Submission Template for ACM Papers

Trust in collaborative intelligence

A recurring phase framework of trust in human-AI collaboration

Melanie McGrath*

CSIRO, melanie.mcgrath@data61.csiro.au

Andreas Duenser

CSIRO, andreas.duenser@data61.csiro.au

Justine Lacey

CSIRO, Justine.lacey@csiro.au

Cécile Paris

CSIRO, cecile.paris@data61.csiro.au

Collaborative intelligence (CINTEL) combines the unique skills and capabilities of humans and machines in sustained teaming relationships leveraging the strengths of each. In tasks involving regular exposure to novelty and uncertainty, collaboration between adaptive, creative humans and powerful, precise artificial intelligence (AI) promises new solutions and efficiencies. User trust is essential to creating and maintaining these collaborative relationships. Established models of trust in AI are limited in their capacity to account for the particular requirements of collaborative human-technology relationships, context-dependency, and temporal dynamics of trust formation and maintenance. Drawing on both the psychological and computer science literature, the recurring phase framework of trust presented in this paper adopts the tripartite structure of antecedents established by earlier models, while incorporating teamwork processes and performance phases to capture the dynamism inherent to trust in teaming contexts. These features enable active management of trust in CINTEL systems, with practical implications for the design and deployment of human-AI teams.

CCS CONCEPTS • Human-centered computing • Human computer interaction • HCI theory, concepts and models

Additional Keywords and Phrases: Collaborative intelligence, trust, human-AI teaming, artificial intelligence

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

^{*} Corresponding author.

1 INTRODUCTION

Although artificial intelligence (AI) can be highly effective where the problem space is clearly defined and specified, it remains challenged by situations and applications that demand not only powerful processing, but a capacity to respond adaptatively to ambiguity or novelty. Consequently, interest is growing in how humans and AI can work together in such situations to maximize the relative strengths of each [1]. Collaborative intelligence (CINTEL) is one approach to meeting this challenge. CINTEL applications involve teams comprising both human(s) and AI(s) with a shared objective, working interdependently over a sustained period to meet that objective. It leverages the notion of complementarity; the empirically verified observation that under many conditions the performance of human-AI teams exceed those of either human or machine teams alone [2]–[4]. The conditions under which complementarity is observed, and for which a CINTEL approach is likely to be most beneficial, are those with a task component where humans continue to hold a comparative advantage over machines, for example, moral reasoning, strategic and critical thinking.

Trust is a critical element in appropriate reliance on, and compliance with, any automated or autonomous system e.g., [5], [6]. Appropriate trust in CINTEL applications, however, will be necessary to provide a foundation not only for appropriate reliance and compliance, but for the sustained working relationship that characterizes collaborative teaming. Given the unique requirements of collaboration and CINTEL's novel pairing of human and machine capabilities, it is likely that the dynamics of trust formation and outcomes in these applications may diverge from established models of trust in automation and AI (A/AI). In this paper we introduce a framework for the antecedents, processes, and outcomes of trust in collaborative human-AI teaming, and discuss the implications of this framework for design and research.

2 MODELS OF TRUST IN HUMANS AND TECHNOLOGY

Trust is a complex, multi-faceted construct investigated in many contexts, and with many conceptualizations. The most established and widely used, however, 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 characterized by risk and vulnerability [7].

Mayer, Davis and Schoorman's [8] model of trust is one of the most widely cited in the literature on trust between humans. In this model, trust is predicted by specific characteristics of both the trustor and the trustee. The relevant characteristic of the trustor is their trait-based tendency to offer trust, independent of the target of that trust [9]. This is referred to as *trust propensity* by Mayer et al [8], and elsewhere in the literature as *dispositional trust* (e.g., [5]). The primary characteristic of the trustee in this model is their perceived trustworthiness along three dimensions; benevolence, integrity, and ability [8], [10], [11]. Many approaches to trust in A/AI draw on the framework established by Mayer et al [8], in particular its differentiation between the influence of trustor and trustee characteristics in trust development [12]. Various attempts have also been made to map the ability, benevolence, and integrity dimensions of trustworthiness to perceptions of A/AI systems [13].

Lee and See's [6] influential exploration of the role of trust in automated systems echoes Mayer et al.'s [8] tripartite categorization of trust antecedents into individual, contextual, and trustee characteristics. The individual characteristics identified as contributing to trust development are trust propensity and the nature of an individual's previous experience with automated systems. Contextual factors are linked to characteristics of the organization and culture in which the system is embedded, for example social norms and expectations and the attitude of third parties. Finally, trustee characteristics are described exclusively in terms of perceived trustworthiness and linked explicitly to Mayer et al's [12] dimensions of ability, integrity and benevolence. The authors nonetheless recognize that ability, integrity, and benevolence as articulated by

Mayer et al [12] have uniquely human implications and recast their dimensions of trustworthiness in terms more applicable to machine agents, that is characteristics of performance, process, and purpose.

In 2015, Hoff and Bashir [5] updated and extended Lee and See's [6] work with their three-layer model of trust in automation. 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).

Across these models we see a clear chain of inheritance from investigations of trust between humans to those of trust in A/AI. Other significant models of trust in A/AI diverge further from theories of interpersonal trust, retaining the tripartite structure of antecedents, but breaking with efforts to fit technological factors into Mayer et al's [12] human dimensions of perceived trustworthiness. Meta-analyses of the factors contributing to trust specifically in robots [14] and AI [15] each identify human-related, technology-related, and environment-related categories of antecedent factors consistent with the tripartite framework. Notably, however, the factors within those categories that contribute to trust formation vary across technological targets (robots or AI) [14], [15]. Consequently, just as we see a shift in antecedents between automation, robots, and AI, we may expect to see a novel constellation of factors predict trust in expressly collaborative forms of AI.

3 FROM TRUST IN A/AI TO TRUST IN COLLABORATIVE INTELLIGENCE

Many of the models described in section 2.0 share limitations when it comes to extending our understanding of trust in technology to incorporate collaborative intelligence, specifically: the contribution of individual differences to trust formation, the context dependency of antecedent factors, and the static nature of many trust models.

3.1 Contribution of Individual Differences

While human-related factors frequently feature in models of trust in A/AI, in empirical research the role of individual differences is often given negligible weight or treated as noise [17]. A failure to account appropriately for individual differences in trust formation and maintenance could have particular implications for CINTEL applications. Research on human-human team trust has shown that an individual's propensity to trust impacts perceptions of trustworthiness, trust itself, and in some cases, trust behaviors [16]. When we sideline these systematic differences among the human partners in a team, we undermine our ability to get a full picture of trust in human-AI collaboration.

3.2 Context-dependency of Antecedent Factors

Existing models of trust in A/AI tend to put forward a single set of antecedent factors, yet meta-analyses show that factors influencing trust formation differ between targets [14], [15]. We anticipate that factors that contribute to trust formation in CINTEL contexts may differ, at least to some extent, from factors that contribute to trust in A/AI. Factors shown to be relevant to human-teaming, such as communication capability and team tenure, are likely to take on more prominence in a CINTEL environment [17], [18]. No single set of factors and processes can be expected to capture the formation and maintenance of trust across the range of AI systems currently in use and the diverse range of CINTEL systems in development. To meet the needs of this proliferation, a model of trust must have the capacity to respond to the specific contextual conditions of the trustor, the trustee and the collaborative task-environment.

3.3 Static Versus Dynamic Models

Although many models recognize the role of feedback and recalibration, in practice there has been little exploration of the temporal dynamics of trust formation and maintenance, especially in the context of longitudinal teaming [19]. Mayer et al's [12] early model of trust between humans includes explicit feedback loops between the outcomes of trust and trustworthiness perceptions, yet empirical testing of pathways in this model more often than not neglect these feedback loops [20]. We see similar disjunctions between theory and empirical investigation for models of trust in A/AI [19]. Yet research specifically exploring temporal trajectories of trust development towards both human and A/AI targets is clear that this process is highly dynamic. Understanding the specific trajectories of trust formation and maintenance across a range of CINTEL applications will be critical to developing long-term teaming capacity and ensuring appropriate trust calibration.

4 A RECURRING PHASE FRAMEWORK OF TRUST IN COLLABORATIVE INTELLIGENCE

In this section we present a novel framework of trust in CINTEL that builds on and extends models of trust in A/AI. What we propose here is not a model of trust in CINTEL, but rather a framework of components that may come together in different ways and different contexts to produce and sustain trust. Models can become bloated or lack ecological validity if they attempt to include all possible antecedent factors [21]. We propose categories of trust antecedents in the tripartite framework established in the literature, without specifying an exhaustive list of factors under each of those categories. It is highly likely that many aspects of the trust development and maintenance process will vary on numerous dimensions, across time, and across applications. The factors that contribute to trust in an embodied robot, for example, are likely to be different to the factors relevant to the formation of trust in an algorithm supporting collaborative digital data curation. Each of these applications requires its own 'recipe' for trust, and each of these recipes may work best with particular ingredients [22]. To allow for this variation across CINTEL applications we recommend researchers select factors that are theoretically supported for inclusion in models of trust for specific applications. To extend our analogy, these factors can be likened to ingredients in a *pantry* that are included in specific trust models based on the requirements of the *recipe*.

In our framework of trust in CINTEL, we also recognize that trust serves a purpose, and these purposes may vary across the life cycle of the CINTEL team. Trust will be important not just to encourage appropriate reliance on CINTEL technologies, but to facilitate the sustained working relationship that characterizes collaborative intelligence. Therefore, under the category of 'outputs' (see Figure 1), we allow for trust to influence various metrics of team effectiveness. Again, these represent an indicative list and selection from this list will be based on the specific context of the target application.

4.1 Process Model

To incorporate the temporal dynamics of trust into our framework, we draw on decades of literature in the fields of management and organizational psychology on team processes. The incorporation of well-established team processes distinguishes the framework of trust in CINTEL from earlier models of trust in automation and AI. Team processes are the interdependent activities that convert inputs to outputs, often indirectly via mediating constructs like trust. They are fundamental to facilitating collaboration and enhancing team effectiveness [23]. If we conceive of trust in a CINTEL application requiring a certain recipe and antecedent factors as ingredients for this recipe, then the team processes of the model are, quite straightforwardly, the processes by which those ingredients are prepared and transformed into a particular dish. A wide range of teamwork processes have been investigated in the organizational literature, with the recommendation being to include those most relevant to the specific teams and tasks under study [24]. The four team processes identified

in the present framework are selected for their compatibility with trust in human-AI collaboration from an empirically supported taxonomy of 10 team processes [24], [25].

4.2 Recurring Phase Framework

Although a process framework of this nature introduces a degree of dynamism to modelling trust that exceeds that of many previous models, it does not fully capture the trajectory of trust over time. To do this we need to take account of the effect of sustained interaction with a CINTEL system over a series of performance episodes. Marks and colleagues [24] define an episode as a "distinguishable period of time over which performance accrues and feedback is available [...] Episodes are most easily identified by goals and goal accomplishment periods, and the conclusion of one period normally marks the initiation of another" (p. 359). In a recurring phase model, the outputs from one episode become one of the inputs for the next cycle (see Figure 1). In this way, mediating constructs like trust and its outputs alter subsequent iterations of episodes in a cyclical pattern that allows for dynamic change in trust and its outcomes over time. This recurring phase structure also makes it possible to capture the trajectory of trust development, maintenance, and influence over time, and supports the potential for factors and processes to vary in effect at different points in the life cycle of a CINTEL team.

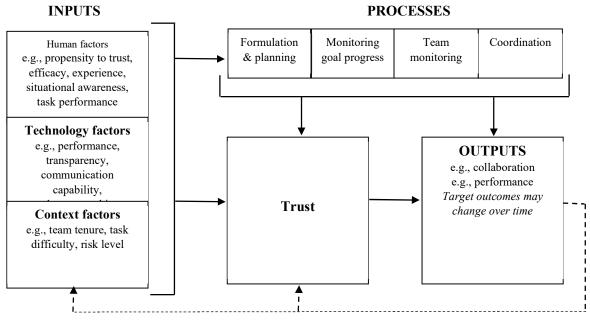


Figure 1: The full Recurring Phase Framework of Trust in Collaborative Intelligence

5 IMPLICATIONS AND APPLICATIONS OF THE RECURRING PHASE FRAMEWORK

5.1 Theoretical Implications

The incorporation of team processes and temporal dynamics permits the integration of a range of empirical findings into a single unifying framework of trust. It is widely recognized, for example, that transparency plays a role in trust formation and maintenance, but that the nature of the relationship between them is not straightforward [22], [26]. While transparency may foster trust under some circumstances, at other times greater transparency may interact with human cognitive capacity with the effect of ultimately reducing trust [27]. The trust framework allows for transparency to operate via different processes and with different effects at different stages in the life cycle of the team. For example, the contribution of transparency to team monitoring processes may increase trust with beneficial outcomes for collaboration in the early 'teambuilding' phases of the human-CINTEL relationship. At later stages when team performance becomes a higher priority outcome, transparency information may be of lower importance to both trust and team functioning.

In addition to incorporating existing findings, the framework may also be used as a foundation to investigate and unravel outstanding empirical questions. One such question is the nature of individual differences in trust calibration. Numerous researchers have identified that people vary in the circumstances under which they revise their trust in an automated system and the speed with which they do so [12], [28], [29]. Using our framework, changes in trust across performance episodes can be identified, and the relative contribution of various antecedent factors in individual trust trajectories can be quantified.

5.2 Practical Implications

Application of the recurring phase framework of trust in CINTEL will enable active management of the trust process to achieve optimal levels of trust within a given CINTEL team. For any system, optimal levels of trust are not equivalent to maximal levels of trust. Trust in a machine system is considered to have reached its optimal level when the human trustor's expectations are aligned with the capabilities of the system [12], [30]. When optimal levels of trust are exceeded, that is, when humans expect more of a system than its capabilities warrant, the resulting overreliance and lack of monitoring can have catastrophic outcomes. A failure to reach optimal levels of trust, on the other hand, can result in disuse of technological systems, representing a waste of time and resources [31]. The field currently lacks well-developed methods for aligning expectations with performance, and identifying when optimal levels of trust have been achieved [12], [30]. The recurring phase framework of trust in CINTEL advances these efforts in a number of ways. The first is by capturing changes in trust levels over time. In doing so, it permits the identification of factors in the interaction between human and machine that increase or decrease levels of trust. Secondly, the framework explicitly identifies team-work processes expected to facilitate this calibration of expectation and performance, including the monitoring of progress towards goals, monitoring the performance of team members (including the CINTEL system), and communication of current status. In this way, the framework can be drawn on to inform the design of system capabilities, collaborative workflows, and training of human team members in a way that supports trust calibration.

6 CONCLUSION

Trust has been shown to be integral to human interaction with automation and AI. Models of trust in A/AI predominantly draw on a tripartite structure of antecedent categories inherited from investigations of trust among humans; human or trustor characteristics, technology or trustee characteristics, and characteristics of the environment or context. These models are limited in their capacity to adapt to specific technological applications, to represent the effect of interaction or

teaming between human and machine, and to capture the temporal dynamics of trust. Such limitations are particularly significant when exploring the nature and function of trust in CINTEL systems.

CINTEL represents a new approach to human-AI teaming that complements the advantages of traditional AI with the unique strengths of human capabilities and values in a process of augmentation rather than automation. Human trust in CINTEL may be more likely to draw on the dynamics of human-human trust to a greater extent than human trust in other technologies. The recurring phase framework of trust in CINTEL is intended to both address the limitations of earlier models of trust in A/AI, and respond to the particular requirements of collaborative human-technology teams. It is primarily distinguished from these earlier models of trust by its incorporation of teamwork processes and recurring performance phases, across which trust and its outputs are influenced by previous team experiences and outcomes. These features enable active management of trust calibration in CINTEL teams, with practical implications for the design of team processes, system capabilities, and human training. In doing so, the recurring phase framework of trust in CINTEL offers a single unifying approach to modelling the trust necessary for human-AI collaboration across a range of applications.

REFERENCES

- C. Paris and A. Reeson, 'What's the secret to making sure AI doesn't steal your job? Work with it, not against it.', The Conversation, Nov. 30, [1] 2021. https://theconversation.com/whats-the-secret-to-making-sure-ai-doesnt-steal-your-job-work-with-it-not-against-it-172691 (accessed Dec. 12,
- B. N. Patel et al., 'Human-machine partnership with artificial intelligence for chest radiograph diagnosis', Npj Digit. Med., vol. 2, no. 1, Dec. [2] 2019, doi: 10.1038/s41746-019-0189-7.
- [3] P. J. Phillips et al., 'Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms', Proc. Natl. Acad. Sci., vol. 115, no. 24, pp. 6171–6176, Jun. 2018, doi: 10.1073/pnas.1721355115.
- M. Steyvers, H. Tejeda, G. Kerrigan, and P. Smyth, 'Bayesian modeling of human-AI complementarity', 2022, doi: 10.1073/pnas
- K. A. Hoff and M. Bashir, 'Trust in automation: Integrating empirical evidence on factors that influence trust', Hum. Factors, vol. 57, no. 3, 2015, [5] doi: 10.1177/0018720814547570.
- J. D. Lee and K. A. See, 'Trust in automation: Designing for appropriate reliance', Hum. Factors J. Hum. Factors Ergon. Soc., vol. 46, no. 1, pp. [6] 50-80, 2004, doi: 10.1518/hfes.46.1.50_30392.
- S. Castaldo, K. Premazzi, and F. Zerbini, 'The meaning(s) of trust. A content analysis on the diverse conceptualizations of trust in scholarly research on business relationships', *J. Bus. Ethics*, vol. 96, no. 4, 2010, doi: 10.1007/s10551-010-0491-4. [7]
- [8] R. C. Mayer, J. H. Davis, and F. D. Schoorman, 'An integrative model of organizational trust', Acad. Manage. Rev., vol. 20, no. 3, pp. 709-734, 1995, doi: 10.5465/AMR.1995.9508080335.
- [9] G. M. Alarcon et al., 'The effect of propensity to trust and perceptions of trustworthiness on trust behaviors in dyads', Behav. Res. Methods, vol. 50, no. 5, pp. 1906–1920, Oct. 2018, doi: 10.3758/s13428-017-0959-6.
- C. Breuer, J. Hüffmeier, F. Hibben, and G. Hertel, 'Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in [10] face-to-face and virtual teams', Hum. Relat., vol. 73, no. 1, pp. 3-34, 2020, doi: 10.1177/0018726718818721.
- F. D. Schoorman, R. C. Mayer, and J. H. Davis, 'An integrative model of organizational trust: Past, present, and future', Acad. Manage. Rev., vol. 32, no. 1, pp. 344-354, 2007.
- [12] K. E. Schaefer, B. S. Perelman, G. M. Gremillion, A. R. Marathe, and J. S. Metcalfe, 'A roadmap for developing team trust metrics for humanautonomy teams', 2020, pp. 261-300. doi: 10.1016/B978-0-12-819472-0.00012-5.
- G. M. Alarcon, A. M. Gibson, and S. A. Jessup, 'Trust repair in performance, process, and purpose factors of human-robot trust', in 2020 IEEE [13] International Conference on Human-Machine Systems (ICHMS), 2020, pp. 1-6. doi: 10.1109/ICHMS49158.2020.9209453.
- P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. De Visser, and R. Parasuraman, 'A meta-analysis of factors affecting trust in human-robot interaction', *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011, doi: 10.1177/0018720811417254.

 A. D. Kaplan, T. T. Kessler, J. C. Brill, and P. A. Hancock, 'Trust in artificial intelligence: Meta-analytic findings', *Hum. Factors J. Hum. Factors* [14]
- [15] Ergon. Soc., p. 001872082110139, May 2021, doi: 10.1177/00187208211013988.
- J. A. Colquitt, B. A. Scott, and J. A. LePine, 'Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with [16] risk taking and job performance', *J. Appl. Psychol.*, vol. 92, no. 4, 2007, doi: 10.1037/0021-9010.92.4.909.

 N. Wang, D. V. Pynadath, and S. G. Hill, 'Trust calibration within a human-robot team: Comparing automatically generated explanations', in
- [17] ACM/IEEE International Conference on Human-Robot Interaction, Apr. 2016, vol. 2016-April, pp. 109-116. doi: 10.1109/HRI.2016.7451741
- [18] K. T. Wynne and J. B. Lyons, 'An integrative model of autonomous agent teammate-likeness', Theor. Issues Ergon. Sci., vol. 19, no. 3, pp. 353-374, 2018, doi: 10.1080/1463922X.2016.1260181.
- [19] E. J. de Visser et al., 'Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams', Int. J. Soc. Robot., vol. 12, no. 2, pp. 459-478, May 2020, doi: 10.1007/S12369-019-00596-X/TABLES/2.

 A. C. Costa, C. A. Fulmer, and N. R. Anderson, 'Trust in work teams: An integrative review, multilevel model, and future directions', *J. Organ.*
- [20] Behav., vol. 39, no. 2, pp. 169-184, Feb. 2018, doi: 10.1002/job.2213.
- [21] T. Saßmannshausen, P. Burggräf, M. Hassenzahl, and J. Wagner, 'Human trust in otherware – a systematic literature review bringing all antecedents together', *Ergonomics*, vol. 0, no. 0, pp. 1–23, Sep. 2022, doi: 10.1080/00140139.2022.2120634.

 A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, 'Toward an understanding of trust repair in human-robot interaction: Current research
- [22] and future directions', ACM Trans. Interact. Intell. Syst., vol. 8, no. 4, Nov. 2018, doi: 10.1145/3181671.
- S. W. J. Kozlowski and G. T. Chao, 'Unpacking Team Process Dynamics and Emergent Phenomena: Challenges, Conceptual Advances, and Innovative Methods', Am. Psychol., 2018, doi: 10.1037/amp0000245.supp.

- [24] M. A. Marks, J. Mathieu, and S. J. Zaccaro, 'A temporally based framework and taxonomy of team processes', Acad. Manage. Rev., vol. 26, no. 3, pp. 356–376, 2001.

 J. A. LePine, R. F. Piccolo, C. L. Jackson, J. E. Mathieu, and J. R. Saul, 'A meta-analysis of teamwork processes: Tests of a multidimensional
- [25] model and relationships with team effectiveness criteria', Pers. Psychol., vol. 61, pp. 273-307, 2008, doi: https://doi.org/10.111/j.1744-6570.2008.00114.x.
- S. Lockey, N. Gillespie, D. Holm, and I. A. Someh, 'A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions', in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2021, vol. 2020-January, pp. 5463–5472. doi: [26] 10.24251/hicss.2021.664.
- M. Dikmen and C. Burns, 'The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending', Int. [27] J. Hum.-Comput. Stud., vol. 162, 2022, doi: 10.1016/j.ijhcs.2022.102792.
- M. D. Coovert, E. E. Pavlova Miller, and W. Bennett Jr., 'Assessing trust and effectiveness in virtual teams: Latent growth curve and latent change score models', Soc. Sci., vol. 6, no. 3, p. 87, 2017. [28]
- J. Lee and N. Moray, 'Trust, control strategies and allocation of function in human-machine systems', *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992, doi: 10.1080/00140139208967392.

 A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, 'Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in [29]
- [30]
- Al', presented at the FAccT, Canada, 2021.
 R. Parasuraman and Riley, Victor, 'Humans and automation: Use, misuse, disuse, abuse', *Hum. Factors*, vol. 39, no. 2, pp. 230–253, 1997, doi: [31] https://doi.org/10.1518/001872097778543886.