PREFACE

Multidisciplinary perspectives on Human-AI team trust

Nicolo' Brandizzi,[1,2] Morgan E. Bailey,[3]
Carolina Centeio Jorge,[4] Myke C. Cohen,[5,6]
Francesco Frattolillo,[7] and Alan R. Wagner[8]
[1] Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) | [2] Lamarr Institute for Machine Learning and Artificial Intelligence | [3] University of Glasgow | [4] Delft University of Technology | [5] Arizona State University | [6] Aptima, Inc. | [7] Sapienza University of Rome | [8] The Pennsylvania State University

1. The need for multidisciplinary perspectives on human-AI team trust

Artificial intelligence is no longer confined to isolated tools or stand-alone systems; it increasingly functions as an active participant in collaborative settings (Seeber et al., 2020). From healthcare and business decision-making to defense operations and everyday workplace environments, humans now share tasks, responsibilities, and risks with AI agents. This shift toward human-AI teams (HATs) offers new opportunities for efficiency, creativity, and problem-solving, but it also raises challenges. Central among these is the question of trust: how, when, and under what conditions humans choose to rely on artificial teammates, and how artificial systems can be designed to act in ways that are worthy of such reliance (Rahwan et al., 2019).

Trust has long been recognized as a critical enabler of cooperation in human teams (Mayer et al., 1995). It shapes coordination, reduces uncertainty, and provides the basis for delegation and shared decision-making. In HATs, these dynamics become more complex. Unlike traditional automation, contemporary AI agents often act with varying degrees of autonomy (Hoff & Bashir, 2015). More recently, AI capabilities have advanced in interpreting natural language, generating creative solutions, and influencing both task outcomes and the social dynamics of collaboration. Such capabilities amplify both the benefits and the risks of misplaced trust. A failure of trust calibration (either excessive reliance on fallible AI systems or unwarranted skepticism toward reliable ones; Parasuraman and

https://doi.org/10.1075/is.00025.edi
Interaction Studies 26:2 (2025), pp. 151–163. ISSN 1572-0373 | E-ISSN 1572-0381

Manzey, 2010) can undermine individual performance, team effectiveness, and ultimately, safety.

The study of trust in HATs, therefore, demands a multidisciplinary perspective. Psychology and organizational science highlight the cognitive and affective mechanisms through which humans extend or withdraw trust (Glikson & Woolley, 2020). Human-computer interaction and design research focus on how transparency, communication, and interface choices shape user perceptions (Wang et al., 2020). Computer science and engineering explore computational models for estimating, predicting, or simulating trust dynamics (Lu et al., 2009). Philosophy and ethics interrogate questions of responsibility, bias, and fairness in designing trustworthy systems (Jobin et al., 2019). While the concept of trust has been studied for decades across psychology, sociology, computer science, and organizational behavior, its dynamics within mixed human-AI teams remain only partially understood. Unlike dyadic interactions between a single user and a system, teams involve multiple actors — human and artificial — working toward shared goals, negotiating interdependence, and adapting to shifting contexts (Lee & See, 2004). Trust in such settings is not a static attribute but an evolving, multidirectional process that shapes and is shaped by team performance. Thus, despite the breadth of knowledge about trust, research progress often proceeds in parallel across fields, with limited multidisciplinary integration. What is lacking are consolidated perspectives that treat trust not as an isolated variable, but as a dynamic, relational, and context-dependent phenomenon.

The central aim of this special issue is to provide a platform where diverse perspectives on human-AI team trust can converge. By bringing together contributions from multiple disciplines, this special issue seeks to advance both conceptual understanding and practical guidance on trust in HATs. The articles collected here demonstrate how different research traditions approach the same fundamental question: how can we build, measure, and appropriately sustain trust when humans and artificial agents work side by side? We consider the contributions of this special issue around four pressing challenges in HAT trust research: (1) trust development and measurement; (2) accounting for the complexity of psychological and interaction dynamics; (3) transparency, communication, and design; and (4) ethics, bias, and organizational considerations.

1.1 Trust development and measurement

Trust in HATs commonly looks at the attitude of humans and understanding that an agent will help achieve an individual's goals even when under a vulnerability level (Lee & See, 2004), highlighting its role in guiding decision-making. reliance. Crucially, trust is not static; it develops over time and can fluctuate with

experience. Early interactions may breed overconfidence in a new AI teammate's capabilities, followed by a calibration perio d as the human learns the AI's actual reliability (Schmutz et al., 2024). Such dynamics underscore the need to measure and manage trust continually.

Researchers have proposed various empirical measures (e.g. psychometric scales and behavioral indicators) to quantify a user's trust level and track how it changes with system performance. These efforts are motivated by findings that an appropriately calibrated trust can enhance team effectiveness. For example, in a decision-aid study, humans with higher trust in an AI agent achieved greater performance gains than those with lower trust (Carragher et al., 2024). Conversely, miscalibrated trust, whether excessive or insufficient, can lead to misuse of automation or disuse, ultimately harming joint human-AI performance (Wen et al., 2025). Ensuring that trust develops in step with an AI's true capabilities, often termed trust calibration, has therefore become a key goal for HAT designers, as calibrated trust is linked to more optimal reliance and better task outcomes (Carragher et al., 2024; Lee & See, 2004).

1.2 Psychological and interaction dynamics

Human cognition, perception, and emotion all influence and are influenced by AI teammates' behaviors during collaboration. Trust in an AI agent has both cognitive foundations (e.g. beliefs about the AI's competence) and affective components (e.g. feelings of security or rapport), similar to interpersonal trust (Lee & See, 2004; Malle & Ullman, 2021). Studies show that people develop trust toward AI team partners differently than toward human partners. For instance, participants tend to report lower cognitive trust in an AI teammate and weaker team identification, compared to teams with only humans (Georganta and Ulfert, 2024; Musick et al., 2021; cf. McNeese et al., 2021). Recent works suggest that these differences are more profound along affective dimensions (Duan et al., 2025; Georganta & Ulfert, 2024), as predicted by theoretical models of human trust in automation (Madhavan & Wiegmann, 2007; Malle & Ullman, 2021).

Psychological nuances can impact team coordination (and therefore team performance): if a human feels uncertain or uneasy about an AI's decisions, they may unnecessarily hesitate to align with the AI's actions, disrupting HAT functioning (Snow, 2021). Indeed, adding an AI member to a team can introduce communication challenges and reduce overall cohesion until the human members adjust (Schmutz et al., 2024). On the other hand, when the human understands the AI's working style and the AI meets the human's trust expectations, a shared mental model can develop, allowing for smoother collaboration. High trust has been linked to better information exchange and collective problem-

solving in teams (Duan et al., 2025). A reservoir of trust can make the team more resilient. When errors or surprises occur, a trusted partnership is more likely to weather the incident without total breakdown, enabling the human-AI team to adapt and continue performing effectively. Because AI teammates are bound to fail at some point, interaction dynamics must reinforce this psychological adaptability (Naikar et al., 2025). Thus, the interplay of cognitive appraisals and emotional responses towards AI partners fundamentally shapes how well humans and AI coordinate their actions and maintain robust teamwork under stress.

1.3 Transparency, communication, and design

The design of an AI teammate and how it communicates are pivotal in fostering trust. Transparent systems that provide interpretable insights into their reasoning or uncertainty tend to be perceived as more trustworthy (Wen et al., 2025). By contrast, "black-box" AI behavior can leave users guessing at motives or reliability, which undermines trust. Effective communication strategies include not only explaining decisions but also acknowledging mistakes. An AI agent that can admit errors or express self-confidence appropriately helps calibrate the human's trust — research indicates that users value an AI agent's willingness to recognize and correct its faults, which in turn builds trust resilience (i.e. trust that endures small failures; Afroogh et al., 2024).

Design cues and interaction modalities also play a role. For example, giving an AI human-like characteristics or social behaviors can influence users' trust. Studies have found that anthropomorphic agents (e.g. using a human avatar or conversational style) are associated with greater resistance to trust breakdowns when errors occur (de Visser et al., 2016). Such human-like communication and trust repair behaviors (e.g. apologies, assurances) can make the AI agent seem a more relatable teammate, thereby sustaining trust through difficulties (de Visser et al., 2020). However, designers must strike a balance: over-humanizing an AI or conveying an illusion of infallibility can backfire. Users may develop unrealistically high expectations and then feel betrayed by any mistake (Shneiderman, 1989). Recent works suggest the importance of transparency in communicative AI agents; i.e., they must explicitly set accurate expectations, explain their actions, and engage human counterparts in a collaborative dialogue. At the same time, these options are limited by design guidelines from industry (e.g., Cohen et al., 2025) and regulatory bodies (e.g., EU AI Act, 2021). Although transparency and communication remain important, technologically-oriented perspectives anchored on these alone are insufficient for HAT trust research and design.

1.4 Ethics, bias, and organizational considerations

Trust in AI systems is deeply entwined with ethical and societal factors. If an AI agent is perceived as biased or unfair, trust can evaporate quickly, as users recoil from decisions that violate social norms of justice (Afroogh et al., 2024; Rezaei Khavas et al., 2024). Empirical work confirms that algorithmic discrimination undermines trust, whereas demonstrable fairness can enhance it (Afroogh et al., 2024). This has led to a strong intersection of trust research with AI ethics and policy: frameworks for "Trustworthy AI" typically encompass principles like fairness, accountability, transparency, and safety, on the premise that these qualities are prerequisites for justified trust (Afroogh et al., 2024; Ryan, 2020).

From an organizational perspective, businesses recognize that maintaining user and stakeholder trust is key to the sustainable adoption of AI technologies. Scandals or failures involving biased AI outputs or opaque algorithms not only create public backlash but also erode the confidence of employees and decision-makers in deploying AI tools. In contrast, aligning AI systems with ethical guidelines and openly addressing their limitations builds a reputation of trustworthiness that can differentiate companies in the marketplace. Moreover, industry standards and regulations are increasingly requiring evidence of bias mitigation and transparency, effectively institutionalizing trust-related criteria. Ultimately, the long-term integration of AI into society hinges on trust: people need to feel that AI agents are not only effective but also aligned with human values and expectations. Achieving this means rigorously evaluating AI for unfair biases, respecting privacy and consent, and being transparent about how decisions are made. Organizations that prioritize these ethical considerations tend to foster greater trust in the deployment of AI systems, which in turn facilitates user acceptance and successful team integration (Afroogh et al., 2024). In sum, the pursuit of trustworthy AI is as much a sociotechnical endeavor as a technical one; it requires balancing business objectives with ethical responsibility to ensure that human-AI team trust can be sustained in real-world applications.

2. Overview of contributions

The seven papers in this special issue reflect the diversity of approaches and perspectives that characterize the emerging field of human-AI team trust. Taken together, they demonstrate how trust can be studied as a psychological construct, a computational phenomenon, an organizational requirement, and a societal concern. We group these seven articles into three clusters: (1) trust signals and com-

munication; (2) computational and modeling approaches; and (3) conceptual, ethical, and organizational perspectives.

2.1 Trust signals and communication

Two contributions considered the linguistically complex relationship between people's tendencies to treat AI teammates as social actors, both as initiators and end-recipients of human-AI communication loops.

Bailey et al. (2026) examine how emojis and AI reliability shape trust and team performance. Through an experimental study, they show that emotional cues embedded in communication can influence perceptions of AI teammates, highlighting the role of seemingly minor signals in trust calibration. Similarly, Cohen et al. (2026) explore how personifying and objectifying language can function as subtle indicators of trust in team communication with AI teammates. Their findings suggest that people's use of objectifying language can indicate perceptions of poor trust and low anthropomorphism. However, they also noted that personifying language does not exhibit the opposite correlations.

Together, these studies indicate that trust in HATs is not only a matter of technical reliability but also of how communication is framed and interpreted. Emojis, metaphors, and linguistic choices may appear peripheral, yet they shape whether an AI is perceived as a capable partner or a fallible tool. Language and symbols can have a crucial role in how trust develops, fluctuates, or erodes within HATs.

2.2 Computational and modeling approaches

The second cluster demonstrates how computational methods can advance the study of trust in human-AI teams.

Kucukosmanoglu et al. (2026) investigate the application of fine-tuned natural language processing (NLP) models to analyze sentiment in expert military team communications during AI-supported combat simulations. By adapting DistilBERT and GPT-2 to the military context, they improve sentiment classification accuracy and show that sentiment measures correlate with both team trust and trust in AI systems. Their work highlights how NLP techniques can be adapted to specialized domains to provide unobtrusive indicators of trust dynamics. Conversely, Nguyen et al. (2026) explore the potential of LLM-based human digital twins (HDTs) that approximate human socio-emotional and cognitive reactions for modeling trust in HATs. They examine how trust can be operationalized and measured in HDT experiments, compare different LLMs for their ability to simulate trust-related dynamics, and propose experimental designs that incorporate individual trust tendencies alongside AI transparency and competence. Their

results suggest that HDTs could serve as computational proxies for investigating how trust emerges and fluctuates in teaming contexts.

These two papers illustrate how computational approaches — whether through domain-adapted NLP models or human digital twin simulations — can open new avenues for studying trust at scale, complementing traditional behavioral and self-report methods.

2.3 Conceptual, ethical, and organizational perspectives

The third set of contributions broaden the perspective on trust, situating it within conceptual, ethical, and organizational contexts. These papers demonstrate that trust in HATs is not only a matter of micro-level interactions but also a socio-technical concern. Conceptual clarity, ethical evaluation, and organizational practices are all necessary to ensure that trust is justified and sustainable as AI systems become embedded in high-stakes domains.

Momen et al. (2026) investigate how people perceive the moral competence of generative AI systems, using Delphi AI (Jiang et al., 2021) as a test case. Participants evaluated the AI when embodied as a humanoid robot or deployed as a web client, rating it highly for moral competence and perceived morality. Yet, the absence of justifications for its judgments tempered participants' willingness to rely on it, suggesting the limitations of ethical robot advisors. Coester et al. (2026) turn to the organizational adoption of AI, reporting findings from the TrustKI project. Their study shows that users demand holistic transparency: technical information and evidence of expertise alone are insufficient. Instead, trust also depends on contextual information about the company providing the AI system. From these findings, they propose an extended set of information requirements that providers should meet to document AI trustworthiness more effectively. Finally, Tielman et al. (2026) present outcomes from three editions of the MULTITTRUST workshop series that inspired the current special issue, synthesizing insights across psychology, computer science, philosophy, and human-computer interaction. Their article provides a shared language of key concepts and definitions, identifies major open challenges, and offers methodological guidelines. These elements form a foundational roadmap for advancing research on trust in human-AI team interactions.

Altogether, these seven contributions highlight the breadth of current research on human-AI team trust, ranging from micro-level interaction cues to large-scale organizational and societal implications, and spanning both theoretical and applied works. As a collective, they show the necessity of a multidisciplinary combination of empirical, computational, and conceptual approaches to

capture human-AI team trust as a multifaceted and dynamic phenomenon across application domains.

3. Looking forward

The papers in this special issue illustrate that trust in human-AI teams is a complex, dynamic phenomenon, shaped by communication cues, computational reasoning, and organizational and ethical considerations. The relevance of human-AI team trust is growing in parallel with the rapid deployment of AI across domains. In healthcare, the UK's National Health Service has begun piloting AI systems to accelerate patient discharges and monitor safety concerns, with the promise of reducing strain on staff and improving outcomes (The Guardian, 2025b). Yet the same technology also raises concerns about trust calibration: clinicians must weigh AI advice against professional judgment, knowing that both over-reliance and undue skepticism can harm patients. Broader policy analyses underline that trust and adoption are closely linked in the deployment of AI systems, stressing that public confidence is essential for successful integration (European Commission, 2021; High-Level Expert Group on Artificial Intelligence, 2019). Looking ahead, research should continue to integrate these perspectives, developing AI agents that can adapt to human teammates' characteristics, such as skills and preferences while maintaining transparency and interpretability.

Future research should keep identifying behavioural cues that can indicate human trust and trustworthiness during interaction, as well as algorithms that allow for contextual and dynamic trust learning over time. All this must support human agency without imposing excessive cognitive or operational effort. Advances in adaptive interfaces, guided communication, and context-sensitive models offer promising avenues to address this tension and can be brought to the future research of human-AI team trust. For example, in mental health, therapists warn that unregulated AI chatbots risk creating emotional dependency or amplifying anxiety, highlighting the dangers of misplaced trust in seemingly empathic systems (The Guardian, 2025c). Commentators have gone further, comparing AI-driven health advice to a new form of digital quackery, where plausible but unverified outputs are mistaken for scientific truth (The Guardian, 2025a). Such examples illustrate the risks that international policy frameworks have sought to pre-empt by stressing transparency, accountability, and human oversight in AI deployment (Tabassi, 2023; UNESCO, 2021). Similar tensions are visible in business and education. Small companies remain hesitant to integrate AI into sensitive processes, citing uncertainty over reliability and data security (The Guardian,

2025d). In schools, leaders have voiced unease about the pace of AI adoption, emphasizing a lack of trust in technology firms to safeguard student welfare (The Guardian, 2023).

Finally, human-AI team trust cannot be studied in isolation from societal and organizational contexts. Ensuring that AI agents support collaboration without fostering over-reliance, exclusion, or unintended consequences will require ongoing attention to ethics, regulation, and human well-being. Future work should strive for AI that is not only effective but also socially responsible and human-centred, which is only possible by keeping efforts towards multidisciplinary theory building.

We conclude by reiterating a common theme across the contributions in this special issue: trust is not a secondary issue but a central condition for the responsible use of AI. When trust fails the promised benefits of human-AI collaboration collapse. At present, research on trust remains fragmented, with psychology, engineering, ethics, and organizational science pursuing parallel directions. The timeliness of this special issue lies in bridging these disciplinary gaps. By drawing together diverse perspectives, it situates trust as a foundational concern for the next phase of AI integration in society.

References

Afroogh, S., Akbari, A., Malone, E., & Langarizadeh, M. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11* (1), 1568.

Bailey, M. E., Gancz, B., & Pollick, F. E. (2026). The effect of emojis and AI reliability on team performance and trust in human-AI teams. *Interaction Studies*, (Special Issue on Multidisciplinary Perspectives on Human-AI Team Trust), *26* (2).

Carragher, D. J., Sturman, D., & Hancock, P. J. (2024). Trust in automation and the accuracy of humanalgorithm teams performing one-to-one face matching tasks. *Cognitive Research: Principles and Implications*, *9* (41).

Coester, U., Anderle, L., & Pohlmann, N. (2026). Trustworthiness needs for the use of AI solutions in business: Ethical and empirical considerations. *Interaction Studies*, (Special Issue on Multidisciplinary Perspectives on Human-AI Team Trust), *26* (2).

Cohen, M. C., Chiou, E. K., & Cooke, N. J. (2026). Trusting machine teammates: The role of personifying and objectifying language in team communication. *Interaction Studies*, (Special Issue on Multidisciplinary Perspectives on Human-AI Team Trust), *26* (2).

Cohen, M. C., Kim, N., Ba, Y., Pan, A., Bhatti, S., Salehi, P., Sung, J., Blasch, E., Mancenido, M. V., & Chiou, E. K. (2025). PADTHAI-MM: Principles-based approach for designing trustworthy, human-centered AI using the MAST methodology. *AI Magazine*, *46*(1), e70000.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams. *International Journal of Social Robotics*, *12* (2), 459–478.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22* (3), 331–349.

Duan, W., Zhou, S., Scalia, M. J., Freeman, G., Gorman, J., Tolston, M., McNeese, N. J., & Funke, G. (2025). Understanding the processes of trust and distrust contagion in human-ai teams: A qualitative approach. *Computers in Human Behavior*, *165*, 108560.

European Commission. (2021). *Coordinated plan on artificial intelligence 2021 review* (tech. rep.) (Accessed: 2025-09-04). European Commission. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

Georganta, E., & Ulfert, A.-S. (2024). Would you trust an ai team member? team trust in human-ai teams. *Journal of Occupational and Organizational Psychology*, *97* (3), 1212–1241.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, *14* (2), 627–660.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy ai* (tech. rep.) (Accessed: 2025-09-04). European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57* (3), 407–434.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., & Choi, Y. (2021). Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature machine intelligence*, *1* (9), 389–399.

Kucukosmanoglu, M., Johnson, C. J., Pollard, K., Chhan, D., Lakhmani, S. G., Forster, D., Conklin, S., Brooks, J., Crowell, H. P., & Krausman, A. (2026). Exploring trust in AI-supported military teams using sentiment analysis. *Interaction Studies*, (Special Issue on Multidisciplinary Perspectives on Human-AI Team Trust), *26* (2).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46* (1), 50–80.

Lu, G., Lu, J., Yao, S., & Yip, Y. J. (2009). A review on computational trust models for multi-agent systems. *The open information science journal*, *2*, 18–25.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automat trust: An integrative review [Publisher: Taylor & Francis _eprint: *Theoretical Issues in Ergonomics Science*, *8* (4), 277–301.

Malle, B. F., & Ullman, D. (2021, January). Chapter 1 — A multidimensional conception and measure of humanrobot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 3–25). Academic Press.

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

McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Trust and Team Performance in Human-Autonomy Teaming [Publisher: Routledge _eprint:]. *International Journal of Electronic Commerce*, *25* (1), 51–72.

Momen, A., Tossell, C. C., Walliser, J. C., Niemyer, R., Tolston, M., Funke, G. J., & de Visser, E. J. (2026). Perceived trustworthiness and moral competence of a GenAI-enabled ethical robot advisor. *Interaction Studies*, (Special Issue on Multidisciplinary Perspectives on Human-AI Team Trust), *26* (2).

Musick, G., O'Neill, T. A., Schelble, B. G., McNeese, N. J., & Henke, J. B. (2021). What Happens When Humans Believe Their Teammate is an AI? An Investigation into Humans Teaming with Autonomy. *Computers in Human Behavior*, *122*, 106852.

Naikar, N., Hoffman, R., Roth, E. M., Klein, G., Militello, L. G., & Dominguez, C. (2025). Should we Make AI More Tool-like or Teammate-Like? [Publisher: SAGE Publications]. *Journal of Cognitive Engineering and Decision Making*, 15553434251346904.

Nguyen, D., Cohen, M. C., Kao, H.-T., Engberon, G., Penafiel, L., Lynch, S., & Volkova, S. (2026). Exploratory models of human-AI teams: Leveraging human digital twins to investigate trust development. *Interaction Studies*, (Special Issue on Multidisciplinary Perspectives on Human-AI Team Trust), *26* (2).

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration [Publisher: SAGE Publications Inc]. *Human Factors*, *52* (3), 381–410.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, *568* (7753), 477–486.

Rezaei Khavas, Z., Kotturu, M. R., Ahmadzadeh, S. R., & Robinette, P. (2024). Do Humans Trust Robots that Violate Moral Trust? *J. Hum.-Robot Interact.*, *13* (2), 25:1–25:30.

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26* (5), 2749–2767.

Schmutz, J. B., Outland, N., Kerstan, S., Georganta, E., & Ulfert, A.-S. (2024). Ai-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology*, *58*, 101837.

Seeber, I., Bittner, E., Briggs, R. O., De Vreede, T., De Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-ReiB, S., Randrup, N., et al. (2020). Machines as teammates: A research agenda on ai in team collaboration. *Information & management*, *57* (2), 103174.

Shneiderman, B. (1989). A nonanthropomorphic style guide: Overcoming the humpty-dumpty syndrome. *The Computing Teacher*, *16* (7), 5.

Snow, T. (2021). From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm [Publisher: Cambridge University Press]. *Data & Policy*, *3*, e3.

Tabassi, E. (2023, 2023-01-26 05:01:00). Artificial intelligence risk management framework (ai rmf 1.0).

The Guardian. (2023, May). Uk schools bewildered by ai and do not trust tech firms, headteachers say [Accessed: 2025-09-04].

The Guardian. (2025a, July). Medical charlatans have existed through history. *but ai has turbocharged them* [Accessed: 2025-09-04].

The Guardian. (2025b, August). Nhs to trial ai tool that speeds up hospital discharges [Accessed: 2025-09-04].

The Guardian. (2025c, August). Therapists warn ai chatbots risk harming mental health support [Accessed: 2025-09-04].

The Guardian. (2025d, May). Yes, ai will eventually replace some workers. but that day is still a long way off [Accessed: 2025-09-04].

Tielman, M., Bailey, M. E., Frattolillo, F., Centeio Jorge, C., Ulfert, A.-S., & Meyer-Vitali, A. (2026). Multidisciplinary Perspectives on Human-AI Team Trust. *Interaction Studies*.

UNESCO. (2021). *Recommendation on the ethics of artificial intel ligence* (tech. rep.) (Accessed: 2025-09-04). United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 1–6.

Wen, Y., Wang, J., & Chen, X. (2025). Trust and ai weight: Human-ai collaboration in organizational management decision-making. *Frontiers in Organizational Psychology*, *3*, 1419403.

Address for correspondence

Nicolo' Brandizzi
Schloss Birlinghoven
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS Schloss Birlinghoven
Konrad-Adenauer-Straße 1
53757 Sankt Augustin
Germany
nicolo.brandizzi@iais.fraunhofer.de

Biographical notes

Nicolo' Brandizzi is a Research Scientist at Fraunhofer IAIS in Bonn, Germany. His work focuses on large-scale language model development, data governance, and reproducibility in European AI initiatives such as OpenGPT-X, EuroLingua, TrustLLM, and DKZ.2R. He cleads the Fraunhofer team for the JUPITER AI Factory and contributes to the Lamarr Institute. Previously, he conducted research on human-AI interaction and model-based reinforcement learning at Sapienza University of Rome, where he earned his Ph.D. in Computer Science Engineering. His publications span language modeling, reinforcement learning, and human-AI trust, and he serves on the steering committee of the MultiTTrust workshop series.

Morgan E. Bailey received her PhD from the University of Glasgow, she currently works as Behavioral Scientist at Social Machine LtD. Her Ph.D. work focused on the forming of Human-AI Teams and how anthropomorphism, emotional intelligence and AI framing impact trust and performance in these teams. Currently her work focuses on ensuring calibrated trust in

human-AI teams and rebuilding trust after AI failure. Morgan also holds a MSc in Psychological Research Methods from the University of Plymouth.
m.bailey.1@research.gla.ac.uk

Carolina Centeio Jorge is Senior Data Scientist at Glintt Global, in Portugal, and a Ph.D. candidate at Delft University of Technology, in the Netherlands. Her Ph.D. focuses on mental models in the context of human-AI teams, such as enabling artificial agents to understand and predict human teammates and effectively act accordingly. In particular, she explored artificial trust, i.e., how an AI should trust a human teammate. Carolina previously received her MSc in Computer Science and Engineering from University of Porto, in Portugal. During her MSc, Ph.D., and in between, she spent periods of her life studying and/or working in different countries such as Spain, The Netherlands, Japan, and USA.
c.jorge@tudelft.nl

Myke C. Cohen is a PhD student in Human Systems Engineering at Arizona State University and an Associate Scientist at Aptima, Inc. His research focuses on the collective decision-making processes, social dynamics, ethics, and design considerations in collaborations involving people and technology. He holds an MS in Human Systems Engineering from Arizona State University and a BS in Industrial Engineering from the University of the Philippines Diliman.
myke.cohen@asu.edu

Francesco Frattolillo received his Ph.D. in Computer Engineering from Sapienza University of Rome. His research focuses on reinforcement learning (RL) in multi-agent systems, with an emphasis on improving sample efficiency and coordination. He has developed hierarchical and model-based solutions to enhance multi-agent RL performance. Additionally, he explores RL applications in hybrid human-agent teams, where he is particularly interested in formalizing trust dynamics and designing trust-aware RL algorithms to improve collaboration and decision-making. Before, he received his MSc in Artificial Intelligence and Robotics from Sapienza.
frattolillo@diag.uniroma1.it

Alan R. Wagner is an associate professor in the department of Aerospace Engineering and a senior research associate with The Rock Ethics Institute at The Pennsylvania State University.
alan.r.wagner@psu.edu