Understanding Perceived Trustworthiness as a Basis for Human Oversight in the Wild

Evgenia Christoforou[1,*], Jahna Otterbacher[1,2] and Styliani Kleanthous[1]

[1]Open University of Cyprus, Nicosia, Cyprus
[2]CYENS Centre of Excellence, Nicosia, Cyprus

Abstract

Regulatory frameworks, such as the European Union's AI Act, emphasize the need for "trustworthy" AI systems and continuous, post-deployment monitoring, particularly for high-risk applications. In practice, however, a tension emerges. System trustworthiness is primarily conceptualized and assessed through technical and normative criteria such as robustness, fairness, and accountability, while users, as key stakeholders, form perceptions of an AI system's trustworthiness through casual interaction in real-world contexts. As a result, a gap arises between evaluating AI systems as trustworthy and experiencing them as trustworthy in real use contexts. The ultimate goal of Trustworthy AI is often framed as fostering appropriate trust between users and AI systems. User trust can be understood as the outcome of an implicit evaluation of an AI system's *perceived trustworthiness*, an evaluation that does not necessarily map directly onto formal trustworthiness principles, but instead, reflects an amalgamation of personal standards, expectations, and contextual cues. In this paper, we argue that understanding how perceived trustworthiness is formed and expressed by users can provide a basis for human oversight of AI systems in the wild. Users already produce trust-related judgments through everyday interaction with AI systems, even in the absence of expertise or awareness of governance frameworks. These judgments are continuous, multifaceted, and probabilistic in nature, unlike static or centralized evaluation mechanisms. Rather than treating such judgments as noise, we aim to leverage them as a resource. We conceptualize user trust judgments as distributed human oversight evaluations that can inform post-deployment monitoring when they are appropriately and systematically measured, interpreted, and aggregated. This work discusses key conceptual challenges in aligning trust, trustworthiness, and human oversight. Building on this, we describe our vision that shapes trust-related evaluations into oversight-relevant information, considering both casual users and crowdsourced participants. We conclude by discussing how capturing, measuring, and aggregating perceived trustworthiness can be incorporated into oversight processes without substituting accountability mechanisms.

Keywords

trust, trustworthiness, perceived trustworthiness, human oversight, crowdsourcing

1. Introduction

The concept of Trustworthy AI was introduced to establish and maintain trust between humans and AI systems.[1] Regulatory frameworks, such as the European Union's AI Act, explicitly frame trustworthiness as a prerequisite for responsible deployment, adoption, and continued reliance on AI, particularly in high-risk domains. Without mechanisms to sustain trust over time, the very premise of Trustworthy AI collapses: systems that are not trusted will not be used, and systems that are used without warranted trust [1] pose significant societal risks.

At the same time, trust in AI systems cannot be assumed to persist once established. AI systems evolve, are deployed across diverse contexts, and interact with users in ways that may diverge from their original design assumptions. As a result, maintaining the warranted trust is dependent on continuous monitoring, reassessment, and human oversight. Trustworthy AI, therefore, implies a feedback cycle in which trustworthiness is intended to foster user trust, user trust enables continued system use, and sustained use calls for ongoing oversight to ensure that trust remains warranted.

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
evgenia.christoforou@ouc.ac.cy (E. Christoforou); jahna.otterbacher@ouc.ac.cy (J. Otterbacher); styliani.kleanthous@ouc.ac.cy (S. Kleanthous)
0000-0003-0455-9357 (E. Christoforou); 0000-0002-7655-7118 (J. Otterbacher); 0000-0003-1594-1340 (S. Kleanthous)

[1]https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Within this cycle, a key tension emerges between how trustworthiness is formally assessed and how it is experienced in practice by users. Governance frameworks primarily evaluate trustworthiness through technical and normative criteria such as robustness, fairness, transparency, and accountability applied primarily through technical assessments conducted during development, deployment, and institutional governance processes [2]. In contrast, users encounter AI systems through everyday interaction in real-world contexts, forming perceptions of trustworthiness shaped by experience, expectations, and contextual cues [3, 4]. This tension underscores the importance of understanding how users perceive trustworthiness and integrating these perceptions into human oversight processes. Considering such an approach to human oversight in post-deployment AI systems aligns with the Trustworthy AI framing of the EU High-Level Expert Group on AI, which emphasizes enabling human agency and ensuring alignment with societal and ethical values throughout the ongoing use of AI systems.

2. Challenges in Aligning Trust, Trustworthiness & Human Oversight

As AI systems increase in complexity, become highly adaptive and widely deployed, maintaining meaningful human oversight is becoming progressively more challenging. AI systems are no longer confined to bounded tasks or standard contexts of use; they operate across domains, user populations, and time, often undergoing continuous updates and interacting with humans in open-ended ways. Furthermore, the speed at which new Foundation Models are being developed makes Generative AI (GenAI) in particular a “moving target"; it is simply impossible to predict the range of innovations and applications that will result from the use of GenAI and its Foundation Models [5]. This evolution poses fundamental challenges for existing and emerging governance mechanisms, which are largely designed around pre-deployment assessment and static evaluation criteria.

Generative AI-based conversational agents exemplify these challenges. As highly interactive systems embedded in everyday activities, which require a high degree of human-system symbiosis to perform well [6, 7], they prompt users to rapidly form multifaceted trust judgments spanning perceptions of robustness, competence, social behavior, ethical alignment, and personal standards. These trust judgments emerge through day-to-day use and interaction, rather than through the technical and normative criteria used to assess the trustworthiness of the system. As a result, a gap persists between systematic post-deployment evaluations of trustworthiness and how AI systems are experienced and trusted in practice. Although human oversight is widely recognized as a critical mechanism for monitoring AI system behavior after deployment, it remains unclear how user trust can be aligned with system trustworthiness to support warranted trust. The challenges outlined below formalize this tension and motivate the need for new ways of thinking about post-deployment human oversight.

Challenge 1. Aligning user trust with system trustworthiness. Trustworthy AI frameworks aim to foster appropriate user trust by defining system trustworthiness through technical and normative criteria. In practice, however, user trust does not mimic or align with these criteria of assessment [3]. This misalignment relates to the fact that trust is formed through interaction and experience, and is mediated by user personal characteristics [8, 9], whereas trustworthiness is assessed through abstract, system-level properties [10]. For example, the relevant ISO standard (ISO/IEC TR 24028:2020) defines trustworthiness as “the ability to meet stakeholder expectations in a *verifiable* manner," suggesting the need to somehow operationalize and measure “trustworthiness" for verification. It is important to note that the challenge of not only aligning user trust with system trustworthiness, but also sustaining this alignment over time, opens up questions about how deviations between the two may be observed and interpreted once systems are deployed.

Challenge 2. Understanding perceived trustworthiness as the basis for user trust. User trust emerges through perceptions of trustworthiness formed during interaction with AI systems and through exposure to contextual and institutional cues [3]. These perceptions might not always map or map directly onto formal trustworthiness principles. The process by which users form perceptions of trustworthiness is currently understudied, with few examples of empirical work (e.g, [11]). Thus, more

research is needed, particularly in natural, out-of-the-lab settings [12], as to better understand signals of (dis)trust, how they might change over time with repeated human-AI interaction and, if/how they can be used in studying/verifying trustworthiness.

Challenge 3. Positioning local trust experiences in globally deployed AI systems. Although there is recognition of the need to include members of the public in decisions concerning AI development, deployment, and regulation ("Participatory AI"), practices for engaging users in "everyday" contexts are ad hoc, with little consensus as concerns methodology [13]. Likewise, perceptions of trustworthiness are inherently local and context-dependent, shaped by specific use cases, user goals, interaction environments, and personal standards, or even an ethics code. In contrast, many AI systems are developed and deployed at a global scale across diverse populations and domains [14].

Challenge 4. Interpreting distributed trust signals for post-deployment human oversight. Following deployment, users continuously express trust-related judgments through behavioral interaction patterns with the AI system. They may even engage in unconscious monitoring when distrust is present to the point of disengagement if trust is completely broken. While such signals are rarely treated as part of formal oversight, they constitute a rich source of experiential evidence that becomes visible only during ongoing use. Tapping into these trust-related judgments, either through deliberate elicitation methods or by enabling users to become more reflective of their own interaction patterns, can present an alternative way of supporting post-deployment human oversight.

Challenge 5. Diversity and AI literacy as conditions for meaningful human oversight. Trust-related signals gain meaning through their distribution across diverse users and contexts. Diversity does not yield a holistic assessment of trustworthiness, but it reveals disagreement, disparate impacts, and context-specific risks. At the same time, as users have differing abilities to access and understand an AI system's trustworthiness cues [3], the quality, interpretability, and length of use of trust signals depend on users' understanding of the system and the potential risks associated with it. Thus, appropriately educating the public becomes pivotal not only for supporting trust calibration, but also for improving the quality and interpretability of distributed trust evaluations.

3. Human Oversight via Perceived Trustworthiness Vision

Having outlined both the challenges and the opportunities associated with aligning user trust and system trustworthiness, we now present our vision for post-deployment human oversight. We first introduce a set of working definitions that ground our proposed vision, and then argue how trust evaluations can be interpreted as oversight-relevant signals. Finally, we discuss how aggregation of trust evaluations can empower the ongoing monitoring of AI systems.

3.1. Working Definitions

We define *trust* as the user's willingness to rely on an AI system under conditions of uncertainty, based on expectations about the system's behavior and potential impact [1]. We consider trust to be dynamic, context-dependent, and shaped through interaction over time, rather than established once and for all. On the other hand, *system trustworthiness* denotes the set of technical, normative, and organizational properties through which an AI system is evaluated against governance requirements, such as robustness, fairness, transparency, and accountability [2]. Assessments of system trustworthiness are typically conducted through formalized processes, including documentation, testing, red teaming, and institutional review, and operate largely independently of everyday system use.

Perceived trustworthiness refers to how users interpret and evaluate the behavior of an AI system in practice, using the cues available, including both primary and secondary cues [3] (those based on direct system use, or judgments or certifications from other stakeholders [15]), relative to their expectations, values, and situational goals. It may partially overlap with formal trustworthiness criteria, but also encompasses social and contextual dimensions that are not fully captured by technical or normative

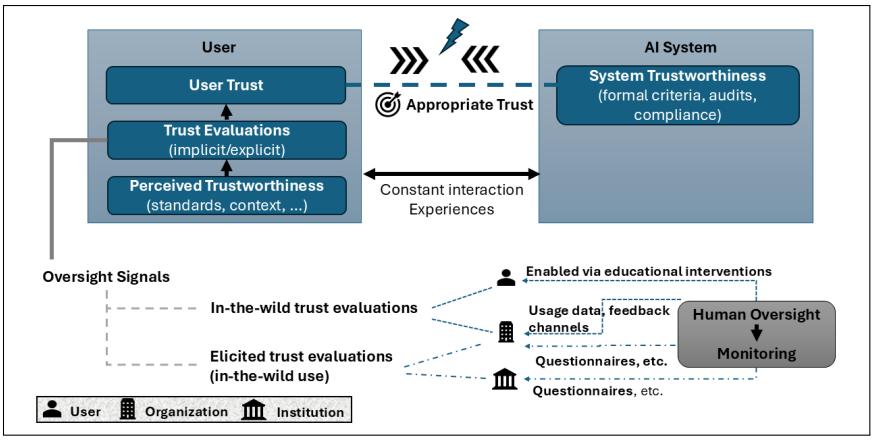

Figure 1: Vision of human oversight of an AI System capitalizing on understanding perceived trustworthiness.

assessments (e.g., secondary cues such as recommendations from a friend). Perceived trustworthiness constitutes the primary basis upon which user trust is formed, maintained, or withdrawn.

Building on this understanding, we conceptualize *trust evaluations* as the user's ongoing assessment of whether an AI system's behavior appears to meet their implicit trustworthiness standards in a given context. These assessments are shaped through interaction, may occur consciously or unconsciously, and can change over time as use, expectations, systems, and technologies evolve.

Post-deployment human oversight is understood as the set of processes through which human actors monitor, interpret, and intervene in the behavior of AI systems after deployment, aiming to support continued alignment with intended goals and acceptable use [16]. Oversight in this sense is ongoing and interpretive, rather than a one-off evaluative act, and is particularly concerned with behaviors and risks that emerge during real-world use.

3.2. From Perceived Trustworthiness to Oversight Signals

In this work, rather than addressing the full range of possible oversight interventions, we examine how trust-related evaluations produced through user interaction can support ongoing monitoring of AI systems in use. It is also important to note that different actors engage in oversight from distinct positions, which shape both what can be observed and how trust-related evaluations can be interpreted. Institutional actors, such as regulators or auditors, typically have access to aggregated and mediated information about system behavior. Organizational actors, including developers, may observe structured feedback, usage patterns, and internal monitoring data. By contrast, users observe AI systems through direct, situated interaction, forming trust judgments that may remain contextual, and inaccessible to external observers. Recognizing these differences is essential for our proposed vision. Not all trust evaluations are observable or actionable from every oversight position, and no single actor has complete visibility into how trustworthiness is experienced in use.

Understanding a user's perceived trustworthiness as a basis for AI system monitoring can be approached through two distinct modes: *capturing expressions of trust evaluations in the wild* (e.g., indirectly via digital traces) or *purposefully eliciting trust evaluations based on in-the-wild use* (e.g., via direct evaluation). In the first mode, trust evaluations surface organically through interaction with the system, for example via patterns of reliance, hesitation, disengagement, increased monitoring, or formal and informal feedback. Such expressions are often implicit, situated, and tightly coupled to specific contexts of use, making them primarily accessible to users themselves and, in some cases, to system providers through usage data or feedback channels. They are rarely directly observable by institutional actors.

In the second mode, trust evaluations are captured deliberately through structured elicitation processes, such as surveys [17, 18] or crowdsourcing tasks designed to solicit and subsequently analyze users' perceptions of trustworthiness grounded in their in-the-wild interactions with the system. These approaches make trust-related assessments externally observable by prompting users to articulate their judgments explicitly, often using standardized measures of trust or perceived trustworthiness.

Not all observations of the expression of trust evaluations are equally informative for system monitoring. Within our proposed vision, the expressions that are relevant for monitoring are conceptualized as *oversight signals*. Oversight signals may reflect trust, distrust, or the justifications underlying (dis)trust, as well as shifts in how trust is justified over time. They function as attention-guiding inputs in the monitoring process, rather than as direct assessments of system trustworthiness or compliance. In this sense, oversight signals constitute the conceptual bridge between individual trust evaluations and post-deployment monitoring activities. Figure 1. presents an overview of our vision for enabling post-deployment monitoring through oversight signals that stream from user's trust evaluations.

3.3. Aggregation and Monitoring

While oversight signals may emerge from individual instances of system use or from more deliberate and structured elicitation, there remains a need to meaningfully aggregate these signals for monitoring at organizational and institutional levels. Such aggregation must account for oversight signals across users, contexts, and time, not as a holistic measure of system trustworthiness, but as a mechanism for detecting potential risks or failures in AI systems as they are used. For aggregation to be informative, however, it requires a robust understanding of how users form and express perceptions of trustworthiness. In this sense, effective aggregation depends on the ability to interpret trust evaluations, shaped by human cognitive, social, and contextual processes, and translate them into representations that make visible the gap between experienced trust and formally assessed trustworthiness, thereby enabling systematic post-deployment monitoring.

4. Discussion & Future Directions

In presenting our vision, we also highlight an important limit of post-deployment human oversight: not all trust-related signals can be meaningfully externalized or aggregated. In many highly interactive AI systems, trust evaluations formed through everyday use remain accessible only to users and may never surface as organizational or institutional oversight signals. When such distributed signals cannot be reliably assessed at scale, alternative mechanisms for supporting oversight become necessary.

If oversight signals in the wild are primarily accessible to users, then enabling users to interpret system behavior, reflect on the cues that shape their trust evaluations, and recognize when trust is warranted or unwarranted becomes essential. Educational interventions that support trust calibration and the development of critical thinking in interactions with AI systems can therefore function as a form of user-centered human oversight, complementing institutional monitoring rather than replacing it. This perspective positions user education beyond a tool for increasing acceptance or usability, instead treating it as a mechanism for sustaining appropriate trust and mitigating emerging risks in deployed AI systems where centralized monitoring alone is insufficient.

As future work, we plan to further develop our vision into a conceptual framework and to ground it in empirical studies examining trust evaluations formed through the use of AI systems in-the-wild. In parallel, we aim to design and study mechanisms for deliberately eliciting trust evaluations and even doing so through educational interventions. Finally, we plan to explore how structured and elicited trust evaluations can be meaningfully combined with in-the-wild trust evaluations to support post-deployment monitoring and human oversight.

Acknowledgments

This project is partially funded by the EU's Horizon Europe Programme, agreement No. 101214000 (TRUMAN), the Cyprus RIF Programme, agreement No. BRIDGE2HORIZON/0823E/0203 (PINNACLE) and the EU's Erasmus+ Programme, agreement No. 2025-1-CY01-KA220-HED-000364525 (AI4Everyone).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.2 in order to check grammar and spelling. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 624–635.

[2] J. Mökander, J. Schuett, H. R. Kirk, L. Floridi, Auditing large language models: a three-layered approach, AI and Ethics 4 (2024) 1085–1115.

[3] N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, M. Langer, How do we assess the trustworthiness of ai? introducing the trustworthiness assessment model (tram), Computers in Human Behavior 170 (2025) 108671.

[4] P. Bisconti, L. Aquilino, A. Marchetti, D. Nardi, A formal account of trustworthiness: Connecting intrinsic and perceived trustworthiness, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, volume 7, 2024, pp. 131–140.

[5] X. Amatriain, A. Sankar, J. Bing, P. K. Bodigutla, T. J. Hazen, M. Kazi, Transformer models: an introduction and catalog, arXiv preprint arXiv:2302.07730 (2023).

[6] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, L. Chen, Generative ai and chatgpt: Applications, challenges, and ai-human collaboration, 2023.

[7] L. Yan, V. Echeverria, G. M. Fernandez-Nieto, Y. Jin, Z. Swiecki, L. Zhao, D. Gašević, R. Martinez-Maldonado, Human-ai collaboration in thematic analysis using chatgpt: A user study and design recommendations, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–7.

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

[9] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, S. Sousa, A systematic literature review of user trust in ai-enabled systems: An hci perspective, International Journal of Human–Computer Interaction 40 (2024) 1251–1266.

[10] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM computing surveys (CSUR) 55 (2022) 1–38.

[11] N. Schlicker, F. Lechner, K. Wehrle, B. Greulich, M. C. Hirsch, M. Langer, Trustworthy enough? examining trustworthiness assessments of large language model-based medical agents (2025).

[12] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, M. L. Tielman, A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges, ACM Journal on Responsible Computing 1 (2024) 1–45.

[13] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, S. Mohamed, Power to the people? opportunities and challenges for participatory ai, in: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2022, pp. 1–8.

[14] M. Young, U. Ehsan, R. Singh, E. Tafesse, M. Gilman, C. Harrington, J. Metcalf, Participation versus scale: Tensions in the practical demands on participatory ai, First Monday (2024).

[15] N. Scharowski, M. Benk, S. J. Kühne, L. Wettstein, F. Brühlmann, Certification labels for trustworthy ai: Insights from an empirical mixed-method study, in: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 2023, pp. 248–260.

[16] K. Kyriakou, J. Otterbacher, Modular oversight methodology: a framework to aid ethical alignment of algorithmic creations, Design Science 10 (2024) e32.

[17] S. Bayer, H. Gimpel, M. Markgraf, The role of domain expertise in trusting and following explainable ai decision support systems, Journal of Decision Systems 32 (2022) 110–138.

URL: https://doi.org/10.1080/12460125.2021.1958505. doi:10.1080/12460125.2021.1958505. arXiv:https://doi.org/10.1080/12460125.2021.1958505.

[18] M. J. McGrath, O. Lack, J. Tisch, A. Duenser, Measuring trust in artificial intelligence: validation of an established scale and its short form, Frontiers in Artificial Intelligence 8 (2025) 1582880.