A. Capiola and M. Pfahler. 2020. The role of human personality on trust in human-robot interaction. *Trust in human-robot interaction* (pp. 159-178). Academic Press. doi: 10.1016/B978-0-12-819472-0.00007-1.

[33] M. Coovert, E. Pavlova Miller, W. Benett Jr. 2017. Assessing trust and effectiveness in virtual teams: Latent growth curve and latent change score models. *Social Sciences*, vol. 6, no. 3. Doi:

[34] M. Dzindolet, L. Pierce, H. Beck, L. Dawe and W. Anderson. 2001. Predicting misuse and disuse of combat identification systems. *Military Psychology*, vol. 13, no. 3, pp. 147-164. doi: 10.1207/S152327876MP1303_2.