The Trust-Safety Divide: A Critical Gap in Human-Robot Interaction Research

Dr. Asieh Salehi Fathabadi
University of Southampton
Southampton, UK
a.salehi-fathabadi@soton.ac.uk

Abstract—**Human-Robot Interaction (HRI) research faces a fundamental divide: technical safety verification develops independently from robot trust requirements, creating a critical barrier to real-world robot deployment. While formal methods can mathematically prove robot safety, these assurances consistently fail to translate into human trust in robots. Conversely, HRI trust approaches lack the rigorous foundations required for safety-critical robotic applications. This position paper argues that the field must move beyond this artificial separation toward computational robot trust verification - systematic approaches that formally integrate robot safety with human trust properties. We examine the state-of-the-art across both domains, identify critical gaps preventing integration, and outline a research agenda for bridging this divide. Without addressing this trust-safety gap, technically sound robots will continue to face deployment barriers, limiting their potential to benefit human-robot collaboration.**

Index Terms—**human-robot interaction, robot trust, formal verification, robot safety, trust calibration**

I. Introduction

The promise of robots to transform human society, from life-saving surgical robots to collaborative manufacturing systems, remains largely unfulfilled despite remarkable technical advances. We can now mathematically prove that autonomous vehicles will never violate traffic safety constraints [1], formally verify that surgical robots operate within precise tolerance bounds [2], and demonstrate that industrial robots maintain safety invariants under all specified conditions [3]. Yet these technically excellent robots face persistent deployment barriers that have little to do with their actual safety or capability.

The fundamental challenge lies in a divide that the autonomous systems research community has inadvertently created between technical excellence and social acceptance. On one side, formal verification researchers have developed increasingly sophisticated mathematical methods for proving system safety. Their work provides mathematical certainty about system behavior under specified conditions - a remarkable achievement that enables deployment in safety-critical domains from aerospace to nuclear power. On the other side, human-computer interaction researchers have developed deep understanding of how people form trust in technological systems, identifying the psychological, social, and cultural factors that determine whether people will accept and appropriately rely on autonomous systems.

These two research universes rarely intersect in meaningful ways, creating what we term the ”trust-safety divide.” Technical safety verification proceeds without systematic consideration of how mathematical proofs translate into stakeholder confidence. Social trust research advances without connection to the formal assurance methods required for safety-critical deployment. The result is a persistent gap between technical capability and social adoption that limits the beneficial impact of autonomous systems.

This divide has real consequences. Industry surveys reveal that only 44% of people globally feel comfortable with businesses using AI, while 91% of organizations doubt their preparedness to implement AI safely and responsibly [4], [5]. High-profile cases of technically sound robotic systems facing human resistance and formally verified medical robots struggling with clinician acceptance demonstrate the practical costs of this disconnect [6], [7]. Research has shown that social credibility factors directly impact safety performance in human-robot interaction [8].

The emergence of comprehensive AI regulation - from the EU AI Act [9] to national AI safety institutes [10], further emphasizes the urgency of bridging this divide. These regulatory frameworks explicitly require both technical compliance and social acceptability for robotic systems.

Recent research has begun to recognize this challenge, emphasizing the need for socio-technical approaches that integrate formal verification with human trust considerations [11]. Work on socio-technical frameworks for trustworthy defense systems emphasizes that ”trust equals less death” - highlighting the life-critical importance of getting trust-safety integration right [12]. Advances in formal specification of trust in multiagent systems demonstrate the feasibility of mathematical trust modeling [13], [14]. However, these efforts remain isolated and have not yet coalesced into a systematic approach for computational trust verification.

This position paper argues that the field must move beyond treating technical safety and social trust as separate concerns toward a unified approach we term **computational trust verification**. This paradigm shift involves extending formal verification methods to prove not just that systems are safe, but that they behave in ways that warrant stakeholder trust. Rather than asking ”How can we make people trust our safe system?” we propose asking ”How can we prove our system behaves in ways that justify trust?”

We examine the current state of both technical safety verification and social trust research, analyze why existing bridge-building efforts remain insufficient, and outline a comprehensive research agenda for computational trust verification. Our goal is not merely to identify this gap, but to catalyze the interdisciplinary collaboration needed to systematically address it.

The remainder of this paper examines the state-of-the-art divide between robot safety verification and HRI trust research (Section II), introduces computational robot trust verification as the missing link (Section III), outlines a research agenda (Section IV), presents a call to action (Section V), and concludes (Section VI).

II. The Problem: A Field Divided

The autonomous systems research community has inadvertently created two parallel universes. In one, formal verification researchers develop sophisticated mathematical proofs of system safety. In another, human-computer interaction specialists study how people form trust in AI systems. These universes rarely intersect, creating a fundamental problem for real-world deployment.

Previous editions of the SCRITA workshop have consistently highlighted this challenge, emphasizing how huge advances have been made in studying trust and acceptance factors in controlled settings, while significant gaps remain in developing robots that can maintain trust in real-world deployment [15]–[17]. The workshop series has identified the persistent need for metrics that allow effective assessment of people's trust towards robots and the development of robots that can proactively adapt their behaviors to maintain human trust [18].

A. The Technical Safety Universe

Formal verification has achieved remarkable success in proving safety properties of autonomous systems. Formal specification, model checking, and theorem proving provide mathematical guarantees about system behavior. The aerospace and automotive industries rely on these methods for safety certification.

The problem: These proofs remain incomprehensible to virtually everyone except formal methods experts. A railway operator doesn't understand temporal logic formulas. A hospital administrator can't interpret formal refinements. A regulatory inspector struggles with theorem prover outputs.

B. The Social Trust Universe

Meanwhile, HCI researchers have developed sophisticated understanding of trust formation in human-AI interaction. We know trust depends on perceptions of competence, benevolence, and integrity. We have validated measurement instruments and understand how transparency and explainability influence trust formation.

Their problem: These approaches often lack the mathematical rigor required for safety-critical applications. Trust-building techniques for recommendation systems may be insufficient for autonomous surgical robots. Social trust without technical foundations cannot provide the assurance required for life-critical decisions.

C. The Real-World Consequences

This divide has real consequences:

- Only 44% of people globally feel comfortable with businesses using AI [4]
- 91% of organizations doubt their preparedness to implement AI safely and responsibly [5]
- Multiple autonomous system deployments have faced public resistance despite technical safety demonstrations

The pattern is clear: technically verified systems fail to gain stakeholder acceptance, while socially trusted systems lack safety assurance. This blocks deployment of beneficial technologies in healthcare, transportation, and other critical domains including the robotics.

III. State-of-the-Art: Two Ships Passing in the Night

A. Technical Safety Verification: Impressive but Isolated

Current formal verification approaches excel at proving technical properties, providing mathematical certainty about system behavior. Event-B refinement [19], temporal logic model checking [20], and theorem proving have enabled verification of complex systems from avionics to nuclear control systems. Tools like PRISM [21] for probabilistic verification and Rodin [22] for Event-B development have proven their worth in industrial settings. However, these approaches assume that mathematical proof equals stakeholder trust - an assumption that is demonstrably false.

Example: An autonomous train control system can be formally verified to never cause collisions under specified track conditions using Event-B refinement [23]. Yet train operators may distrust the system because they cannot understand why it makes specific decisions, leading to manual overrides that compromise both safety and efficiency.

B. Social Trust Research: Insightful but Insufficient

HCI research has revealed crucial insights about trust formation, developing deep understanding of psychological trust mechanisms and validated measurement approaches. Foundational work by Lee and Moray [24] established the critical role of trust in human-automation systems, showing that trust significantly affects operators' decisions to rely on or intervene in automated systems. Research on trust calibration demonstrates the importance of aligning user trust with actual system capabilities [25]. The Technology Acceptance Model and its extensions provide frameworks for understanding technology adoption, while validated instruments like the Human-Computer Trust Scale [26] enable systematic trust measurement.

However, these approaches rarely connect to mathematical safety assurance. Trust-building techniques focus on user experience and interface design but typically lack the rigorous foundations required for safety-critical deployment.

Example: An explainable AI system for medical diagnosis might build physician trust through clear reasoning explanations, but without formal verification of diagnostic accuracy bounds, the system cannot be deployed in critical care scenarios where incorrect diagnoses have life-or-death consequences.

C. Attempted Bridges: Promising but Incomplete

Explainable AI (XAI): Focuses on making AI decisions interpretable through techniques like LIME, SHAP, and attention mechanisms [27]. While valuable for transparency, XAI rarely connects explanations to formal safety properties or provides mathematical guarantees about explanation correctness.

Trustworthy AI frameworks: Initiatives like the EU's Ethics Guidelines for Trustworthy AI provide high-level principles but lack systematic integration methods [28]. These frameworks identify important dimensions but don't provide computational approaches for verification.

Human-centered verification: Emerging work considers human factors in formal verification through approaches like human-machine interface verification [29]. However, these efforts typically model humans as components rather than formally representing trust properties.

These efforts are valuable but remain piecemeal. They don't address the fundamental challenge: **How can we systematically prove that technical safety mechanisms actually build stakeholder trust?**

IV. The Missing Link: Computational Trust Verification

We argue that the HRI field needs a new research direction: **computational trust verification** - systematic integration of formal verification methods with empirically-grounded robot trust models, creating unified frameworks that can prove both robot safety and human trustworthiness perceptions.

A. Core Principles

Formal Robot Trust Modeling: Trust properties in HRI should be modeled as mathematically precise predicates that can be formally verified alongside robot safety properties. Rather than treating trust as purely subjective, we can identify objective, verifiable robot behaviors that reliably produce trust in specific human user groups.

For example, context-dependent trust properties might be formalized as:

$$\begin{aligned} \text{TrustWorthy}(action, context, user) &\equiv \\ \text{Explainable}(action, user.expertise) & \\ \text{Predictable}(action, context.task) & \\ \text{Safe}(action, context.constraints) & \end{aligned}$$

where predicates are parameterized by situational and individual factors; and each predicate has formal definitions based on robot behaviors that empirical HRI research has shown to correlate with human trust. This approach builds on recent advances in formal specification of trust properties [13] and computational models of actual trust in multiagent systems [14], extending these foundations to safety-critical human-robot interaction.

Acknowledging Trust Complexity: Trust varies significantly across contexts, tasks, and individuals. We propose parameterized trust models that explicitly represent situational factors and user characteristics, recognizing trust's multi-dimensional nature while maintaining mathematical precision for safety-critical applications.

Integrated Robot Verification: Verification processes should simultaneously check robot safety and human trust properties, ensuring consistency between what the robot provably does and what humans need to trust it.

Trust Evidence Translation: Mathematical proofs should be automatically translated into human-appropriate evidence, bridging the communication gap between formal verification and human understanding of robot behavior.

B. Research Challenges

Formalizing Robot Trust: How can we create mathematical models that capture trust's situation-dependent and individual-variant nature while remaining formally verifiable? Current trust research identifies factors like competence and benevolence, but these manifest differently across contexts (surgical vs. manufacturing robots) and users (expert vs. novice operators). We need parameterized formal models that can represent trust's multi-dimensional complexity—including cultural factors, task-specific requirements, and individual differences—while maintaining the mathematical precision required for verification in safety-critical robotic systems.

Robot Verification Integration: How can we extend existing formal verification tools to handle trust properties alongside robot safety properties? This involves both theoretical work (extending formal logics) and practical tool development (modifying verification engines for robotic systems).

Trust Evidence Communication: How can we automatically generate compelling, accurate trust evidence from formal proofs about robot behavior? This requires advances in natural language generation, robotic interface design, and human-robot interaction.

Validation and Measurement: How do we validate that formal robot trust verification actually improves real-world human trust in robots? This involves developing new evaluation methodologies that bridge formal verification validation with empirical HRI trust measurement.

V. Research Agenda: Building the Bridge

A. Foundational Research Needs

Mathematical Robot Trust Foundations: Develop formal logics for representing trust properties in HRI, create systematic mappings between robot safety features and human trust indicators, and establish theoretical relationships between formal robot verification and trust formation. Key questions include: How can temporal logics be extended to express trust properties that evolve during human-robot interaction? How can we formalize trust calibration - the alignment between actual robot capabilities and human trust levels?

Robot Verification Method Innovation: Extend existing formal verification tools to handle trust properties in robotic systems, develop new algorithms for integrated robot safety-trust verification, and create automated reasoning techniques for trust property satisfaction. This involves extending Event-B with robot trust refinements, or developing model checkers that can verify probabilistic trust properties alongside deterministic robot safety properties.

Empirical HRI Validation: Conduct systematic studies linking formal robot properties to measured human trust, validate trust models across different human user groups and robotic application contexts, and develop standardized metrics for computational robot trust verification effectiveness. This requires advancing beyond traditional safety analysis approaches toward systematic ethical hazard analysis methods that can capture the complex interplay between trust and safety [30].

B. Application Domains for Validation

Surgical Robotics: Where both technical precision and surgeon trust are life-critical requirements. The regulatory environment provides clear safety standards, while operating room settings offer opportunities to measure trust impacts on patient outcomes and surgical performance.

Collaborative Manufacturing: Where worker trust in robotic systems affects both safety compliance and productivity. Factory environments allow controlled studies of robot trust-safety relationships with measurable outcomes in human-robot collaboration.

Service Robotics: Where public trust is essential for deployment in homes, hospitals, and public spaces. This domain offers opportunities to study how formal robot safety proofs can be translated into public confidence in everyday human-robot interaction.

Assistive Robotics: Where elderly or disabled users must trust robots for daily activities. High-stakes environments where robot failure could cause physical harm provide critical test cases for trust verification approaches.

C. Community Building Needs

Interdisciplinary Collaboration: The HRI field needs systematic collaboration between formal methods researchers, robotics engineers, and human factors specialists. We propose creating new venues specifically for robot trust-safety integration research and developing joint funding programs that require interdisciplinary teams.

Robot-Specific Tool Development: We need open-source tools that make computational robot trust verification accessible to robotics practitioners. This includes extending existing verification tools (Rodin, TLA+, SPIN) with robot trust modeling capabilities and robotic system interfaces.

Robot Trust Standards Development: The robotics industry needs standardized approaches for trust verification, similar to existing robot safety standards but encompassing human trust requirements. This involves working with standards bodies (ISO, IEEE, IEC) to develop robot trust verification standards.

VI. Call to Action: Why Now?

The urgency for this research direction has never been higher. The EU AI Act emphasizes trustworthy AI requirements with specific provisions for high-risk AI systems, including robots, that must demonstrate both technical compliance and social acceptability [9]. The UK AI Safety Institute focuses on public trust in AI systems [10]. These regulatory frameworks create immediate demand for computational approaches that can demonstrate both robot safety and human trustworthiness.

The formal methods community has developed powerful verification tools, while the HRI community has sophisticated understanding of robot trust formation. The convergence of these capabilities creates unprecedented opportunities for integration in robotics.

For the Formal Methods Community: Your mathematical rigor is essential for robot safety, but insufficient alone for robot deployment. Trust is a critical system property that affects real-world robot adoption. Consider: How can your verification techniques be extended to prove robot trust properties? Recent advances in formal trust specification [13], [14] provide starting points, but extending these to safety-critical human-robot interaction requires your mathematical rigor and tool-building expertise.

For the HRI Community: Your insights about robot trust formation are crucial, but need formal foundations for safety-critical robotic applications. Consider: How can robot trust theories be formalized as verifiable properties? How can HRI trust measurement be integrated with formal robot verification processes?

For Robotics Industry: The robot trust-safety gap is costing real money and delaying beneficial robot deployments. Early adopters of systematic robot trust verification approaches will gain competitive advantages in regulated markets where human-robot interaction is critical.

For Policymakers: Regulation focusing only on robot technical compliance or only on human acceptance will fail. Effective robot governance requires frameworks that integrate both technical verification and human trust assurance in HRI.

VII. Conclusion: Toward Unified Robot Trust-Safety Assurance

The divide between robot safety verification and HRI trust research is artificial and counterproductive. Computational robot trust verification represents a paradigm shift toward engineering trust as an integral robot property, extending mathematical rigor to encompass human factors that determine real-world success. By bridging this divide, we can accelerate deployment of robotic systems that are both mathematically safe and genuinely trusted by humans, ensuring robots actually benefit human-robot collaboration.

References

[1] M. Fisher, L. Dennis, and M. Webster, "Verifying autonomous systems," *Communications of the ACM*, vol. 56, no. 9, pp. 84-93, 2013.

[2] M. Webster, C. Dixon, M. Fisher, M. Salem, J. Saunders, K. L. Koay, K. Dautenhahn, and J. Saez-Pons, "Toward reliable autonomous robotic assistants through formal verification: A case study," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 2, pp. 186-196, 2016.

[3] M. Luckcuck, M. Farrell, L. Dennis, C. Dixon, and M. Fisher, "Formal specification and verification of autonomous robotic systems: A survey," *ACM Computing Surveys*, vol. 52, no. 5, pp. 1-41, 2019.

[4] Edelman Trust Barometer, "The AI Trust Imperative: Navigating the Future with Confidence," Edelman Trust Institute, 2025.

[5] McKinsey Global Institute, "Insights on responsible AI from the Global AI Trust Maturity Survey," McKinsey & Company, 2025.

[6] A. F. Winfield, K. Michael, J. Pitt, and V. Evers, "Machine ethics: The design and governance of ethical AI and autonomous systems," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 509-517, 2019.

[7] M. Salem and K. Dautenhahn, "Evaluating trust and safety in HRI: Practical issues and ethical challenges," in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 297-298, 2015.

[8] P. Holthaus, C. Menon, and F. Amirabdollahian, "How a robot's social credibility affects safety performance," in *Proc. International Conference on Social Robotics (ICSR)*, 2019, pp. 740-749.

[9] European Commission, "Artificial Intelligence Act: Regulation (EU) 2024/1689," Official Journal of the European Union, 2024.

[10] UK Department for Science, Innovation and Technology, "Introducing the AI Safety Institute," HM Government, 2024.

[11] M. Akintunde, V. Young, V. Yazdanpanah, A. Salehi Fathabadi, P. Leonard, M. Butler, and L. Moreau, "Verifiably Safe and Trusted Human-AI Systems: A Socio-technical Perspective," in *Proc. First International Symposium on Trustworthy Autonomous Systems (TAS)*, 2023, pp. 45-62.

[12] A. Salehi Fathabadi, P. Leonard, K. E. H. C. A. C. Dixon, and M. Fisher, "Trust equals less death - it's as simple as that: developing a socio-technical framework for trustworthy defence and security automated systems," in *Proceedings of the ACM Conference on Human-Factors in Computing Systems (CHI)*, pp. 156-171, 2024.

[13] M. Akintunde, V. Yazdanpanah, A. Salehi Fathabadi, C. Cirstea, M. Dastani, and L. Moreau, "Formal specification of actual trust in multiagent systems," *Formal Aspects of Computing*, vol. 36, no. 2, pp. 1-28, 2024.

[14] M. Akintunde, V. Yazdanpanah, A. Salehi Fathabadi, C. Cirstea, M. Dastani, and L. Moreau, "Actual Trust in Multiagent Systems," in *Proc. 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2024, pp. 45-53.

[15] A. Rossi, S. Papadakis, T. Koulouri, and S. Sabanovic, "IEEE Trust, Acceptance and Social Cues in Human-Robot Interaction – SCRITA 2022 Workshop," arXiv preprint arXiv:2208.11090, 2022.

[16] A. Rossi, S. Papadakis, T. Koulouri, and P. Holthaus, "Trust, Acceptance and Social Cues in Human-Robot Interaction – SCRITA 2021," arXiv preprint arXiv:2108.08092, 2021.

[17] A. Rossi, S. Papadakis, T. Koulouri, and P. Holthaus, "SCRITA 2023: Trust, Acceptance and Social Cues in Human-Robot Interaction," arXiv preprint arXiv:2311.05401, 2023.

[18] A. Rossi, "Trust, Acceptance and Social Cues in Human–Robot Interaction (SCRITA)," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1815-1816, 2021.

[19] J.-R. Abrial, *Modeling in Event-B: System and Software Engineering*. Cambridge University Press, 2010.

[20] E. M. Clarke, T. A. Henzinger, H. Veith, and R. Bloem, *Handbook of Model Checking*. Springer, 2018.

[21] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of probabilistic real-time systems," in *Proc. 23rd International Conference on Computer Aided Verification (CAV)*, 2011, pp. 585-591.

[22] J.-R. Abrial, M. Butler, S. Hallerstede, T. S. Hoang, F. Mehta, and L. Voisin, "Rodin: An open toolset for modelling and reasoning in Event-B," *International Journal on Software Tools for Technology Transfer*, vol. 12, no. 6, pp. 447-466, 2010.

[23] A. Iliasov, E. Troubitsyna, L. Laibinis, A. Romanovsky, K. Varpaaniemi, D. Ilic, and T. Latvala, "Formal development of critical systems with Event-B," in *Proceedings of the 18th International Workshop on Formal Methods for Industrial Critical Systems (FMICS)*, 2013, pp. 69-78.

[24] 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.

[25] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50-80, 2004.

[26] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53-71, 2000.

[27] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1-42, 2018.

[28] L. Floridi, C. Cowls, M. King, and M. Taddeo, "Translating the European principles for trustworthy AI into practice," *Philosophy & Technology*, vol. 32, no. 4, pp. 679-700, 2019.

[29] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal verification of ethical choices in autonomous systems," *Robotics and Autonomous Systems*, vol. 77, pp. 1-14, 2016.

[30] C. Menon, A. Rainer, P. Holthaus, S. Moros Espanol, and G. Lakatos, "X-HAZOP: A family of techniques for ethical hazard analysis of assistive robots," *IEEE Robotics and Automation Magazine*, 2025.