

# Formal Verification for Human-Centred Trust in AI: A Critical Examination of Current Paradigms

Asieh Salehi Fathabadi

University of Southampton, UK

## ABSTRACT

As artificial intelligence systems increasingly permeate critical societal infrastructures, the gap between technical verification and human-centred trust has become a fundamental challenge. This position paper argues that current formal verification approaches for AI systems are fundamentally inadequate to foster genuine public trust, particularly in settings involving human interaction and socio-technical complexity. We advance three critical arguments: (1) the Trust Verification Paradox: static verification approaches fail to capture the dynamic and adaptive nature of trust; (2) the Public Technical Trust Divide: technical correctness without human understanding risks “certification theatre”; and (3) the Distributed Responsibility Crisis: existing verification paradigms struggle to account for collective outcomes and accountability. We propose a shift toward Participatory Verification, in which formal methods are extended to embed stakeholder values, support verification of trust evolution, and enable responsibility attribution. Through a formal and illustrative autonomous vehicle coordination case study, we demonstrate the expressive power of Participatory Verification and outline how trust evolution, stakeholder values, and responsibility attribution can be embedded into verification frameworks. This vision paper calls for a research agenda that bridges formal methods, human-AI interaction, and social science to support AI systems that are not only technically correct, but genuinely trustworthy.

**Keywords:** Formal methods, Responsible AI, Trust, Human-centred, Participatory design

## INTRODUCTION

Artificial Intelligence systems are embedded in socio-technical environments where humans interact with, rely upon, and are affected by algorithmic decisions. In domains such as autonomous transportation, healthcare, and public infrastructure, AI operates in settings where public trust is essential (Wing, 2021). Yet despite advances in formal verification, public trust in AI remains fragile (Pew Research Center, 2025).

Formal verification has traditionally focused on correctness, safety, and robustness (Clarke, 2018). These approaches prove effective for closed systems with stable specifications. However, when AI systems interact with humans and adapt over time, trust becomes dynamic and socially mediated (Mayer, 1995). Verification guarantees that remain invisible or misaligned with stakeholder concerns fail to translate into trust (Amershi et al., 2019). Consider medical diagnosis AI: a system may be verified for accuracy and

fairness, yet physicians distrust it because they cannot understand how diagnoses are generated, and administrators hesitate because liability remains unclear.

Recent work has begun to address this challenge through socio-technical approaches that integrate formal verification with human trust considerations (Akintunde et al., 2023; Salehi Fathabadi, 2025). Research on defence and security automated systems emphasizes that trust and safety are inseparable, as one study memorably states, “trust equals less death” (Salehi Fathabadi, and Leonard, 2025). However, these efforts remain nascent and have not yet coalesced into systematic verification frameworks for human-centred AI.

Regulatory frameworks such as the EU AI Act (European Parliament, 2024) emphasize compliance and transparency. While necessary, these mechanisms risk conflating compliance with trustworthiness (Wing, 2021). Empirical and design research shows that public trust depends on perceived fairness, transparency, accountability, and the ability to contest decisions (Slovic, 1993, Friedman et al., 2017). These human-centred properties remain largely external to current verification frameworks (Clarke, 2018).

This paper argues that the limitation is structural. Verification has been conceived as a purely technical activity, detached from social contexts where trust forms. Human trust evolves through interaction; stakeholders hold diverse, conflicting values; and verification artifacts remain opaque to non experts. As AI systems become more interactive and consequential, this separation becomes untenable.

We propose Participatory Verification as a socio-technical process that explicitly engages with trust dynamics, stakeholder values, and collective responsibility. This paper: (1) identifies fundamental mismatches between current verification paradigms and human-centred trust; (2) proposes a framework integrating formal methods with stakeholder engagement; and (3) demonstrates feasibility through a case study. Our goal is to extend verification, ensuring AI systems are not only correct but genuinely trustworthy.

This vision paper is intentionally forward looking: it proposes a conceptual and methodological framework, supported by formal illustrations, rather than reporting large scale empirical deployment or validation.

## THE TRUST VERIFICATION PARADOX

**The Static Trust Fallacy.** Most existing approaches to trust verification model trust as a formally specified property or invariant, abstracted from human and social dynamics (Drawel et al., 2022). This conflicts with decades of empirical and theoretical research demonstrating that trust is adaptive, context dependent, and shaped by interaction history (Rousseau et al., 1998, Falcone and Castelfranchi, 2012). Trust increases through successful interaction, decreases following failure, and may recover under appropriate conditions.

In learning enabled AI systems, trust relationships evolve continuously. Users revise expectations and adapt reliance based on observed behaviour. A verification framework that assumes static trust relationships effectively verifies that trust cannot adapt paradoxically making the system less trustworthy in dynamic human-AI interaction contexts.

**Trust Evolution vs. Trust State.** We propose shifting focus from trust states to trust evolution. Rather than verifying that trust remains above a threshold, verification should reason about whether trust adapts appropriately in

response to evidence including responsiveness to failures, proportional degradation, opportunities for recovery, and resistance to manipulation.

We sketch Trust Evolution Temporal Logic (TETL), extending temporal logics (Baier and Katoen, 2008) with operators that explicitly model trust adaptation. Using standard temporal logic notation,  $G$  denotes “globally” and  $F$  denotes “eventually”.

$$G(E(\text{user}_i, \text{evidence}_e) \rightarrow F(A\varphi(\text{user}_i, \text{evidence}_e, \text{criteria}_c)))$$

Here,  $E(\text{user}_i, \text{evidence}_e)$  denotes that agent  $i$  incorporates trust relevant evidence  $e$  about the system, while  $A\varphi(\text{user}_i, \text{evidence}_e, \text{criteria}_c)$  denotes the subsequent adaptation of trust according to an update function  $\varphi(\text{user}_i, \text{evidence}_e, \text{criteria}_c)$  constrained by normative criteria  $c$ . This formulation captures properties such as responsiveness, non-discrimination, and recovery, shifting verification from stability to appropriateness of adaptation.

**Emergent Trust in Human-AI Ecosystems.** Many trust relevant properties arise at the collective level. In collaborative AI and recommendation systems, trustworthiness emerges from interaction patterns rather than individual components. Traditional compositional verification (Baier and Katoen, 2008) struggles in such settings, as system level trust properties emerge from collective interaction patterns rather than component level correctness (Selbst et al., 2019). Individual AI assistants may behave correctly while collectively exhibiting biased patterns when interacting with diverse user populations. Trust is shaped by long term collective dynamics, not local correctness alone.

## HUMAN-CENTRED TRUST: VALUES AND RESPONSIBILITY

**The Public Technical Trust Divide.** Formal verification is increasingly used in certification and regulatory assurance processes (U.S. Administration 2021, ISO 2022), but certification does not guarantee trust. Verification artifacts often remain accessible only to technical experts, creating a gap between technical assurance and perceived legitimacy (Rudin, 2019). This produces what we term “certification theatre”: systems satisfying formal requirements while failing to address stakeholder concerns about understanding, control, and accountability. Public trust depends critically on transparency and explainability, not just technical correctness (Pew Research Center, 2025).

The problem is that verification treats human understanding as external. Even when efforts translate verification results, such translations are post hoc and may not address stakeholders’ actual concerns. What is needed is verification that engages with stakeholder concerns from the outset.

**Values as First Class Verification Concerns.** Trust is grounded in values such as fairness, transparency, privacy, safety, and agency (Friedman, 1996). We define values as stakeholder prioritized normative requirements constraining acceptable behaviour. Stakeholders may hold conflicting values: efficiency may conflict with equity, privacy with accountability, or automation with human control.

Value Sensitive Design (Friedman et al., 2017) and related frameworks (Floridi et al., 2018) emphasize embedding values in design but rarely extend to formal verification. Existing verification frameworks seldom make value trade-offs explicit, a limitation we address through Participatory Verification. Participatory Verification embeds values directly into specifications, enabling

verification to reason about normative requirements alongside functional properties. This involves: (1) structured elicitation through workshops and deliberation; (2) formalization as constraints within verification models; (3) verification that systems respect value constraints; and (4) presentation of results in stakeholder understandable terms.

**The Human Accountability Gap.** In AI systems, outcomes may be harmful even when components satisfy specifications. This creates an accountability gap: when harm occurs, it is unclear who should be held responsible (Matthias, 2004; Nissenbaum, 1996). Traditional verification focuses on correctness without explanation (Baier and Katoen, 2008; Clarke et al., 2018), providing no basis for determining whether fault lies with training data, model architecture, deployment context, or human operators. Trust erodes when responsibility cannot be attributed.

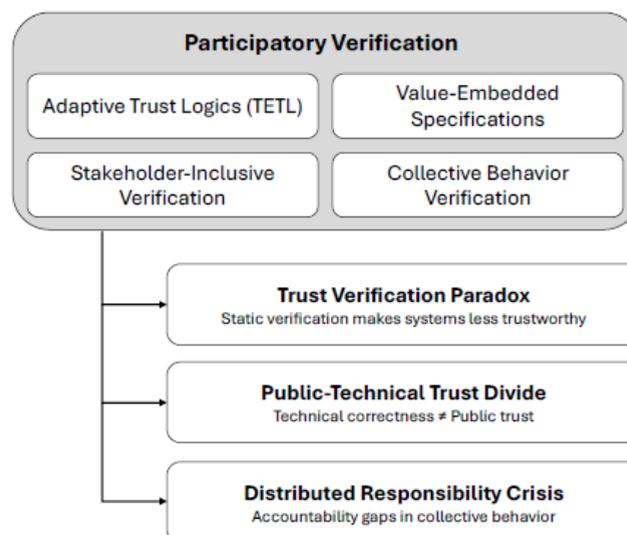
Drawing on causal reasoning frameworks (Pearl, 2009; Halpern, 2016), we propose extending verification with responsibility semantics. A responsibility function

$$R: A \times E \rightarrow [0,1]$$

maps each agent  $a \in A$  and event  $e \in E$  to a normalized value in  $[0,1]$  representing the agent's degree of causal contribution to the event, subject to normalization and causal relevance constraints. This enables explanation, accountability, and stakeholder facing justification. Importantly, responsibility attribution is not about blame but about understanding and learning.

## PARTICIPATORY VERIFICATION FRAMEWORK

Participatory Verification reframes verification as a socio-technical process integrating formal methods with stakeholder engagement (Schuler and Namioka, 1993). Rather than treating verification as purely technical, it engages stakeholders throughout the verification process itself.



**Figure 1:** Participatory Verification framework integrating trust evolution, stakeholder values, and responsibility attribution into formal verification for human-centred AI.

The framework comprises the following conceptual components:

**Trust Evolution Specifications.** Rather than static trust invariants, we verify properties of trust adaptation using TETL. This models how users should adjust reliance based on system performance, error patterns, and contextual factors. We can specify that trust decreases proportionally to error severity, then recovers with consistent good performance. Trust evolution specifications also address fairness (different user groups experience similar trust dynamics for similar experiences) and resistance to manipulation (trust responds to genuine evidence, not superficial signals).

**Value Embedded Formal Models.** Stakeholder values are elicited through structured engagement and formalized as constraints. The process involves: *value elicitation* through workshops and deliberation; *value formalization* translating abstract values into concrete constraints (stakeholders review representations to ensure accuracy); *value conflict resolution* through participatory deliberation determining acceptable trade-offs; and *value verification* proving the model respects value constraints across all behaviours.

**Stakeholder Engagement Protocols.** Verification is envisioned to involve stakeholders in defining acceptable behaviour, resolving value conflicts, and interpreting results. Protocols include: *accessible specification review* in multiple formats (formal notation, natural language, visual diagrams, scenarios); *assumption negotiation* where stakeholders challenge and revise verification assumptions; and *verification result interpretation* presenting results with rich context explaining what was proven, under what assumptions, and with what confidence.

**Responsibility Attribution.** Verification outputs include causal responsibility assignments, serving multiple purposes: *explanation* of how outcomes occurred; *accountability* clarifying which entities should be held responsible; *learning* by identifying weak points for improvement; and *trust calibration* helping users understand system limitations and capabilities through responsibility distributions across scenarios.

## **CASE STUDY: AUTONOMOUS VEHICLE COORDINATION**

We illustrate the expressive power and intended application of Participatory Verification using an autonomous vehicle coordination scenario involving vehicle agents, intersection controllers, and emergency response mechanisms. This case study is illustrative: its purpose is to demonstrate how trust evolution, stakeholder values, and responsibility attribution can be formally specified and verified, rather than to report a completed implementation or empirical validation.

### **Limitations of Traditional Verification**

Traditional verification approaches for autonomous vehicle systems focus primarily on safety and efficiency properties such as collision avoidance, deadlock freedom, and traffic throughput. While these properties are necessary, they capture only a narrow subset of what makes such systems trustworthy from a human-centred perspective. Concerns related to

explainability, fairness, accountability, privacy, and appropriate trust calibration are typically left unaddressed.

### Illustrative Trust Evolution Specification

Within Participatory Verification, trust is treated as a dynamic process rather than a static invariant. We model this using Trust Evolution Temporal Logic (TETL), which enables reasoning about how trust should adapt in response to evidence. An example trust responsiveness property is:

$$G(E(v, \text{signal\_performances}) \rightarrow F(A_\phi(v, \text{signal\_performances}, \text{fairness\_and\_reliability})))$$

The formula states that whenever a vehicle  $v_i$  incorporates new *evidence signal\_performance*, the system must eventually adapt the vehicle's trust in those signals in a manner consistent with fairness and reliability criteria. This captures appropriate trust calibration in which vehicles revise reliance after inconsistencies, while ensuring that similar evidence leads to similar trust adaptations across vehicles.

### Value Embedded Specifications

Stakeholder values can be formalized as first class verification constraints rather than informal design goals. The following examples illustrate how common concerns in autonomous vehicle coordination may be expressed formally.

#### Pedestrian Safety Priority

$$G(\text{pedestrian\_present}(\text{intersection}_i) \rightarrow \text{Priority}(\text{pedestrian\_safety}, \text{traffic\_efficiency}))$$

This property ensures that whenever pedestrians are present, safety takes precedence over traffic efficiency.

#### Explainable and Overridable Decisions

$$G(\forall d, v : \exists e (\text{comprehensible}(e) \wedge \text{supports}(e, d) \wedge \text{overridable}(d, v)))$$

This specification requires that every routing decision affecting a vehicle is accompanied by an explanation that is comprehensible to the user and remains subject to human override.

#### Privacy Preservation

$$G(\text{anonymized}(l) \wedge \text{minimal\_retention}(l) \wedge \text{purpose\_limited}(l))$$

This property ensures that location data is anonymized, retained only as long as necessary, and used solely for explicitly defined purposes.

### Illustrative Responsibility Attribution

To address accountability for collective outcomes, Participatory Verification incorporates responsibility attribution using a causal contribution function:

$$R : A \times O \rightarrow [0, 1]$$

where  $A$  is the set of agents and  $O$  the set of observable outcomes. A normalization constraint ensures interpretability:

$$\forall o \in O : \sum_{a \in A} R(a, o) = 1$$

In an autonomous vehicle coordination setting, this enables attribution of responsibility for outcomes such as congestion or unsafe manoeuvres across vehicle agents, infrastructure components, and coordination mechanisms. Responsibility attribution supports explanation, accountability, and system improvement rather than blame.

### Role of the Case Study

These formal illustrations demonstrate that Participatory Verification can express adaptive trust behaviour, encode stakeholder values as verifiable constraints, and support responsibility attribution for collective outcomes. They establish a formal foundation for future tool development and empirical evaluation without presupposing completed implementation or validation.

*All specifications in this section were illustrative and intended to demonstrate expressiveness rather than claim completed implementations or empirical validation.*

## EVALUATION AND CHALLENGES

**Human-Centred Evaluation.** Evaluation must combine technical metrics with human-centred measures. We propose: *quantitative trust measures* using validated scales (Mayer 1995) assessing perceived competence, integrity, benevolence, alongside measures of fairness, transparency, and control; *qualitative understanding* through semi structured interviews capturing how stakeholders interpret results and whether values are reflected; *longitudinal studies* tracking how trust adapts over weeks or months of interaction; and *behavioural indicators* including usage patterns, override rates, and complaint frequencies providing objective measures.

**Scalability and Complexity.** Participatory Verification introduces challenges of scalability, specification complexity, and coordination. Embedding values increases verification overhead; responsibility attribution may be computationally expensive. Current tools like PRISM (Kwiatkowska et al., 2011) and Event-B (Abrial 2010) require extension for trust evolution operators and value constraints. Stakeholder engagement presents scalability challenges: illustrative applications of Participatory Verification typically involve a limited number of stakeholder representatives, whereas real world deployments may involve millions of users with heterogeneous and potentially conflicting values. We are exploring representative sampling, value clustering,

and hierarchical engagement, though these introduce challenges regarding representativeness and fairness.

**Stakeholder Diversity and Value Conflicts.** Real populations are heterogeneous with conflicting values and varying technical literacy. Effective engagement requires facilitation, value negotiation protocols, and minority representation mechanisms. Future evaluations should include efficiency oriented users clashed with safety prioritizing users, privacy advocates objected to data collection while planners wanted extensive data. We are developing structured methods including value ranking exercises, scenario based discussions, and Pareto frontier visualization to balance inclusivity with tractability.

**Technical Innovation Needs.** Realizing Participatory Verification requires: verification algorithms for trust evolution properties; specification languages for value constraints; automated tools that scale to millions of states and users; and responsibility attribution algorithms balancing accuracy with feasibility.

**Integration with Machine Learning.** Many AI systems employ machine learning with learned rather than specified behaviour. Formal verification of neural networks remains limited in scalability. Integrating Participatory Verification with machine learning poses unique challenges: verifying trust evolution in continuously adapting systems, embedding value constraints in learned models, and attributing responsibility when behaviour emerges from training data. We envision hybrid approaches combining formal verification for high level properties with statistical verification and runtime monitoring for learned components.

## RELATED WORK

**Trust Models and Computational Trust.** Organizational trust models (Mayer et al., 1995; Rousseau et al., 1998) identify trust dimensions including competence, integrity, and benevolence, emphasizing trust as dynamic and relationship based. Research on trust in automation (Lee and See, 2004) has shown how trust calibration affects human reliance on automated systems. Computational trust research (Falcone and Castelfranchi, 2012) focuses on modelling trust in multi-agent systems but typically uses simple update rules failing to capture rich adaptive dynamics. Recent work has begun to formally model trust in autonomous systems, including delivery vehicles (Altamimi et al., 2025). Our TETL framework extends these efforts by formalizing trust evolution as verifiable property in human-AI interaction contexts.

**Explainable and Transparent AI.** Work on explainable AI (Arrieta et al., 2020) addresses transparency through post hoc explanation of model decisions, treating explanation as separate from verification. Rudin (Rudin, 2019) argues for inherently interpretable models in high stakes domains. Our work extends this to verification rather than verifying opaque properties and explaining afterward, we advocate verifying human-centred properties directly.

**Value Centred Design and Ethics.** Value Sensitive Design (Friedman, 1996; Friedman et al., 2017) and participatory design (Schuler and Namioka, 1993)

emphasize stakeholder engagement in development, ensuring technologies reflect human values. These frameworks operate at design time; our work extends them to the verification phase itself, ensuring values are formally verified.

**Formal Verification of Trust Properties.** Formal verification of trust in multi-agent systems (Drawel et al., 2022) typically uses temporal logics to specify trust invariants, assuming static models. Our work challenges this, arguing static trust verification is mismatched to human-AI interaction where trust must evolve appropriately.

**Responsibility and Causality.** Work on responsibility in AI (Matthias, 2004; Nissenbaum, 1996) addresses the “responsibility gap” in autonomous systems. Causal reasoning frameworks (Pearl, 2009; Halpern, 2016) provide formal foundations but are rarely integrated with verification. Our responsibility semantics embed causal frameworks within verification.

**Human-AI Interaction and Regulation.** Human-AI interaction research (Amershi et al., 2019) produces guidelines emphasizing transparency and control. Regulatory frameworks (European Parliament, 2024, U.S. Administration, 2021; ISO, 2022) mandate transparency documentation, and assurance obligations. Our work provides formal foundations for these principles, making them verifiable properties rather than aspirational guidelines.

## CONCLUSION

This paper has argued that formal verification, while essential, is insufficient for building trustworthy AI systems in human-centred contexts. Trust is adaptive, value laden, and socially mediated, and is therefore fundamentally mismatched with traditional verification paradigms focused on static properties and technical correctness.

We have identified three structural gaps: the Trust Verification Paradox, the Public Technical Trust Divide, and the Human Accountability Gap. In response, we proposed Participatory Verification as a forward looking framework that embeds trust evolution, stakeholder values, and responsibility attribution directly into formal verification. Through formal and illustrative specifications, we demonstrated the expressive power of this approach and outlined how it can be applied in realistic socio-technical settings.

While this work does not report empirical deployment or validation, it establishes the formal and methodological foundations necessary for future tool development, large scale evaluation, and real world application. Realizing Participatory Verification in practice will require both technical advances in verification algorithms and methodological innovation in stakeholder engagement and evaluation.

We call on the AI research community to embrace this challenge, to develop the tools and techniques needed for Participatory Verification, and to ensure that the AI systems we create are worthy of the public trust they require. The stakes could not be higher: the future of AI in society depends on our ability to bridge the gap between technical verification and public trust. The AI research community has the opportunity and the responsibility to lead this transformation.

## REFERENCES

- Abrial, J.-R. (2010) *Modeling in Event-B: System and Software Engineering*. Cambridge: Cambridge University Press.
- Akintunde, M. et al. (2023) Verifiably safe and trusted human-AI systems: A socio-technical perspective, *TAS*, pp. 56:1–56:6.
- Altamimi, M. Salehi Fathabadi, A. and Yazdanpanah, V. (2025) Formal modeling of trust in AI-driven autonomous delivery vehicles, *Lecture Notes in Computer Science*, vol. 16194. Springer, pp. 234–251.
- Amershi, S. et al. (2019) Guidelines for human-AI interaction, in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York: ACM, pp. 1–13.
- Arrieta, A.B. et al. (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*.
- Baier, C. and Katoen, J.-P. (2008) *Principles of Model Checking*. Cambridge, MIT Press.
- Clarke, E.M. Henzinger, T.A. Veith, H. and Bloem, R. (2018) *Handbook of Model Checking*. Springer.
- Drawel, N., Bentahar, J., Laarej, A. and Rjoub, G. (2022) Formal verification of group and propagated trust in multi-agent systems, *AAMAS*, 36(1), p. 19.
- European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 on artificial intelligence. *Official Journal of the European Union*, L 1689.
- Falcone, R. and Castelfranchi, C. (2010) *Trust Theory: A Socio-Cognitive and Computational Model*. Chichester: John Wiley & Sons.
- Floridi, L. et al. (2018) AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, *Minds and Machines*, 28(4), pp. 689–707.
- Friedman, B. (1996) Value-sensitive design, *Interactions*, 3(6), pp. 16–23.
- Friedman, B., Hendry, D.G. and Borning, A. (2017) A survey of value sensitive design methods, *Foundations and Trends in Human-Computer Interaction*, 11(2), pp. 63–125.
- Halpern, J.Y. (2019) *Actual Causality*. MIT Press.
- ISO (2022) Road vehicles—Safety of the intended functionality. *ISO/PAS 21448:2022*.
- Kwiatkowska, M. Norman, G. and Parker, D. (2011) PRISM 4.0: Verification of probabilistic real-time systems, Springer, pp. 585–591.
- Lee, J.D. and See, K.A. (2004) Trust in automation: Designing for appropriate reliance, *Human Factors*, 46(1), pp. 50–80.
- Matthias, A. (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata, *Ethics and Information Technology*, 6(3), pp. 175–183.
- Mayer, R.C. Davis, J.H. and Schoorman, F.D. (1995) An integrative model of organizational trust, *Academy of Management Review*, 20(3), pp. 709–734.
- Nissenbaum, H. (1996) Accountability in a computerized society, *Science and Engineering Ethics*, 2(1), pp. 25–42.
- Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*. 2nd edn. Cambridge: Cambridge University Press.
- Pew Research Center. (2025) How the U.S. public and AI experts view artificial intelligence. Available at: <https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/>
- Rousseau, D.M., Sitkin, S.B., Burt, R.S. and Camerer, C. (1998) Not so different after all: A cross-discipline view of trust, *Academy of Management Review*, 23(3), pp. 393–404.

- Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence*, 1(5), pp. 206–215.
- Salehi Fathabadi, A. (2025) The trust-safety divide: A critical gap in human-robot interaction research, in *Proceedings of the SCRITA Workshop*.
- Salehi Fathabadi, A. and Leonard, P. (2024) “Trust equals less death – it’s as simple as that”: Developing a socio-technical framework for trustworthy defence and security automated systems, *TAS*, pp. 20:1–20:10.
- Schuler, D. and Namioka, A. (1993) *Participatory Design: Principles and Practices*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Selbst, A.D., et al., (2019) Fairness and abstraction in sociotechnical systems, *ACM*, pp. 59–68.
- Slovic, P. (1993) Perceived risk, trust, and democracy, *Risk Analysis*, 13(6), pp. 675–682.
- U.S. Food and Drug Administration (2021) *Artificial Intelligence in Software as a Medical Device*. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
- Wing, J.M. (2021). Trustworthy AI. *Communications of the ACM*, 64(10), pp. 64–71.