Designing Trustworthy AI Systems for Human-Centric Collaboration

Mario Mirabile[a,b,*]

[a]University of Santiago de Compostela, International PhD School (EDIUS)
[b]University of Bologna, Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI)

Abstract. Artificial Intelligence (AI) systems are increasingly embedded in complex, high-stakes sectors such as healthcare, finance, and telecommunications, fundamentally altering the nature of human-AI collaboration. Although intelligent technologies hold great promise for efficiency and innovation, their widespread adoption faces critical challenges, notably around the establishment and maintenance of human trust. Trust is widely recognised as a cornerstone of effective human-AI interactions, as users must rely on AI systems to perform reliably, ethically, and transparently. However, trust in AI remains inadequately understood, although an increasing number of scholars are actively investigating its definitions, dimensions, and implications. Addressing these challenges is particularly timely given regulatory developments, such as the EU AI Act, which prioritise transparency, explainability, fairness, and accountability. There is a compelling need for standardised frameworks and tools to systematically understand, evaluate, and improve trustworthiness in human-AI collaborations. This doctoral research aims to conceptually unpack trust, develop standardised evaluation metrics, and propose actionable design strategies for trustworthy AI systems, ultimately fostering more effective and widely accepted human-AI collaborations.

1 Introduction and motivation

Artificial Intelligence (AI) systems are increasingly embedded in complex high-stakes sectors such as healthcare, finance, and telecommunications, fundamentally altering the nature of human-AI collaboration. Although intelligent technologies hold great promise for efficiency and innovation, their widespread adoption faces critical challenges, notably around the establishment and maintenance of human trust. Trust is widely recognised as a cornerstone of effective human-AI interactions, as users must rely on AI systems to perform reliably, ethically and transparently [1, 3, 5, 6, 7].

However, trust in AI remains inadequately understood, although an increasing number of scholars are actively investigating its definitions, dimensions, and implications in various contexts and user groups, as efforts continue to overcome barriers to adoption and make everyone aware of the potential benefits of AI [2, 4, 7, 8]. Addressing these challenges is particularly timely given regulatory developments, such as the EU AI Act, which explicitly prioritise transparency, explainability, fairness, and accountability in AI systems. As intelligent technologies increasingly mediate critical decisions in human-centric contexts, there is a compelling need for standardised frameworks and tools to systematically understand, evaluate, and enhance trustworthiness in human-AI collaborations.

2 Research questions

The doctoral research aims to address the following questions:

1. How does the scientific community currently define and conceptualise trust within the context of human-AI interaction?
2. What tools and methodologies already exist for measuring and evaluating trustworthiness in human-AI team collaboration scenarios?
3. What design principles and strategies can effectively enhance trustworthiness in AI systems, taking into account diverse user needs and contexts?

3 Methodology

This research adopts a Design Science approach, systematically integrating interdisciplinary insights from cognitive psychology, ethics, engineering, and management.

- **Phase 1 - Systematic Literature Review (current):** the first phase involves conducting a comprehensive systematic literature review following PRISMA guidelines to identify influential concepts and thematic clusters.
- **Phase 2 - Multi-Method Validation:** subsequent phases will employ both qualitative and quantitative methods to validate the conceptual understanding. Qualitative techniques include structured interviews with AI developers and end-users, and stakeholder workshops using thematic analysis. Quantitative methods encompass experimental designs manipulating trust antecedents (explanation detail, anthropomorphic features), behavioural data collection tracking actual reliance patterns versus self-reported trust, and longitudinal studies examining trust evolution through formation, violation, and repair cycles.
- **Phase 3: Iterative Prototyping:** this phase focuses on developing AI agents incorporating Human-Centric XAI techniques (LIME, SHAP), fairness-aware algorithms, and adaptive user interfaces. The prototype wants to demonstrate how targeted explanations and adaptive interface designs can tangibly enhance user trust across different expertise levels and contexts.

[*] Corresponding Author. Email: mario.mirabile@rai.usc.es.

4 Preliminary results and contributions

Initial findings from the systematic literature review with a focus on financial contexts indicate substantial variability and a lack of standardisation in the conceptualisation and measurement of trust between disciplines.

4.1 Interdisciplinary conceptualisations of trust and methods to measure it

The ongoing analysis is revealing trust operates through distinct but interconnected dimensions:

1. **Cognitive Trust** is conceptualized as belief in AI competence and reliability, measured primarily through Technology Acceptance Model (TAM) constructs and Structural Equation Modeling (SEM).
2. **Affective Trust**, defined as emotional connection to AI systems, emerges particularly through anthropomorphic design and is measured via warmth/competence perceptions scales and emotional response questionnaires. Experimental studies manipulate social cues (voice, avatar appearance) to assess impact.
3. **Procedural Trust**, grounded in transparency and explainability, is measured through user comprehension tests of AI explanations, decision confidence ratings, and task performance metrics when using XAI tools like LIME or SHAP.
4. **Infrastructural Trust** emerges from system architecture (blockchain, zero-trust) and is measured through audit trail completeness, verification success rates, and system resilience metrics.

4.2 Key gaps identified in the literature

The review identifies five critical gaps in current research. First, methodological fragmentation is evident in the over-reliance on cross-sectional surveys with limited experimental, or longitudinal designs, preventing causal inference and understanding of trust dynamics over time. Second, there is a level disconnection with minimal integration between micro-level user perceptions, meso-level organizational practices, and macro-level governance mechanisms. Studies typically focus on single levels without examining cross-level interactions. Third, the field suffers from a lack of measurement standardization, with an absence of validated multilevel instruments. Existing tools fail to capture trust's context-dependent and dynamic nature, with most scales adapted from interpersonal trust measures not validated for AI contexts. Fourth, a trust calibration gap reveals limited understanding of appropriate trust levels, as studies focus on increasing trust without addressing risks of over-reliance or the need for calibrated skepticism in high-stakes decisions. Finally, geographic and cultural bias is evident with research concentrated in Western and East Asian contexts, with minimal representation from Global South, limiting generalizability across diverse regulatory and cultural environments.

5 Directions for the remaining work

Building upon these initial contributions, future work will focus on three integrated streams:

5.1 Conceptual framework development (months 1–12)

The immediate priority involves completing the systematic review in finance. This will inform the development of a multi-level socio-technical framework integrating micro (cognitive-affective), meso (organizational design), and macro (infrastructure-governance) dimensions. Concurrently, we will create a standardized trust measurement battery combining self-report, behavioural, and system-level metrics to address the identified measurement gaps.

5.2 Empirical validation (months 12–24)

The validation phase will employ multiple methodological approaches to address the limitations identified in current research. We will conduct 2×2 factorial experiments manipulating explanation detail and user control to identify optimal XAI configurations for different user groups. In parallel, a longitudinal field study in financial advisory contexts will track trust trajectories across a 6-month period, providing insights into trust dynamics over time. This empirical work will support the development of computational models predicting trust formation, calibration, and repair based on system behaviour and user characteristics.

5.3 Design Guidelines and Implementation (Months 24–36)

The final phase focuses on translating research insights into practical applications. This includes creating adaptive multi-agent prototypes with specialized roles (guardian agents for oversight, advisory agents for decision support) that demonstrate trust-enhancing design patterns. The research will culminate in formulating actionable guidelines aligned with EU AI Act requirements for transparency, accountability, and human oversight, alongside developing an open-source toolkit for trustworthy AI assessment available to researchers and practitioners.

6 Conclusion

This doctoral research contributes to the field by systematically unpacking the concept of trust in human-AI interactions, identifying gaps in existing methods, and proposing rigorous design strategies to enhance trustworthiness. Ultimately, the research seeks to foster AI systems that align with ethical standards, regulatory frameworks, and diverse human needs, thus enabling more effective and widely accepted human-AI collaborations.

Acknowledgements

The author gratefully acknowledges the guidance and support of the supervisor, Prof. José Maria Alonso-Moral, and colleagues from the University of Santiago de Compostela and the University of Bologna.

References

[1] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99: 101805, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101805. URL https://www.sciencedirect.com/science/article/pii/S1566253523001148.

[2] R. Confalonieri and J. M. Alonso-Moral. An Operational Framework for Guiding Human Evaluation in Explainable and Trustworthy Artificial Intelligence . *IEEE Intelligent Systems*, 39(01):18–28, Jan. 2024. ISSN 1941-1294. doi: 10.1109/MIS.2023.3334639. URL https://doi.ieeecomputersociety.org/10.1109/MIS.2023.3334639.

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

[4] R. R. Hoffman. Trusting as an emergent: Implications for design. *Ergonomics in Design*, 0(0):10648046241295270, 0. doi: 10.1177/10648046241295270.

[5] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445923. URL https://doi.org/10.1145/3442188.3445923.

[6] 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 54th Hawaii International Conference on System Sciences*, pages 5463–5472, Honolulu, HI, United States, 2021. Hawaii International Conference on System Sciences. ISBN 9780998133140. doi: 10.24251/hicss.2021.664. URL https://doi.org/10.24251/hicss.2021.664.

[7] D. Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human Computer Studies*, 146, Feb. 2021. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2020.102551. Publisher Copyright: © 2020.

[8] S. Thiebes, S. Lins, and A. Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464, Jun 2021. ISSN 1422-8890. doi: 10.1007/s12525-020-00441-4. URL https://doi.org/10.1007/s12525-020-00441-4.