On the Establishment of Trust: Rethinking Trust in the Age of Artificial Intelligence

Christoph Lipps[1] and Siphesihle Sinthungu[2]
[1]Intelligent Networks Research Group, German Research Center for Artificial Intelligence, Kaiserslautern, Germany
[2] Academy of Computer Science and Software Engineering, University of Johannesburg, South Africa

Christoph.Lipps@dfki.de
siphesihles@uj.ac.za

Abstract: Trust has always been a fundamental element of organisational and social structures, enabling cooperation in situations where uncertainty or vulnerability cannot be fully eliminated. As societies and infrastructures have become increasingly interconnected, the foundations of trust have shifted from interpersonal relationships toward complex institutional and technological arrangements. This evolution has reached a new turning point: digital infrastructures, autonomous systems, and data-driven decision-making processes increasingly mediate interactions that once relied on human reasoning. Artificial intelligence (AI) is reshaping expectations of reliability, transparency, and accountability, prompting a reassessment of what it means to trust in environments where human and machine actions are intertwined. In this context, the notion of trustworthy AI has emerged as a key prerequisite for the responsible integration of AI systems into social and technical ecosystems. Core principles such as robustness, explainability, fairness, and security are widely regarded as essential for fostering a balanced relationship of trust. These requirements become especially significant in domains where AI supports or controls essential services. This is particularly true for critical infrastructures - energy grids, transportation systems, healthcare services, and communication networks-, where AI-driven decisions may trigger cascading effects that extend far beyond individual users. Ensuring trustworthy behaviour in such settings requires not only technical reliability but also governance structures that provide oversight, recourse, and continuous assurance. Despite broad agreement on the importance of trustworthiness, a substantial trust gap persists. Users often struggle to understand how AI systems reach their decisions, leading to scepticism or disengagement. Conversely, operators may place unwarranted confidence in automated outputs, especially in high-pressure or resource-constrained environments. Organisations face challenges translating abstract principles into operational practices, and regulatory frameworks remain inconsistent across sectors and jurisdictions. These tensions illustrate that trust cannot be created through technical means alone; it must be cultivated through transparent design, institutional accountability, and alignment with societal values.

Keywords: Trust**,** Trustworthiness, Trusted AI, Artificial intelligence

1. Introduction and Motivation

Trust is a foundational mechanism that enables coordinated action under uncertainty. Classical accounts frame trust as a way of reducing complexity in situations where full information or control is not possible (Luhmann, 1979; Gambetta, 1988). In modern societies, this function has increasingly been delegated to large-scale socio-technical systems, such as financial platforms, communication networks, healthcare systems, and energy infrastructures, where trust is no longer placed solely in individuals, but in the reliability and governance of the systems themselves.

This shift has reached a critical point with the rise of Artificial Intelligence (AI) methods. Unlike traditional software systems, AI models are inherently data-driven, probabilistic, and often opaque. Their behavior emerges from statistical patterns rather than explicitly defined rules, making it difficult to fully predict, interpret, or verify their decisions. As a result, AI challenges traditional engineering assumptions that underpin trust, namely determinism, transparency, and verifiability. Instead, trust in AI-enabled environments must be established under conditions of uncertainty, approximation, and evolving system behavior.

At the same time, AI is no longer confined to low-stakes applications. It is increasingly embedded in decision-making processes that affect safety-critical and societally significant outcomes. In domains such as healthcare, finance, transportation, and public administration, AI systems influence or automate decisions that were historically made by human experts. Empirical research in human–automation interaction shows that users struggle to calibrate their trust in such systems, often exhibiting over-reliance (automation bias) or under-reliance (algorithm aversion), depending on how system behavior is perceived and communicated (Parasuraman & Riley, 1997; Dietvorst, Simmons and Massey, 2015). This miscalibration highlights a fundamental problem: trust in AI is not simply a function of system performance, but depends on how technical capabilities align with human expectations, organizational practices, and institutional safeguards.

In response, the concept of trustworthy AI has emerged as a central organizing principle in both research and policy. Influential frameworks developed by bodies such as the European Commission, the Organization for Economic Co-operation and Development (OECD), and the National Institute of Standards and Technology (NIST) converge on a set of core requirements, including robustness, reliability, transparency, fairness, accountability, and human oversight. These principles reflect a broad consensus that trustworthiness must be engineered and governed across the lifecycle of AI systems, rather than assumed as an emergent property.

However, despite this convergence at the level of principles, a substantial gap remains between normative guidance and operational practice. Much of the existing work addresses individual dimensions of trustworthiness—such as explainability, fairness, or robustness—in isolation. Yet real-world systems require these properties to coexist and interact under constraints, often involving trade-offs. For example, improving model interpretability may reduce predictive performance, while increasing robustness to adversarial inputs can raise system complexity and reduce transparency. Moreover, improvements in technical metrics do not necessarily translate into appropriate user trust when system behavior remains difficult to interpret or contextualize.

These challenges become significantly more pronounced in the context of critical infrastructures. Systems such as energy grids, transportation networks, and healthcare services operate under strict requirements for reliability, safety, and continuity. They are also deeply interconnected, meaning that failures can propagate across subsystems and produce cascading effects. When AI is introduced into these environments, it does not merely augment existing processes; it becomes part of a tightly coupled cyber-physical system in which errors, uncertainties, or adversarial manipulations may have systemic consequences. Prior work on cyber-physical security demonstrates that such systems are particularly vulnerable to complex interactions between digital and physical components (Cardenas, Amin and Sastry, 2008; Rajkumar et al., 2010).

At the same time, advances in machine learning have revealed fundamental limitations in current approaches to reliability and robustness. Research on adversarial machine learning shows that even highly accurate models can fail under small, carefully crafted perturbations (Goodfellow, Shlens and Szegedy, 2015; Papernot et al., 2016). Studies on distributional shift and uncertainty estimation similarly demonstrate that models often behave unpredictably when exposed to conditions that differ from their training data (Hendrycks & Gimpel 2017; Gal & Ghahramani 2016). These findings challenge the assumption that strong benchmark performance is sufficient for trustworthy deployment in operational environments.

Taken together, these developments point to a fundamental tension. On the one hand, AI offers unprecedented capabilities for optimization, prediction, and automation in complex systems. On the other hand, the very properties that enable these capabilities also introduce new forms of uncertainty and risk. This tension is particularly acute in critical infrastructures, where the consequences of failure are systemic rather than local.

The motivation for this work is to address the fragmentation that currently characterizes research on trustworthy AI. Rather than treating trustworthiness as a collection of independent technical properties, this paper adopts an integrated socio-technical perspective that situates trust within organizational practices and institutional governance. Focusing on critical infrastructures as a high-stakes domain, the paper examines trust not as a static property achieved at design time, but as a dynamic relationship that must be continuously established, calibrated, and sustained over the lifecycle of AI-mediated systems (Lipps et al., 2024).

2. Background and State of the Art: Trust and Trustworthy AI

The study of trust naturally appears in multiple disciplines, including sociology, philosophy, economics, and, more recently, computer science. Early conceptualizations characterized trust as a mechanism for reducing social complexity (Luhmann, 1979) or as a rational expectation under uncertainty (Gambetta, 1988). In these accounts, trust is not blind confidence but a calibrated judgement that balances risk, evidence, and prior experience. This foundational view remains relevant in AI-mediated contexts, where uncertainty arises from data dependence, model approximation, and environmental variability.

Before the emergence of modern AI, research on trust in automation established principles that continue to inform current work. Studies in human factors showed that trust in automated systems is dynamic and context-dependent, evolving with system performance and user experience (Lee & See 2004). Inappropriate trust manifests in two forms: misuse (over-reliance) and disuse (underutilization), a framing that closely aligns with contemporary discussions of automation bias (Parasuraman & Riley 1997) and algorithm aversion (Dietvorst, Simmons and Massey, 2015).

Empirical work further demonstrated that users calibrate trust based not only on perceived reliability, but also on transparency and feedback. Even minor observable errors in algorithmic systems can lead to disproportionate loss of trust despite superior overall performance (Dietvorst, Simmons and Massey, 2015). These findings underscore that trust is shaped by how system behavior is communicated and interpreted, not solely by objective accuracy.

A central theme in recent work is the explicit distinction between trust and trustworthiness: Trust denotes a subjective human attitude or a *willingness* to accept vulnerability in light of uncertainty, whereas trustworthiness refers to properties of a system that can, in principle, be specified, evaluated, and justified independently of individual belief states. Recent work in communications engineering has emphasized this distinction, arguing that technical systems cannot be trusted in a moral or intentional sense but can only be made trustworthy by design through measurable system characteristics and governance mechanisms (Fettweis et al., 2025). Clarifying this distinction is crucial, as conflating trust with trustworthiness risks treating technical compliance or high performance as sufficient for the establishment of appropriate human trust.

2.1 From Trustworthy Automation to Trustworthy AI

The transition from traditional automation to AI introduces new challenges. Unlike rule-based systems, machine-learning models often lack explicit, human-interpretable logic, leading to their characterization as “black boxes” (Lipton 2018). This opacity complicates verification, validation, and accountability, particularly in high-stakes domains. In response, the field of explainable AI (XAI) has emerged, proposing techniques such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh and Guestrin, 2016) and Shapley Additive explanations (SHAP) (Lundberg & Lee 2017) to provide local or global explanations of model behavior. While such methods can improve interpretability, their ability to foster genuine understanding and appropriately calibrated trust remains contested. Explanations may increase user confidence without improving decision quality, thereby reinforcing risks of over-trust. Parallel efforts address robustness and reliability, particularly regarding adversarial manipulation and distributional shift. Research has shown that highly accurate models can fail under small perturbations (Goodfellow et al. 2015) or when deployed in environments that differ from training conditions (Hendrycks & Gimpel 2017). Together, these findings challenge the assumption that technical performance alone is sufficient for trustworthy deployment.

2.2 Fairness, Accountability and Ethical AI

Another core dimension of trustworthy AI concerns fairness and bias. Machine-learning systems can inherit and amplify biases present in training data, leading to discriminatory outcomes (O’Neil, 2016; Buolamwini & Gebru, 2018). This has motivated the development of fairness-aware learning approaches (Barocas, Hardt & Narayanan, 2019) and algorithmic auditing frameworks. Accountability extends the focus beyond technical correctness to include traceability, auditability, and responsibility assignment. Documentation mechanisms such as model cards (Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2018) support transparency and governance, but do not by themselves resolve questions of institutional oversight or liability.

2.3 Formalizing Trustworthy AI

Efforts to formalize trustworthy AI have converged around common principles. Frameworks proposed by the European Commission (2019), the OECD (2019), and the NIST AI Risk Management Framework (2023) emphasize robustness, transparency, fairness, accountability, and human oversight across the AI lifecycle. Despite broad alignment, these frameworks remain largely high-level and normative. Translating them into operational practices for complex, real-world systems remains challenging, particularly given trade-offs between transparency, privacy, robustness, and performance.

2.4 Limitations of the Current State of the Art

The current literature reveals several persistent limitations. First, research often addresses individual dimensions of trustworthiness -such as explainability, robustness, or fairness-, in isolation, neglecting their interaction within integrated systems. Second, improvements in technical metrics do not reliably translate into appropriately calibrated human trust, reinforcing the gap between trustworthiness and trust. Third, many approaches implicitly assume static deployment contexts, whereas real-world systems -especially in critical infrastructures-, are dynamic, adaptive, and exposed to evolving risks. Finally, trust is frequently treated as a property of the AI system itself rather than as an emergent phenomenon arising from interactions between technology, users, organizations, and institutions. This system-centric perspective limits the ability of existing

approaches to explain how trust is established, calibrated, and sustained in operational environments, thereby motivating the need for an explicitly socio-technical perspective.

2.5 The Trust Gap in AI -Mediated Systems

Despite growing convergence around principles of trustworthy AI, a persistent gap remains between AI systems that satisfy formal trustworthiness criteria and the way they are actually trusted, relied upon, or resisted in practice. This trust gap refers to a misalignment between objective system properties and the subjective or institutional trust relationships that govern real-world use. Similar gaps have long been discussed in the literature on trust and automation, where improved system performance does not necessarily lead to appropriate reliance (Lee & See, 2004; Hancock et al., 2011).

This misalignment emerges from the interaction of system properties, human judgement, and organizational context, and is conceptually summarized in Fig. 1.

2.6 Conceptualizing the Trust gap

A central source of this gap lies in the distinction between trust and trustworthiness. Trust is commonly defined as a willingness to accept vulnerability under conditions of uncertainty, based on expectations about the behavior of others (Luhmann, 1979; Mayer, Davis & Schoorman, 1995). Trustworthiness, by contrast, refers to properties of the trusted entity—such as ability, integrity, robustness, or reliability—that can be evaluated independently of the trustor's subjective belief. Recent work in both social sciences and engineering highlights the importance of maintaining this distinction, particularly in socio-technical systems where trust is often misattributed to technical artefacts (Bauer, 2021; Fettweis et al., 2025).

The trust gap emerges when improvements in trustworthiness—such as higher accuracy, explainability, or robustness—do not translate into appropriately calibrated trust. Conversely, trust may be placed in systems whose trustworthiness is insufficient. This misalignment reflects not a simple failure of technical performance, but a broader disconnect between system behavior, human expectations, organizational routines, and institutional safeguards.

2.7 Opacity, Uncertainty, and Miscalibrated Trust

The epistemic opacity of AI systems is a key contributor to the trust gap. Machine-learning models, particularly deep learning systems, operate on high-dimensional statistical representations that resist full interpretation or prediction. This opacity is structural rather than accidental, arising from the mathematical and computational foundations of machine learning (Burrell 2016). While explainable AI techniques attempt to mitigate this issue, they typically provide local or approximate explanations rather than comprehensive understanding (Lipton, 2018; Doshi-Velez & Kim, 2017).

Empirical studies suggest that such explanations can increase user confidence without necessarily improving decision quality or understanding of uncertainty (Binns et al., 2018; Ehsan et al. 2021). This creates a risk of illusory transparency, in which systems appear more trustworthy than they actually are. As a result, users often struggle to assess when a system's outputs are reliable, when uncertainty is high, or when human judgement should prevail.

These epistemic limitations translate directly into miscalibrated trust. Research in human–automation interaction shows that users frequently exhibit automation bias, deferring too readily to system recommendations, as well as algorithm aversion, rejecting AI support following salient failures (Parasuraman & Riley, 1997; Dietvorst, Simmons & Massey, 2015). Crucially, over-trust and under-trust are not opposing phenomena but manifestations of the same calibration problem: the absence of meaningful cues for assessing system competence across contexts (Lee & See, 2004).

2.8 Organizational and Institutional Mediation of Trust

Trust in AI-mediated systems is rarely formed at the level of individual cognition alone. AI systems are embedded in organizational settings that strongly shape how trust is constructed and enacted. Standard operating procedures, interface designs, training regimes, time pressure, and incentive structures influence whether and how AI outputs are used (Möllering, 2006). In many cases, reliance on AI reflects organizational compliance rather than genuine belief in system reliability.

This organizational mediation of trust also affects accountability. Responsibility for AI-mediated outcomes is typically distributed across developers, system integrators, operators, and regulators. Such fragmentation

complicates responsibility attribution and creates moral distance between decisions and consequences (Nissenbaum, 1996). When it is unclear who is accountable for errors or harms, trust becomes fragile—even when systems behave as designed.

While "human-in-the-loop" approaches are often proposed as safeguards, empirical and conceptual work suggests that nominal human oversight does not automatically resolve accountability concerns. When intervention authority is limited or poorly supported, human oversight may serve primarily as a symbolic reassurance rather than a substantive control mechanism (Kroll et al., 2017; Pasquale, 2015).

2.9 Trust as a Dynamic and Context-Sensitive Process

A further dimension of the trust gap is its temporal character. Trust in AI-mediated systems evolves over time as systems are updated, retrained, or deployed in new environments, and as organizational practices and regulatory expectations change. Static assessments of trustworthiness at design or deployment time cannot account for distributional shift, emerging vulnerabilities, or changes in operational context (Hoff & Bashir, 2015; Amodei et al., 2016).

Empirical studies show that trust is highly sensitive to experience and feedback, and can erode even in the absence of discrete failures if system behavior becomes less predictable or less intelligible (Hancock et al., 2011). This highlights the inadequacy of one-time certification or compliance-based approaches to trustworthiness. Trust must instead be continuously established, monitored, and recalibrated throughout the system lifecycle.

Taken together, these insights show that the trust gap in AI-mediated systems arises from the interaction of technical opacity, human cognitive limitations, organizational mediation, and institutional accountability structures. Closing this gap requires treating trust not as a static property engineered into systems, but as a socio-technical relationship embedded in context and sustained over time (Lipps, Scharwatz and Collier, 2026).

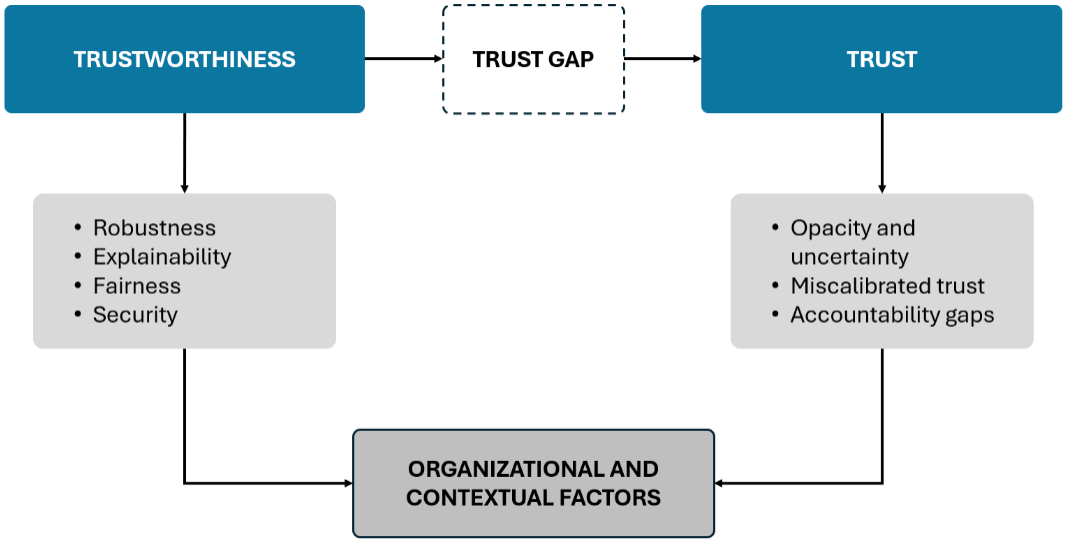

Figure 1: Conceptual illustration of the trust gap in AI-mediated systems

Objective trustworthiness properties of AI systems do not translate directly into established trust. The gap between trustworthiness and trust is mediated by organizational, contextual, and institutional factors that shape how AI systems are understood, relied upon, and governed.

3. Trustworthy AI in Critical Infrastructures

Trustworthy AI is particularly important in critical infrastructure environments because these systems provide services on which society depends, including electricity, water, transportation, healthcare, finance, telecommunications, and public safety. In such contexts, AI failure is not merely a technical inconvenience. It may affect physical safety, service continuity, public confidence, economic stability, and national security. This makes critical infrastructure a distinct setting for trustworthy AI: the central concern is not only whether an AI system performs well under nominal conditions, but whether it can be relied upon under operational stress, uncertainty, cyberattack, and changing environmental conditions.

Critical infrastructures are increasingly cyber-physical systems in which digital decisions produce physical consequences. AI may be used to detect anomalies in power grids, optimize traffic flows, support clinical triage, identify cyber intrusions, allocate scarce resources, or predict equipment failure. These applications can enhance efficiency and resilience, but they also introduce new forms of dependency. When AI becomes embedded in operational decision-making, errors can propagate across interconnected subsystems. A faulty prediction in one component may influence scheduling, control actions, maintenance planning, emergency response, or downstream infrastructure dependencies. For this reason, trustworthy AI in critical infrastructure must be understood as a system-level property, not a model-level attribute.

The NIST AI Risk Management Framework identifies characteristics of trustworthy AI such as validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness with harmful bias managed (NIST, 2023). These characteristics are especially relevant in critical infrastructure, where the cost of misplaced trust is high. A model that is accurate under normal operating conditions may still be untrustworthy if it fails under distribution shift, cannot detect uncertainty, lacks auditability, or cannot be overridden by human operators. NIST's recent work on a dedicated profile for trustworthy AI in critical infrastructure reinforces the need for sector-specific guidance spanning information technology, operational technology, and industrial control systems (NIST, 2026).

A key challenge is that critical infrastructure environments are often safety-critical, time-sensitive, and adversarial. In ordinary commercial settings, AI errors may result in inconvenience, reputational damage, or financial loss. In contrast, failures in critical infrastructure can compromise essential services or endanger human life. For example, AI-driven optimization in an energy grid must not only be efficient but also stable, secure, and robust against sensor faults, data poisoning, adversarial manipulation, and unexpected demand patterns. Similarly, AI used in healthcare or emergency response must support timely decision-making without obscuring responsibility or weakening professional judgement.

Cybersecurity is central in this context. Infrastructure operators increasingly deploy AI for threat detection, anomaly detection, and automated response. However, AI systems themselves become attack surfaces. ENISA distinguishes multiple dimensions of the AI–cybersecurity relationship: using AI to support cybersecurity, securing AI systems themselves, and addressing malicious uses of AI (ENISA, 2020; ENISA, 2025). This distinction is important because it avoids a one-sided view of AI as merely a defensive tool. In critical infrastructures, trustworthy AI must therefore include protection against adversarial examples, data poisoning, model theft, prompt injection, unsafe automation, and compromised supply chains.

Governance and accountability are equally important. The United States Department of Homeland Security emphasizes that responsibility for trustworthy AI in critical infrastructure is distributed across AI developers, cloud and compute providers, infrastructure owners and operators, regulators, and public authorities (US DHS 2024). No single actor can guarantee trustworthiness in isolation. Developers may design robust models, but operators must deploy them safely; infrastructure providers must secure platforms; regulators must define obligations; and affected stakeholders must have access to transparency and recourse mechanisms.

The European Union's AI Act similarly classifies certain AI systems used as safety components in the operation and management of critical infrastructure as high-risk systems. This reflects a broader regulatory judgement: where AI affects essential services, heightened requirements apply for risk management, data governance, technical documentation, human oversight, accuracy, robustness, and cybersecurity (European Union, 2024). The significance of this classification is not merely legal. It confirms that trustworthiness must be demonstrated through evidence, documentation, and lifecycle controls, rather than asserted through general ethical claims.

For this reason, trustworthy AI in critical infrastructure requires continuous assurance. Pre-deployment testing is necessary but insufficient. Models must be monitored after deployment for performance degradation, concept drift, abnormal inputs, and emerging threats. Operators require explanations that are actionable, not merely visual or statistical. They must know when to rely on the system, when to challenge it, and when to revert to fallback procedures or human control.

A growing body of applied research has begun to address these challenges, particularly in cyber-physical infrastructure systems. For example, *Munir, Shetty, and Rawat* (2023) propose a trustworthy AI framework for smart-grid cyber-attack detection that integrates reliability, explainability, transparency, fairness, and accountability into a unified pipeline. Their approach combines machine-learning-based detection with SHAP-based explanations and dynamic risk quantification, demonstrating that trustworthiness can be operationalized as measurable system properties within a real infrastructure context.

Complementing this, conceptual work on governance and accountability highlights that trustworthy AI in infrastructure systems requires clear responsibility allocation and regulatory alignment. Recent analyses of AI-enabled energy systems (Volkova et al., 2024) argue that traceability of decisions and accountability structures must be considered integral to trustworthiness, rather than external constraints. This reinforces the view that trust in critical infrastructures is inherently socio-technical.

In summary, critical infrastructure intensifies the trustworthy AI problem because reliability, security, accountability, and resilience cannot be separated. Trust cannot be reduced to user confidence or predictive accuracy. It must be grounded in demonstrable technical performance, robust cyber-physical safeguards, institutional accountability, and continuous operational assurance. These characteristics make critical infrastructure a key domain for translating trustworthy AI from abstract principles into practical, auditable, and enforceable socio-technical systems.

4. Conclusion and Future Work

Artificial Intelligence systems are increasingly integrated into societal infrastructures and organizational decision-making processes, often in contexts where reliability, safety, and accountability are critical. In response, the notion of trustworthy AI has gained prominence in research, policy, and regulation, resulting in a broad alignment around technical and ethical principles such as robustness, transparency, fairness, and human oversight. Yet, despite this convergence, trust in AI-mediated systems remains fragile, contested, and frequently miscalibrated in practice. This persistent tension indicates that technical trustworthiness alone does not resolve the deeper problem of how trust is established and sustained when human and machine actions are tightly intertwined.

Therefore, this work addresses this gap by examining trust in AI-mediated systems through a socio-technical lens. Drawing on theories of trust, empirical findings from human–automation interaction, and contemporary AI governance frameworks, the analysis demonstrate that trust and trustworthiness are distinct but interdependent concepts. Trustworthiness refers to objective system properties that can be designed, evaluated, and governed, whereas trust describes a willingness to accept vulnerability under uncertainty. The central contribution of this work lies in identifying the trust gap that emerges when these concepts are conflated - when technically trustworthy systems fail to elicit appropriate trust, or when trust is placed despite unresolved limitations.

By analyzing opacity, uncertainty, miscalibrated reliance, organizational mediation, and accountability fragmentation, the work shows that the trust gap is not a marginal or transient phenomenon. Rather, it arises from the interaction of technical system behavior, human judgement, organizational routines, and institutional arrangements. This diagnosis challenges system-centric approaches that treat trust as a property that can be engineered or certified at design time, and instead highlights trust as a dynamic relationship embedded in context.

The analysis further demonstrated that these issues become especially acute in critical infrastructures. In such environments, AI systems operate within tightly coupled cyber-physical systems, where uncertainty, adversarial conditions, and cascading effects are inherent. Here, trust failures have systemic consequences, and ensuring trustworthy behavior requires continuous assurance, clear responsibility allocation, and robust governance across organizational and sectoral boundaries. The discussion of regulatory and institutional perspectives illustrates that trustworthiness must be enacted and maintained over the system lifecycle, not merely asserted through abstract principles.

Several avenues for future research follow from this work: First, empirical studies are needed to investigate how trust in AI-mediated systems evolves over time in real-world organizational contexts, particularly under changing operational conditions. Second, domain-specific analyses of trustworthiness in different critical infrastructure sectors could help translate socio-technical insights into actionable practices. Third, further work should explore governance, auditing, and assurance mechanisms that explicitly address the relational and institutional dimensions of trust identified in this paper.

In conclusion, advancing trustworthy AI requires moving beyond a purely technical understanding of trustworthiness toward an integrated socio-technical perspective on trust. By reframing trust as a dynamic, context-sensitive, and institutionally mediated relationship, this paper provides a conceptual foundation for future efforts to build not only technically sound AI systems, but also resilient and legitimate trust in their deployment.

Acknowledgements

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 16KIS2239K SUSTAINET_guarDian). The authors alone are responsible for the content of the paper.

Ethics and Data Use Statement: This research used publicly available, anonymized secondary data. No identifiable personal information was accessed. Ethical approval was not required in accordance with institutional policies.

AI Assistance Statement: The authors used generative AI tools solely for linguistic editing and improving readability (e.g., grammar, phrasing, and structure). No AI system contributed to the formulation of the research questions, methodology, data interpretation, theoretical reasoning, or conclusions. All content was critically reviewed and approved by the authors, who take full responsibility for the manuscript.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schuhlmann, J., and Mané, D., "Concrete problems in AI safety", 2016, arXiv preprint arXiv:1606.06565.

Barocas, S., Hardt, M. and Narayanan, A. "Fairness and Machine Learning: Limitations and Opportunities", The MIT Press, 2023, ISBN: 978-0262048613.

Bauer, P.C., "Clearing the jungle: Conceptualizing trust and trustworthiness", *Trust Matters: Cross-Disciplinary Essays,* pp. 17 – 34, Hart Publishing (Oxford), 2021.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N., "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions", *CHI Conference on Human Factors in Computing Systems*, pp. 1 – 14, DOI: 10.1145/3173574.317395

Buolamwini, Joy and Gebru, T., "Gender shades: Intersectional accuracy disparities in commercial gender classification", *Proceedings of Machine Learning Research (FAT/ML),* 2018.

Burrell, J. "How the machine 'thinks': Understanding opacity in machine learning algorithms", *Big Data & Society*, vol. *3, no.* 1, 2016, DOI: 10.1177/2053951715622512.

Cardenas, A.A., Amin, S. and Sastry, S., "Research challenges for the security of control systems", *3rd Conference on Hot Topics in Security (HOTSEC) - USENIX Workshop*, pp. 01 – 06, San Jose, CA, USA,

Dietvorst, B.J., Simmons, J.P., and Massey, C., "Algorithm aversion: People erroneously avoid algorithms after seeing them err", *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, DOI: 10.1037/xge0000033

Ehsan, U., Wintersberger, P., Liao, Q.V., Watkins, E.A., Manger, C., Daumé III, H., Riener, A., and Riedl, M.O., "Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI", *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1 – 7, DOI: 10.1145/3491101.3503727

ENISA, 2020. Artificial Intelligence Cybersecurity Challenges. ENISA.

ENISA, 2025. ENISA Threat Landscape 2025. ENISA.

European Commission, 2019. Ethics Guidelines for Trustworthy AI. Brussels: European Commission.

European Union, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Brussels.

Fettweis, G.P., Grünberg, P., Hentschel, T., and Koepsell. S., "Conceptualizing Trustworthiness and Trust in Communications", *IEEE Communications Magazine,* pp. 126 – 132, vol. 63, no. 12, IEEE, 2025, DOI: 10.1109/MCOM.001.2400383.

Gambetta, D. (Ed.), "Trust: Making and Breaking Cooperative Relations" *The Economic Journal,* vol 99, no. 394, pp- 201 – 203, 1989, DOI: 10.2307/2234217

Gal, Y., and Ghahramani, Z., "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning", 33rd International Conference on Machine Learning (ICML), pp. 1050 – 1059, New York, NY, USA, 2016, DOI: 48:1050-1059, 2016.

Gebru, T., Morgenstern, J., Veccione, B., Vaughan, J.W., Wallach, H., Daumé III, H., and Crawford, K., "Datasheets for Datasets", *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden, 2018.

Goodfellow, I., Shlens, J. and Szegedy, C., "Explaining and Harnessing Adversarial Examples", in *International Conference on Learning Representations (ICLR)*, 2014.

Hancok, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., and Parasurman, R., "A meta-analysis of factors affecting trust in human-robot interaction", *Human Factors,* vol. 53, no. 5, 2011, DOI: 10.1177/0018720811417254

Hoff, K.A., and Bashir, M., "Trust in automation: Integrating empirical evidence on factors that influence trust", *Human Factors*, vol. 57, no. 3, pp. 407 – 434, 2015, DOI: 10.1177/0018720814547570

Hendrycks, D., and Gimpel, K., "A baseline for detecting misclassified and out-of-distribution examples in Neural Networks", International Conference on Learning Representations (ICLR), 2017.

Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., and Yu, H., "Accountable Algorithms", *University of Pennsyvania Law Review*, vol 165, 2017.

Lee, J.D. and See, K.A., "Trust in automation: Designing for appropriate reliance", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50 – 80, 2004, DOI: 10.1518/hfes.46.1.50_3039

Lipton, Zachary C., "The mythos of model interpretability", *ACM Queue*, vol. 16, no. 3, pp. 31–57, 2018, [Online: https://queue.acm.org/detail.cfm?id=3241340]

Lipps, C., Herbst, J., Reddy, R:, Rüb, M., and Schotten, H.D., „Authentication in a Hyperconnected World: Challenges, Opportunities and Approaches", *International Conference on Cyber Warfare and Security (ICCWS),* Johannesburg, South Africa, 2024, DOI: 10.34190/iccws.19.1.2070

Lipps, C., Scharwatz, D., and Collier, H., "On the Establishment of Trust: Challenges, Opportunities and Socio- Cultural Factors", *International Conference on Cyber Warfare and Security (ICCWS)*, Wilmington, NC, USA, 2026, DOI: 10.34190/iccws.21.1.4527

Luhmann, N., "Trust and Power", John Wiley and Sons Limited, 1979, ISBN-13: 978-1-5095-1945-3.

Lundberg, S.M. and Lee, S.-I., "A unified approach to interpreting model predictions", in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765 –4774, Long Beach, CA, USA, 2017.

Mayer, R.C., Davis, J.H., and Schoormann, F.D., "An Integrative Model of Organizational Trust", *The Academy of Management Review,* vol. 20, no. 3, pp. 709 – 734, 1995, DOI: 10.2307/258792

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., and Gebru, T., "Model cards for model reporting", in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 220 – 229, Atlanta, GA, USA, 2019, DOI: 10.1145/3287560.3287596

Munir, M., Shetty, S. and Rawat, D.B., "Trustworthy artificial intelligence framework for smart grid cyber-attack detection", *IEEE Winter Simulation Conference (WSC),* San Antonio, TX, USA, 2023,, DOI: 10.1109/WSC60868.2023.10408676

Moellering, G., "Trust: Reason, Routine, Reflexivity", Elsevir, 2006, ISBN: 0-08-044855-0

National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Managament Framework (AI RMF 1.0)", 2023, DOI: 10.6028/NIST.AI.100-1, [Online: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf]

Nissenbaum, H., "Accountability in a computerized society", *Science and Engineering Ethics,*, vol. 2, pp. 25 – 42, 1996, DOI: 10.1007/BF02639315

Organization for Economic Co-operation and Development (OECD), "Recommendation of the Council on Artificial Intelligence", Paris, France, 2019, [Online: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449]

O'Neil, C., "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy", Crown Publishing Group, New York, NY, United States, 2016, ISBN: 978-0-553-41881-1

Papernot, N., McDaniel. P., Jha, S., Fredrikson, M., Celik, Z.B., and Swami, A., "The Limitations of Deep Learning in Adversarial Settings", *IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372 – 387, Saarbruecke, Germany, 2016, DOI: 10.1109/EuroSP.2016.36

Pasquale, F., "The Black Box Society: The Secret Algorithms That Control Money and Information", Harvard University Press, 2015, ISBN: 978-0-674-73606-1

Parasuraman, R. and Riley, V., "Humans and automation: Use, misuse, disuse, abuse", *Human Factors*, vol. 39, no. 2, pp. 230 – 253, 1997, DOI: 10.1518/001872097778543886

Rajkumar, E., Lee, I., Sha, L., and Stankovic, J., "Cyber-Physical Systems: The Next Computing Revolution", *47th Automation Conference,* pp. 731 – 736, Anaheim, California, USA, 2010, DOI: 10.1145/1837274.1837461

Ribeiro, M.T., Singh, S. and Guestrin, C., "Why should I trust you? Explaining the predictions of any classifier'" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135 – 1144, 2016, DOI: 10.1145/2939672.29397

Sithungu, S. and Lipps, C., „Critical Infrastructure Security and the Role of AI: An Overview", European Conference on Cyber Warfare and Security (ECCWS), Kaiserslautern, Germany, 2025, DOI: 10.34190/eccws.24.1.3770.

Sithungu, S. and Lipps, C., „Revisiting Biometrics in Cybersecurity: Do AI Methods and Zero Trust Architectures Drive Innovation? ", International Conference on Cyber Warfare and Security (ICCWS), Wilmington, NC, USA, 2026, DOI: 10.34190/iccws.21.1.4533.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., and Hall, P., "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence", NIST Special Publication 1270, 2022, DOI: 10.6028/NIST.SP.1270

US Department of Homeland Security, "Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure", Washington, DC, USA, 2024. [Online: https://www.dhs.gov/sites/default/files/2024-11/24_1114_dhs_ai-roles-and-responsibilities-framework-508.pdf]

Volkova, A., Hatamian, M., Anapyanova, A., de Meer, H., "Being Accountable is Smart: Navigating the Technical and Regulatory Landscape of AI-based Services for Power Grid", *Proceedings of the 2024 International Conference on Information Technology for Social Good,* pp. 118 – 126, Association for Computing Machinery, New York, NY, USA, 2024, DOI: 10.1145/3677525.3678651.

Yan, Y., Qian, Y., Sharif, H., and Tipper, D., "A Survey on Cyber Security for Smart Grid Communications", *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 998–1010, DOI: 10.1109/SURV.2012.010912.00035.