Five Obstacles to AI Trustworthiness Alignment and How to Deal With Them: Introducing the DELIB-TRUST Framework

Pascal Koenig[1*], **Siddharth Mehrotra**[2*]

[1]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[2]Birla Institute of Technology & Science, Pilani, Pilani Campus, India
p.d.koenig@vu.nl, siddharth.mehrotra@pilani.bits-pilani.ac.in

Abstract

Aligning the perceived trustworthiness and trusting behavior of users with the capabilities and characteristics of AI systems is considered central to the adoption of AI systems in society. However, achieving such a trustworthiness alignment faces major practical obstacles, at least with more complex notions of trust that go beyond mere reliance on an AI system. The present article identifies five fundamental obstacles that arise in the trustworthiness assessment and alignment process and that threaten to make any trustworthiness alignment unfeasible in practice. To address these obstacles, we introduce a novel DELIB-TRUST framework which goes beyond the conventional causal-behavioral framework of trust calibration and incorporates elements of deliberation and reason-giving. We evaluated our framework by conducting an expert study with 12 HCI/AI practitioners and provide actionable steps for its implementation. Through its evaluation, we show how the framework can help close important gaps that exist due to unacknowledged value-laden decisions and interpretative steps in efforts to achieve trustworthiness alignment.

Introduction

Achieving trustworthy AI has become a key goal both for AI developers and policy makers. Trustworthy AI is deemed crucial for the acceptance of the technology, its societal uptake, the emergence of consumer markets, and, not least, responsible development and use of the technology. An important assumption in research on trust in and trustworthiness of AI is that the trustworthiness of a system is a function of design characteristics, such as the performance, the biases, and the transparency of a system (Kaur et al. 2023; Liao and Sundar 2022; Mehrotra et al. 2024a,b; Vereschak, Bailly, and Caramiaux 2021).

Trustworthy AI itself is not enough for societal acceptance of AI, though, as people also need to trust it (Braun, Bleher, and Hummel 2021; Laux, Wachter, and Mittelstadt 2024). Even an AI system that has been carefully designed may be perceived as untrustworthy. If, for instance, a system makes a crucial decision, such as recommending a medical treatment or providing financial advice, but cannot explain the rationale behind that decision in a way that is understandable to users, it may lead to skepticism or rejection. In this sense, people may under-trust certain AI systems. Yet, besides under-trusting AI systems, individuals may also over-trust AI systems, meaning that they place their trust in systems designed in a way that does not warrant. Therefore, a key challenge in shaping human-AI interactions in view of the growing proliferation of AI in people's lives is to align individuals' perceptions of trustworthiness with what has been called 'actual trustworthiness' of AI systems (Tomsett et al. 2020; Wischnewski, Krämer, and Müller 2023; Lucas, Becerik-Gerber, and Roll 2024; Schlicker et al. 2025a). Achieving such 'calibrated' or 'appropriate' trust means avoiding both too much and too little trust in an AI system[1].

While achieving alignment between actual and perceived trustworthiness seems plausible in abstract terms, it involves a demanding process with intricate methodological challenges when it is put into practice. Specifically, it requires: (a) making assessments of actual and perceived trustworthiness *i.e.*, arriving at valid and reliable measures, (b) finding ways to make valid comparisons between these two sides and also (c) to align them. The state of research leaves us with an important gap regarding this process. Previous work has outlined the general challenge of aligning the two sides of actual and perceived trustworthiness (Wischnewski, Krämer, and Müller 2023; Lucas, Becerik-Gerber, and Roll 2024; Mehrotra et al. 2024a; Schlicker et al. 2025a; Liao and Sundar 2022). It has pointed to relevant components that make up actual (Li et al. 2023) and perceived (Schlicker et al. 2025b) trustworthiness. It has furthermore highlighted the need for a communication process of conveying information about an AI system and its actual trustworthiness to influence and align trustworthiness perceptions (Schlicker et al. 2025a; Liao and Sundar 2022). However, prior work has not dealt with the detailed challenges and practical steps of measurement and establishing the equivalence of metrics required in this process.

In this article, we look more closely at *what makes the trustworthiness assessment and alignment process work and what it takes to put it into practice*. We ask what *practical challenges* arise from the HCI perspective in the processes of aligning actual with perceived trustworthiness of

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]For a detailed discussion and distinctions of these concepts, see Mehrotra et al. (2024a)

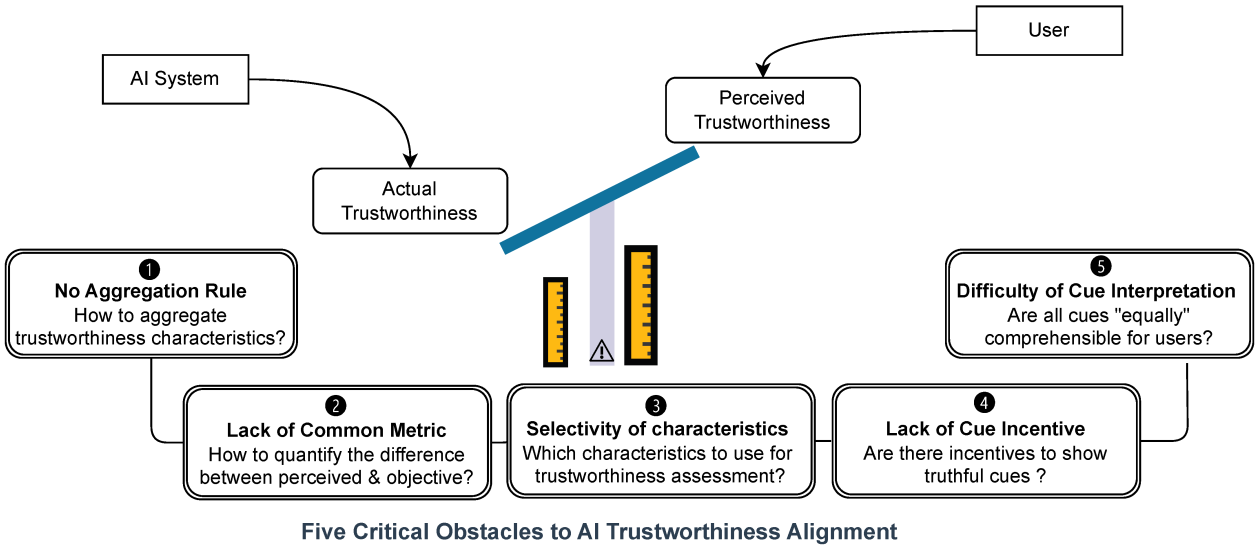


Figure 1: We identify and discuss five critical obstacles which need to be overcome to achieve an alignment between actual and perceived trustworthiness of an AI system. *The imbalance between and use of different scales for assessing actual and perceived trustworthiness represents the challenge of evaluating what it means to have comparable measures and an alignment or misalignment between the two sides. Also, the exclamation mark indicates the problematic nature of the trustworthiness comparison and alignment.*

an AI system. While existing research assumes that calibrating an appropriate level of trust is generally achievable (Mehrotra 2021), our analysis indicates that the practical feasibility remains an open question, at least with more complex AI agents. The conventional trust calibration paradigm operates under the assumption that trustworthiness can be objectively measured and effectively communicated to users through transparent design choices and explainable AI interfaces (Naiseh et al. 2021; Cancro, Pan, and Foulds 2022). This approach typically involves developing standardized metrics for system performance, reliability, and fairness, then translating these technical specifications into user-comprehensible formats such as confidence scores, uncertainty indicators, or explanatory visualizations. Providing users with accurate information on an AI system's capabilities and limitations can help users in forming appropriately calibrated trust levels that align with the system's actual trustworthiness profile (Liao and Sundar 2022; Wischnewski, Krämer, and Müller 2023; De Visser et al. 2020; Okamura and Yamada 2020; Zhang, Liao, and Bellamy 2020; Hoff and Bashir 2015; Mehrotra et al. 2024a, 2025).

However, aligning trusting behavior and perceived trustworthiness with actual trustworthiness faces major practical obstacles, specifically in the effort to make verifiable assessments of actual and perceived trustworthiness and to compare and align the two. This seemingly straightforward process masks significant complexities in both the assessment and communication phases of trust calibration (Jacovi et al. 2021; Ferrario 2025). The assessment of actual trustworthiness presents considerable challenges, particularly when moving beyond narrow technical metrics, such as reliability, to encompass broader notions of ethical AI behavior, contextual appropriateness, and social responsibility (Barocas, Hardt, and Narayanan 2023; Selbst et al. 2019). Even when having quantified actual trustworthiness, there is no clear standard of comparison to match it with measures of perceived trustworthiness. Furthermore, the assumption that users rationally process and correctly interpret trustworthiness information overlooks intricacies of cue interpretation and the inherently subjective and contextualized nature of trust formation. Research on human-AI interaction has already demonstrated that a one-size-fits-all approach to trustworthiness communication may be inadequate (Wang et al. 2021; Miller 2019) and needs to be more differentiated.

Building on and extending previous research, we will first discuss five practical obstacles to realizing an alignment of actual and perceived trustworthiness (Figure 1) and discuss ethical harms, specifically for people's autonomy, that result from ignoring these obstacles. We use this discussion as a basis for developing a complementary framework, DELIB-TRUST. This model departs from the behavioral foundations of the trust calibration process and incorporates elements of reason-giving: Specifically, (a) disclosing the reasons for choosing trustworthiness characteristics, (b) exploring the meaning of actual and perceived trustworthiness scores and discrepancies between them, (c) reflecting on cue interpretability and reasons for cue selection, and (d) reflecting on the working of the trustworthiness assessment and alignment process. With these additional ele-

ments, this framework accommodates the fact that the trustworthiness assessment and alignment process necessarily involves value-laden choices and interpretative steps for which there is no objectively correct solution. These choices remain implicit and unacknowledged with a behavioral approach to trust calibration which sees this calibration challenge as a matter of objective measurements and of using cues to causally influence trustworthiness perceptions.

Overall, we make three key contributions:

- First, we point out blind spots in the form of undertheorized aspects in the process of trustworthiness assessment and alignment by discussing key obstacles that, if left unaddressed, lead to gaps in that process and entail ethical harms.
- Second, from these gaps we derive a novel framework for aligning actual and perceived trustworthiness that takes the deliberative and socially constructed nature of trustworthiness into account.
- Finally, we validate the framework through an expert study with 12 HCI/AI practitioners to examine the descriptive, evaluative, and generative powers (Beaudouin-Lafon 2004) of the framework.

Background

While trustworthiness, as a propensity to not disappoint someone's trust, cannot be directly observed, concrete agent properties can be taken as an indication of trustworthiness. According to Mayer, Davis, and Schoorman (1995)'s model that has been highly influential for HCI research, three central dimensions that reflect trustworthiness are a trustee's benevolence, competence, and integrity. These concepts are tailored to interpersonal relations and presume intentionality of the actors in question. Intentionality here means that the actors do not simply *behave* in certain ways influenced by external causes, but attach subjective meanings to their actions which we cannot presume of AI systems. However, it is possible to adopt a behavioral interpretation of trustworthiness that does not depend on intentionality, as has been done with Lee and See (2004)'s seminal contribution on trust and automation. Rather than speaking of benevolence, competence, and integrity, they characterize AI systems in terms of goal alignment, performance, and consistency, respectively. Similar to this approach, subsequent conceptions of trustworthiness see trustworthiness as multi-dimensional in the sense that trustworthiness is tied to various attributes of an agent. Frameworks of trustworthy AI have similarly elaborated a range of characteristics that are held to constitute trustworthiness (Thiebes, Lins, and Sunyaev 2021; Kaur et al. 2023).

Realizing the alignment between perceived trustworthiness among AI system users and the trustworthiness of an AI system as a function of its characteristics faces various challenges at least with more advanced AI systems that have agent-like qualities. For simpler applications that are designed for performing a specific task, one can narrowly conceptualize the actual trustworthiness of the system as its reliability, a function of its performance and ability to achieve certain pre-defined outcomes. For instance, one can use a measure of the joint performance of an AI system and a human performing a given task (McDermott and Brink 2019). Based on this measure, one can assess how differences in the propensity to trust the AI system lead to different outcomes in the performance measure. What amounts to optimal trust is then what optimizes this measure of joint task performance. Reliability of the AI system and a person's trust in it are thus anchored in an external standard.

However, the trust problem becomes more complex with AI systems and for AI systems performing tasks for which we do not have clearly definable standards. Certain AI systems may show subtle biases in their behavior, meaning that their goals may diverge from that of their users. AI chatbots based on large language models show ethnic, gender, or political-ideological biases in their outputs that may be subtle and hard to detect (Choudhary 2025; Hartmann, Schwenzow, and Witte 2023; Reveilhac and Morselli 2025; Tao et al. 2024). In these cases, not only the performance and reliability of the system are of interest when thinking about their trustworthiness, but also their goal alignment (in analogy to benevolence), and their consistency in adhering to certain standards or principles (in analogy to integrity) are relevant dimensions. AI agents thus present major delegation challenges similar to human agents: they can have hidden goals, information, and behaviors (Krafft, Zweig, and König 2020). These systems may develop misaligned sub-objectives (Russell 2020) or exhibit deceptive behaviors like sycophancy producing outputs designed to please users rather than provide accurate information (Park et al. 2024). These behaviors complicate the notion of trustworthy AI and require more comprehensive trustworthiness assessment as well as cues to align user trust with system capabilities (Liao and Sundar 2022; Schlicker et al. 2025a).

An important approach for achieving an appropriate level of trust lies in assessing "actual" and "perceived" trustworthiness and using cues that shape people's trustworthiness perceptions (Liao and Sundar 2022; Lucas, Becerik-Gerber, and Roll 2024; Mehrotra et al. 2024a; Schlicker et al. 2025a; Wischnewski, Krämer, and Müller 2023). Research that has conceptualized this process to align trustworthiness has pointed to a set of characteristics that are relevant for *actual* trustworthiness, such as robustness, fairness, privacy, and explainability (Li et al. 2023). While these characteristics can be inspected individually, contributions that describe trustworthiness of AI systems recurrently see these individual characteristics as constituting the *overall* trustworthiness of an AI system (Bisconti et al. 2024; Lucas, Becerik-Gerber, and Roll 2024; Kaur et al. 2023; Schlicker et al. 2025a). Similarly, *perceived* trustworthiness is tied to a range of evaluative criteria, such as perceived performance, safety, and usability (Schlicker et al. 2025b).

Crucially, users of the AI systems may not get an idea of the trustworthiness of an AI system without being provided information about characteristics of the system (Schlicker et al. 2025a; Liao and Sundar 2022). When interacting with an AI system, users cannot infer from the outputs alone

what, e.g., the system's performance, biases, or fairness are. This would need to be systematically assessed and communicated to users in the form of information about the system, commonly understood as cues. Furthermore, various cognitive factors and people's mental models shape how people process cues and thus have consequences for the gap between actual and perceived trustworthiness (Bansal et al. 2019; Steyvers and Kumar 2024). To align perceived with actual trustworthiness, it is therefore essential to use cues from which users can form accurate perceptions of AI system characteristics that, in turn, inform their trustworthiness perceptions.

The idea of matching system design properties and perceived trustworthiness provides crucial guidance for AI system design and a communication of this design that caters to the needs and wants of humans. However, the viability of this general framework for achieving an appropriate level of trust depends on various assumptions that need to be met in practice. As we argue in the next section, this effort involves severe obstacles.

The Five Obstacles

Producing a measure of "actual trustworthiness" based on how an AI system is engineered and assessing perceived trustworthiness before making targeted interventions to align the two seems like a straightforward approach. However, making this approach work depends on realizing linkages and translation processes between the truster and trustee, and also between individual trustworthiness characteristics and overall trustworthiness both on the side of the truster and the trustee. Several obstacles emerge when trying to realizing these linkages.

Obstacle 1: Examine actual trustworthiness based the AI systems' properties - *Step 1*. The first obstacle arises when trying to combine individual trustworthiness characteristics. Trustworthiness is a global evaluation of an agent. Yet, the assessment of trustworthiness of an AI system will generally be a composite measure of relevant system characteristics, such as performance, fairness, transparency (Liao and Sundar 2022). When aggregating individual characteristics into an overall trustworthiness score, one faces a major challenge. One can devise or use existing measures, for example, for performance such as AUC-ROC or the F1 score, for the goal alignment or biases of AI systems (Wan et al. 2023), for the transparency (Bhatt et al. 2021; Holzinger, Carrington, and Müller 2020), or even a Language Explanation Trustworthiness (LExT) score (Shailya et al. 2025). However, having these assessments does not provide one with a recipe for combining these measures into a composite score. They all have different units and scales, and one would first need to make them commensurable to be able to add them up.

Yet, there is no single correct way of making individual characteristics commensurable. There is no ground truth to questions like what is the objective trustworthiness value for a given performance or explainability score of an AI system. For instance, having assessed the performance of an AI system at a task as AUC-ROC = 0.78 does not tell us that the system is X percent trustworthy or has a trustworthiness of X on a scale from 0 to 10. Even if this were possible, there would still be the additional problem of adding such new metrics up into overall trustworthiness because we have no true weights for the individual characteristics. Specifying these weights ultimately requires a value judgment. At this point, we thus run into the first obstacle, the *lack of an aggregation rule*.

Making system characteristics commensurable and aggregating them into an overall trustworthiness assessment is a matter of human discretion, since there exists no ground truth for such a translation process. This is a fundamental obstacle when trying to derive the overall "actual trustworthiness" of an AI system from a set of system characteristics. This obstacle is tied to the fact that developers and researchers necessarily bring their own positionality to their evaluation of AI systems, and that this can remain obscured as their own blind spot (Raji et al. 2020; Dobbe, Krendl Gilbert, and Mintz 2021). Any way of assessing actual trustworthiness can, therefore, only be a convention rather than yielding a true, objective underlying property of an AI system. *We thus identify a **first gap** at step 1 of the trustworthiness assessment and alignment process that exists as long as the obstacle of a lack of an aggregation rule is not overcome and value-laden choice is not made explicit.*

Measuring perceived trustworthiness - *Step 2*. For the perceived trustworthiness of an AI system, one can probe people's perceptions regarding trustworthiness characteristics like performance or transparency, and then examine how much each of them adds to the overall perceived trustworthiness among users (e.g., overall trustworthiness rating). This can be done statistically through probing how much evaluations of different characteristics contribute to overall perceived trustworthiness. Notably, these statistically and indirectly estimated *implicit* weights of the characteristics do not necessarily match the consciously stated *explicit* weights that people assign to those characteristics (Busacca and Padula 2005). For instance, users may on average state that transparency is the most important characteristic of an AI system for them (explicit importance), but when inspecting the associations between individual perceived characteristics and overall perceived trustworthiness, perceived transparency may show a comparatively weak association (indicating implicit importance) with this stated overall trustworthiness.

Nonetheless, obtaining scores for perceived trustworthiness of the AI system and of individual system characteristics is generally feasible. One can use semantically intelligible scores, such as through letting individuals rate their perceived trustworthiness on a scale from 1 to 5 (Wischnewski et al. 2024). In this way, they can provide both global evaluations of an agent's trustworthiness and evaluations of individual aspects such as performance and fairness. This step is straightforward and faces no obstacles when users' subjective weights have been determined. Measuring perceived trustworthiness does, however, face certain constraints that have important consequences for the subsequent stage involving the comparison of actual and perceived trustworthi-

ness.

It is not a feasible to ask people about technically detailed perceptions of an AI system, such as its performance (Lucas et al. 2016). We cannot realistically ask people how they perceive the performance of an AI system in terms of its AUC-ROC. They will hardly state that they think its AUC-ROC score is, e.g., 0.74. Rather, people will commonly have to rate performance or other features on intelligible scales, such as from 1 = very low to 5 = very high. One could also ask users openly about their perceptions and they might give less structured answers like "not biased at all" or "performing very well" (see, e.g., (Terblanche et al. 2023)). In any case, there are limits to asking people about their perception of characteristics deemed relevant for trustworthiness. In practice, there will be no constellation in which one obtains both objective measures and perceptions that are on the same scale. Due to these limitations further obstacles arise when trying to compare actual and perceived trustworthiness measures.

Obstacles 2 and 3: Quantify the discrepancy between the assessed and perceived trustworthiness - *Step 3*. At the stage of comparing the two sides of trustworthiness, we run into the second obstacle, the *lack of a common metric*. There is no underlying shared natural metric for both actual and perceived trustworthiness measures that would allow us to interpret the difference between these two measures both regarding the overall trustworthiness assessments and evaluations of individual trustworthiness characteristics. Yet, without a common measurement metric, one cannot properly assess the discrepancy between them.

Numeric measures of the actual and of the perceived trustworthiness may be formally comparable, but we do not know if, e.g., the score 0.8 for actual trustworthiness has the same meaning as 0.8 for perceived trustworthiness. The same holds true for the evaluation of individual characteristics of trustworthiness, such as performance and transparency. For the perceived trustworthiness score, we may know, based on the rating questions posed to users, that a score of 0.8 reflects the semantic content "very trustworthy", which is more than "somewhat trustworthy" and less than "completely trustworthy". For the actual trustworthiness score, one would need to translate the numbers of characteristics like a performance score into such semantic categories to make the scores comparable. Yet, we cannot presume that this works in the same way as for the perceptions of trustworthiness, and the problem of a lacking common metric reappears.

This translation process is again one that involves value judgments and that is necessarily based on someone's interpretation of trustworthiness characteristics like relevant performance scores for an AI system. One must decide about the threshold for saying that a certain performance scores makes a system "very trustworthy" [2]. Comparing the two sides of trustworthiness in our example is further compounded by the fact that our overall perceived trustworthiness score may be informed by other criteria that we have not considered. Perhaps it is also determined by the usability of the system at least to some extent. This equally casts doubt on whether we are actually comparing the right measures.

Another form of discretion enters the act of comparing the two sides of trustworthiness: actual and perceived due to a third obstacle, that of the *selectivity of trustworthiness characteristics and concrete measures*. Which concrete characteristics should one use for a trustworthiness assessment and the eventual comparison in the first place? This is not a problem of commensurable *metrics*, but instead one of comparable *sets* of characteristics concerning the AI system and users' perceptions of the AI system. Developers of the AI systems and those doing the trustworthiness assessment can select from a broad range of concrete characteristics and specific measures for them (Li et al. 2023; Lekadir et al. 2025). Indeed, for measures of aspects such as explanation quality, HCI researchers appear to be divided on which measures are the most suitable (Markus, Kors, and Rijnbeek 2021; Miró-Nicolau, Jaume-i Capó, and Moyà-Alcover 2025). In sum, by not explicitly dealing with obstacles number two (lack of common metric) and three (selectivity of characteristics and measures), we end up with an important **second gap** in the trustworthiness assessment and alignment process. *What is missing at the third stage of this process is a way of making sense of how the measures for actual and for perceived trustworthiness relate to one another and how we can interpret differences between them.*

Obstacle 4: Choose suitable cues for reducing the discrepancy between actual and perceived trustworthiness - *Step 4*. Having registered the difference between actual and perceived trustworthiness, the trustworthiness assessor devises cues for reducing the discrepancy. A central question is where the cues about trustworthiness of an AI system originate. Some contributions suggest that such cues should come from the systems and their developers themselves. This approach is implicit, for instance, in the account by Liao and Sundar (2022), which also posits that 'good' trustworthiness cues must be truthful (Liao and Sundar 2022). In their description of an AI use case, however, they focus on other conditions that need to be met for good cues while simply assuming that the truthfulness condition is satisfied for the creators.

In practice, we cannot take for granted that providers use truthful cues, though. They primarily want to market products (Belanche et al. 2024). If people trust their product more than it is warranted, providers would need to communicate cues which lower trust to achieve trust calibration. Yet, they hardly have an incentive to send signals which are unfa-

[2]This translation problem becomes particularly intricate with measure of complex characteristics. For instance, evaluation metrics used to assess text generation quality of explanations in LLMs are designed to provide an indication of their trustworthiness. Yet, these measures may not comprehensively cover relevant aspects. For instance, BERT embeddings neglect critical dimensions like consistency and robustness to counterfactual or paraphrased prompts, which are essential for a thorough evaluation of explanation reliability (Devlin et al. 2019). This makes it difficult to state how a given score for the evaluation metrics translates into trustworthiness.

vorable for their products and would hurt their sales. Even if providers are legally compelled to provide truthful cues that are relevant in the sense that they reflect trustworthiness characteristics, providers may still do so but mix these cues with other cues that are especially effective in influencing trust without being relevant. The latter kind of cues could be design aspects that enhance usability or anthropomorphizing features of artificial agents that increase people's propensity to use them. Importantly, banning the use of such additional cues hardly seems an option. The provider can justify them with user experience and user satisfaction, which does make them truthful.

This fourth obstacle, the *lack of incentives to select relevant cues*, is not a problem for the system provider as they are the ones selecting cues. Rather, it is a problem for the goal of properly aligning actual and perceived trustworthiness. This obstacle thus has to be addressed by other parties, a point we will come back to when discussing possible ways to address the identified obstacles.

Obstacle 5: Difficulty of Cue Interpretation - *Step 5*. The assessor of the system does, however, generally face the fifth obstacle, the *difficulty of cue interpretation*. When providing individuals with cues that are relevant for the evaluation of trustworthiness, we cannot presume that individuals understand these cues. As situational awareness theory (Endsley 1995) posits, users need to have certain capabilities to process information about technologies to extract actionable knowledge and may need special training to understand AI system properties.

It is precisely because individuals cannot reliably use cues to infer properties of an AI system that many possible cues are useless or can be manipulative and that there may be incentives to provide irrelevant but effective cues. Cues about AI system properties can evoke trust even when trust is not justified, as is the case, for instance, with usability features that make systems more likable and foster affective trust (Ciriello, Chen, and Rubinsztein 2025). Indeed, AI systems can be designed to evoke positive emotions among users that may compensate for functional deficiencies and promote use of the technology (Koenig and Mehrotra 2025). This means that cues may be *causally effective* in bringing about trust in an AI system, but these cues may not reflect underlying characteristics that could be justified as warranting trust.

Further, following cue utilization theory (Koriat 1997), the cues that people actually utilize can differ from valid cues that accurately reflect certain attributes or outcomes. Liao and Sundar (2022)'s account of trustworthiness cues addresses this issue as it posits that good trustworthiness cues need to be not only effective but also relevant, *i.e.*, reflect characteristics that are central to trustworthiness, such as performance and biases. Still, the cues that convey the most relevant information about AI system trustworthiness may be the hardest for users to understand. Such cues may not be understandable to the vast majority even in a basic sense. Technical assessments of characteristics like the performance, fairness, biases, or explainability will hardly be comprehensible to most users. Regarding explainable AI methods, for instance, different ways of providing explanations about AI systems are not equally comprehensibly for different audiences (Brennen 2020).

Not only do there exist different measures (Koyejo et al. 2014), but even the scores for a given measure do not tell one whether the score is "good". For performance scores like the AUC-ROC, it is debatable and ultimately a matter of convention what counts as a "good" performance, what should be the threshold (Bartolucci, Singh, and Bae 2016; Lenat and Feigenbaum 1991; Leushuis et al. 2009; Padhy et al. 2019). Just like there is a certain discretion in choosing a threshold at which to categorize, e.g., performance as "good", there is equally discretion involved in translating such "good" performance into a trustworthiness score.

Finally, it may be acceptable that the assessor uses cues that are effective even while they are not relevant and thus foster trust in ways that are detached from the actual trustworthiness. This may be the only way to reduce a discrepancy between actual and perceived trustworthiness, which is a commendable goal. Doing so can go both ways, *i.e.*, to avoid that users are trusting far too much or too little. However, merely influencing users' trustworthiness perceptions in this way, i.e. causally and without them necessarily aware of it, is potentially deceptive. *In any case, what is missing at the stage of signaling via cues about the AI system is additional information that makes cues interpretable and makes transparent why they have been chosen as signals that should warrant trust.*

The DELIB-TRUST Framework

If the obstacles described further above remain unaddressed, the trustworthiness assessment and alignment process remains incomplete due to several gaps which also create ethical problems. As we argue in the following, additional elements are needed in the process to close these gaps. To close the gap that occurs when assessing actual trustworthiness (first gap), assessors need to go beyond merely applying or positing a certain way of aggregating a select set of assessed AI system characteristics. The step of aggregation involves a use of discretion that must be acknowledged as there are always alternative ways of doing the aggregation. Assessors thus need to disclose the aggregation rule they use and justify it as suitable. As a part of this disclosure, they need to spell out why they assume that the scales of the assessments for different trustworthiness characteristics can be made comparable. They also need to specify why they chose certain weights for these characteristics when creating the composite trustworthiness score even if they simply assumed equal weights. For this purpose, assessors could use use standardized reporting sheets in which they provide reasons for the approach in a structured fashion. This could also entail clarifications of why alternative possible approaches have not been chosen. With this requirement of reason-giving, a deliberative element enters the first stage of the trustworthiness assessment and alignment process (DELIB-TRUST layer 1).

Dealing with the interpretative gap that occurs when trying to assess the discrepancy between actual and perceived trustworthiness (second gap) equally requires an element of reason-riving. The assessor needs to make sense of the

scores resulting from the actual and perceived trustworthiness assessment and how these scores relate to each other. To understand what the scores mean, one needs to obtain users' reasons for why they perceive the trustworthiness of an AI system at a given level and what role different criteria play in this regard. Users would need to be given the opportunity to *consciously* reflect about obtained trustworthiness scores and what they mean in terms of semantic categories. Assessors might, e.g., score the fairness of a system low while users perceive its fairness as high and probe whether users would evaluate the system differently if they had the same information as the assessors. In line with participatory modeling processes (Lee et al. 2019), assessors would thus establish a common understanding of the trustworthiness assessment process and its results both among users and assessors. In this way the assessors can probe whether the registered numeric differences between objective and perceived trustworthiness are understandable for and justifiable toward users. With this additional element of deliberation (DELIB-TRUST Layer 2), assessors can better understand what lies behind the calculated discrepancy between actual and perceived trustworthiness.

Finally, closing the gap that arises when using cues to influence perceived trustworthiness (third gap) requires mechanisms that promote truthful and relevant signaling of trustworthiness characteristics in a way that is also effective. For this to be the case, those using the cues will need to provide additional explanations and justifications, especially if they choose cues that are more effective but less relevant. Specifically, they will need to provide information that makes their cues interpretable and give reasons for the choice of their cues, elaborating on and justifying the chosen trustworthiness cues and why these signals warrant trust (DELIB-TRUST Layer 3). As part of this approach, assessors may also use labels that visually assure a high performance, fairness, and transparency, and that refer to a credible source (Fischer, Liebig, and Morik 2024; Scharowski et al. 2023). While the design of the trustworthiness alignment process matters for whether truthful cues are interpretable for users, overcoming the obstacle of a lack of incentives of providers to choose relevant cues is not possible from within this process. It requires mandating that independent third parties, and not providers, use cues such as certification, labels, and additional information (Fischer, Liebig, and Morik 2024; Laux, Wachter, and Mittelstadt 2024). Note, however, that this pushes the trustworthiness problem one step further as the third parties must be seen as trustworthy themselves for cues to be credible and effective.

All in all, the above-mentioned three elements imply that assessors must disclose their own inevitable positionality and standards when it comes to assessing AI system trustworthiness and devising ways to achieve alignment between actual and perceived trustworthiness. The three deliberative elements that are crucial for complementing the trustworthiness assessment and alignment process introduce moments of reflection on assessments and the cues used for the intended alignment. They ensure that assessing AI trustworthiness and influencing users' perceptions of trustworthiness do not occur *behind users' backs* and safeguard the autonomy of those who must do the trusting.

Precisely because the DELIB-TRUST framework adds elements of reason-giving at various stages that create a space of discretion and value judgment, it creates an occasion for assessing the adequacy of these elements and their contextualized application. Based on the preceding considerations, we thus also identify an overarching meta-layer as a key part of the DELIB-TRUST framework. This fourth layer establishes ongoing reflection and revisability of the entire trustworthiness assessment and alignment process and allows for evaluating the need to make adjustments in the implementation of the other three layers (DELIB-TRUST Layer 4).

Overall, the DELIB-TRUST framework conceptualizes trustworthiness as a socially constructed phenomenon requiring elements of structured interpretation and deliberation besides objective measurement. It thus moves beyond the traditional behavioral paradigm and the "cue $\rightarrow$ trust" model, embracing the unavoidable interpretative and normative dimension of trustworthiness assessments. This framework operates through four interconnected components, each addressing specific obstacles to effective trust calibration:

1. **Disclosure Layer:** This layer forms the foundational transparency component, requiring explicit documentation of assessment methodologies, rationale for characteristic selection, and justification for aggregation approaches. This layer directly addresses the lack of pre-existing aggregation rules by mandating transparent explanation of how diverse trustworthiness characteristics are weighted and combined.
2. **Evaluation Harmonization:** This layer creates semantic bridges between technical metrics and user perceptions, addressing the absence of common measurement metrics. The harmonization process also recognizes that different stakeholder communities may require different interpretive lenses while working toward appropriate trustworthiness perceptions.
3. **Cue legitimation:** This layer establishes mechanisms promoting the effective, truthful and relevant signaling of trustworthiness characteristics. It acknowledges that effective trust calibration requires not just suitable cues but also cue interpretation capabilities among diverse user communities. Third-party certification processes and regulatory frameworks create accountability structures that reward genuine transparency and penalize misleading representations.
4. **Deliberative Processes (Meta-level):** Finally, the deliberative processes layer provides structured mechanisms for multi-stakeholder dialogue, argumentation, and consensus-building for reflecting on the trustworthiness assessment and alignment process. It recognizes that trustworthiness judgments often require contextual interpretation that cannot be predetermined through algorithmic means alone.

Evaluation of the Framework

To evaluate our framework, we utilize the three powers: descriptive, evaluative, and generative of interaction models

described by Beaudouin-Lafon (2004). We conducted an expert review asking HCI/AI practitioners to examine the value of the framework in ways that relate to those three powers. Beaudouin-Lafon (2004) and report on any new insights the experts gained with the framework and any challenges in applying it. Overall, we conducted semi-structured interviews with 12 HCI/AI practitioners who have varying experience with trustworthiness assessment based on their work or prior publications. We recruited participants from our social and professional networks. The interviews lasted approximately 60 minutes and were conducted remotely.

Our sample was composed of 7 women and 5 men aged between 30 to 42 years (M = 33.14). All participants were from Europe. We spoke with two PhD candidates in their last year of PhD, three postdoctoral researchers, a medical doctor who is also a data scientist working at a public health care institute, one AI certification manager and one solution architect at a technical inspection association, two industry design (MS) and user experience (Ph.D.) researchers and one assistant and one associate professor. All of them work in the domain of trust in human-AI interaction [3].

Our interview protocol consisted of two phases. In the first phase, we provided a detailed explanation of the five obstacles to trust calibration with the help of Figure 1 and then discussed the four layers of DELIB-TRUST framework as a means to study and overcome these obstacles. During the second phase of our interview protocol, we asked them to critically examine our framework with questions relating to its descriptive, evaluative and generative power. A list of questions asked can be found in the appendix.

Descriptive power: We assessed our framework's descriptive power by asking participants about how the framework covers existing challenges linked with trustworthiness assessments. We found that through complementing the trustworthiness assessment and alignment process with the layers of the DELIB-TRUST framework, it is possible to comprehensively cover the problem space. As the interviews with practitioners showed, the deliberative elements described above allow practitioners to characterize what occurs during trustworthiness assessment more completely.

One participant, P8 stated that at least to some extent, processes of justification already occur, but often stay hidden. When creating AI systems, developers make choices that they document and for which they give reasons. They choose certain metrics and justify them. Yet, these deliberative elements, according to the interviewee, often remain inchoate, obscuring the expert consensus and value judgments underlying seemingly objective standards. For instance, a justification can simply be a reference to a state of the art. Yet, this state of the art is itself the product of expert deliberations that led to a consensus about, e.g., suitable metrics. The framework thus highlights the conventionalized nature of existing approaches and measures. Similarly, P9 mentioned that fairness metrics can be a contentious issue, and that the choice of a certain metric and also of cut-off points for acceptable scores need to be discussed and justified. Ultimately, without transparency of the way in which trustworthiness is assessed, P12 stated, overall trustworthiness becomes a black box.

Another expert (P1) observed that "*individual standards influence the actual trustworthiness*," noting that fairness, for example, could mean equal distribution or need-based allocation depending on context. The framework recognizes that "*different ways to conceptualize fairness*" exist, and that the Disclosure Layer must make explicit not just which characteristics were chosen, but which definitions of those characteristics were applied. As the participant noted, "*it should be clear what the trustworthiness alignment refers to, [...] that there might be different standards among the stakeholders*." P1, P8 and P9 all pointed to the value of the framework's disclosure layer making hidden choices and their justification explicit, requiring transparent explanation of how diverse trustworthiness characteristics are weighted and combined attesting to the value of the framework to describe real problems.

Experts also found that the identified gaps from Figure 1 describe real practical problems. As one participant explained regarding healthcare AI and the issue of cue selection (P2): "*If my system is just 60% accurate in terms of prediction... is there any incentive, especially for the AI system provider or deployer, to show less accuracy for its users?*" Another expert (P7) elaborated on the problem of cue interpretability among different stakeholders, noting that "*what doctors would need would be very different from what nurses would need, what hospital management staff would need, and what patients would need*". P3 highlighted the framework's sociotechnical focus as both a strength and limitation: "*I think it focuses a little bit more on the process of implementing [AI] systems which is utmost important but... I don't think it covers as much when it comes to how do we build these systems.*" To give an example, this expert noted, "*your solution to the fact that there's no way to aggregate [trustworthiness dimensions] is to be transparent about how you aggregate which makes sense from a sociotechnical perspective while acknowledging the technical aggregation still remains a challenge.*"

Proposed refinements: Participants suggested that the connection between obstacles and framework components should be made more explicit, possibly through direct labeling in the figure itself. Additionally, the role of different actors (service providers, deployers, researchers, users) in each layer could be clarified within the framework to prevent the perception that it represents a linear process rather than concurrent activities by different stakeholders.

Evaluative power: Evaluative power is about the ability to make comparisons in the problem space. The added evaluative power of the DELIB-TRUST framework lies in a better ability to assess and compare the validity of trustworthiness alignments. Particularly the framework's emphasis on context emerged as a key evaluative strength in the interviews. P5 noted, "*it very much emphasizes context here*," contrasting this with traditional trust scales. P5 illustrated this using the Human Robot Trust scale, where dimensions

[3]Within our community, our sample could be readily identifiable. To protect the anonymity of our participants, we do not provide a mapping of demographic information to participant IDs.

like 'ethical' may not always be relevant, despite being crucial in other situations, leading to difficulties with the "not relevant option." P10 agreed that the framework is useful because "*it drills that down a little bit more. That the context there is important for what's relevant*," demonstrating that this contextual sensitivity enables more meaningful and relevant comparisons.

P2 provided a healthcare-specific validation of the evaluative power of the framework. When discussing the disclosure and evaluation harmonization components, P2 emphasized that for clinicians, "*interpretability is important because they wouldn't do something they don't understand, even if they see a very high accuracy.*" This highlights how the framework enables evaluation of whether trustworthiness assessments address the right dimensions for the right stakeholders, rather than applying one-size-fits-all metrics.

Another interviewee, P9, noted that the framework could highlight and help to deal with a lack of comparability of trustworthiness assessments as it would lead to more transparency regarding these assessments while clarifying whether different underlying notions of trustworthiness have been used for different assessments. Similarly, P8 estimated that the framework would not itself lead to comparability of trustworthiness assessments since its layers are themselves of a deliberative sort and thus not fixed. Yet, the framework could guide efforts to achieve comparability of trustworthiness assessments. It would do so by highlighting important aspects that need to be included in the comparison. As P3 noted, the framework helps "*compare existing practices*" by making explicit the normally implicit choices even if it doesn't prescribe universal standards for those explanations.

Proposed refinements: The evaluative power could be strengthened with more guidance on when and how to compare assessments that use different standards or definitions for trustworthiness characteristics. Further, while the sociotechnical focus is appropriate, participants suggested acknowledging more explicitly where technical implementation choices affect comparability, even if the framework may remain agnostic about which technical approaches to use. For example, two organizations might assess "robustness," but one uses adversarial testing on a closed model while another relies on red-teaming an open-source system. These different technical setups can make direct comparison of their trustworthiness claims difficult.

Generative power: The generative power of the DELIB-TRUST framework lies in its ability to generate new questions that can help identify instances in which the alignment process can become derailed. It contributes to a more reflexive design process in which discretion is made explicit and subjected to deliberation. By adding the deliberative elements described above, the framework can inform new approaches at the various stages of the trustworthiness assessment and alignment process, informing standard procedures for implementing this process. These potentials of the framework are reflected in the interviewees' responses. For instance, one participant (P3) noted how applying the Evaluation Harmonization principle can prompt AI system deployers to question whether their definition of "fairness" aligns with how affected communities actually understand fair treatment which in turn can lead them to design or redesign stakeholder consultation processes before finalizing their assessment criteria.

According to one of the interviewees, P9, the framework would raise important questions of who was involved in the trustworthiness assessments and alignment process and made key decisions. Referring to AI in health care, P9 pointed to the importance of including practitioners and ideally also patients in the attempts to make sure that AI systems are trustworthy since AI could alter the trust relationship between patients and health care practitioners. Notably, the interviewee also pointed out that these challenges are not fundamentally new and were already present since practitioners needed to become increasingly capable to assess the reliability of statistical evidence in health care.

P8 stated that the framework emphasized the bigger issue of the discrepancy between actual and perceived trustworthiness while current efforts in practice are directed mainly at detail questions of how to assess individual aspects that are relevant for trustworthiness. P8 furthermore noted that a current focus on technical aspects of trustworthiness assessments means largely sidelining the perception side of trustworthiness. Indeed, for a mere conformity and compliance assessment, P12 noted, assessing characteristics relevant for trustworthiness is already enough from the point of view of an assessing body. Yet, P12 also stated, the perception of trustworthiness in relation to an assessed actual trustworthiness is arguably of particular interest for developers as these want to see their products getting adopted.

Proposed refinements: Interviewees suggested that the role of different stakeholders would need to be part of the deliberation process envisioned by the framework and name of the incentive alignment layer would be more accurate as cue legitimation. Based on the suggestions provided by our experts, we visualize our framework with additional clarifications in Figure 2.

Discussion

Auditing frameworks like the ISO/IEC 42001 and the NIST AI Risk Management Framework posit sets of criteria compliance with which is deemed crucial for trustworthy AI systems. They do not specify, however, how those criteria could be combined to measure overall trustworthiness. Yet, such an assessment would be needed to compare AI systems regarding their trustworthiness. Further, those frameworks are not concerned with comparing assessed trustworthiness of AI systems with users' perceptions of trustworthiness. Yet, this comparison is important for identifying discrepancies that indicate too little or too much trust.

Problematizing such discrepancies, research has conceptualized the assessment of both actual and perceived trustworthiness of AI systems and ways of aligning them (Bisconti et al. 2024; Lucas, Becerik-Gerber, and Roll 2024; Kaur et al. 2023; Schlicker et al. 2025a). However, like existing auditing frameworks, this research leaves the epistemic authority with the assessor of trustworthiness, assuming that the assessor can arrive at a true measure of trustworthiness that justifies altering users' perceptions through

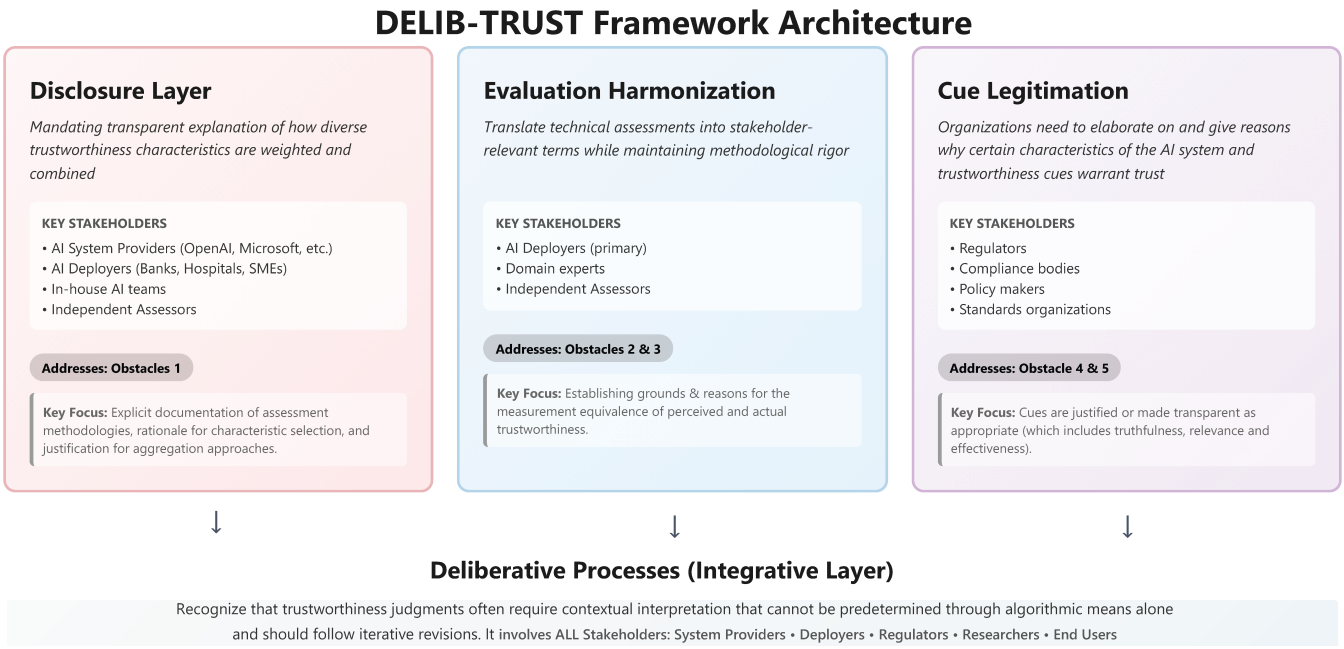


Figure 2: The revised DELIB-TRUST Framework addressing five obstacles to trust calibration through integrated components with key stakeholders and focus.

suitable cues. The DELIB-TRUST framework differs in this regard as it understands any assessment of the worthiness of trust as a claim that needs to be made explicit and validated by the trustees. It adds several deliberative elements to the trustworthiness assessment and alignment process that are needed to address discretion and value-laden choices which are always made, even if only implicitly. These deliberative elements entail that the process of assessing and aligning trustworthiness cannot take place behind users' back. In this way, it goes beyond a behavioral approach of measuring and causally influencing alignment between actual and perceived trustworthiness which can occur without users being aware.

The behavioral model of trust calibration is compatible with a wide range of use cases. If there is a clearly defined joint task and criterion of joint task performance of a human and AI agent, trust calibration can be anchored in this criterion. In other cases, the behavioral model works because trustworthiness can boil down to a question of performance and reliability, such as with an interpretable recommender system that performs a single clearly defined task and with known possible biases. A consensus on how to assess such reliability is then conceivable. The behavioral approach of trustworthiness assessment and alignment will arguably work well, for instance, in aligning reliance on an AI system with the prediction accuracy of the system (Wischnewski, Krämer, and Müller 2023; Klingbeil, Grützner, and Schreck 2024; Schemmer et al. 2023; Huang et al. 2025). Yet, with more complex systems, the goals that the system may realize and the biases that it can show are more varied and intricate, as evidenced by more elaborate forms of deception by certain AI systems (Park et al. 2024).

Assessing such more comprehensive and demanding notions of trustworthiness involves value-laden decisions about which characteristics and concrete measures to choose, and how to weigh and jointly translate them into an overall trustworthiness measure. There is no a priori suitable way of doing this. While it is certainly possible to score an AI system on various characteristics, such as performance and fairness, it is unclear what such measures and corresponding numbers mean in terms of trustworthiness as an inherent, objective property. This is always also a question of interpretation and judgment, not just measurement.

Conclusion

At the heart of the practical challenges in achieving appropriate trust for an AI system lies a fundamental question: how do we conceptualize and understand "actual trustworthiness" of a system in the first place. With simpler AI systems or when assessing the the joint performance of humans and machines, such actual trustworthiness can be reduced to reliability. However, with more complex AI systems that are agent-like and can show more complex behaviors and biases, a more complex notion of actual trustworthiness involving various relevant characteristics is warranted. It then becomes important to acknowledge that any assessment of 'actual trustworthiness' is conventional rather than an inherent objective property of an AI system. Without acknowledging this, we risk treating trustworthiness as a natural, measurable property while it is fundamentally a social and contextual judgment. This is why we propose the DELIB-TRUST framework to overcome the identified obstacles. We hope with this work, the AI ethics community acknowledges the discussed obstacles and our proposed framework paves the path to overcoming them.

Acknowledgments

We would like to thank Nadine Schlicker, Myrthe Tielman, Agathe Balayn, and our interviewees for their comments on earlier versions of the manuscript. We also thank the anonymous reviewers for their comments and suggestions.

Funding

This research has been supported by the Fritz Thyssen Foundation (Grant number 20.25.0.010PO).

Research Ethics and Social Impact

Ethical Considerations Statement

This study was approved by the ethics committee of the researchers' university (approval number 2025-11-24-461). All twelve participants took part voluntarily in semi-structured interviews conducted remotely; informed consent was obtained prior to each session. To protect participant anonymity, demographic characteristics are reported only at the aggregate level, and no mapping of demographic attributes to participant identifiers is provided, given that the relatively small and professionally identifiable sample could render individuals recognizable within their community.We note that participants were recruited through the authors' social and professional networks. While this facilitated access to relevant expertise, it introduces a potential for selection bias toward perspectives already sympathetic to sociotechnical framings of trustworthiness. Participants were exclusively European-based HCI and AI practitioners; the viewpoints of AI users, affected communities, or practitioners from non-Western regulatory contexts were not directly solicited. The framework's evaluation therefore reflects the deliberative norms and professional cultures of a specific geographic and disciplinary community, a limitation that must be borne in mind when generalizing its prescriptions.

Adverse Impact Statement

Several adverse consequences of this work deserve acknowledgment. First, the DELIB-TRUST framework, by formalizing deliberative requirements across four layers, could be absorbed into existing organizational practices as a compliance ritual rather than as a genuine shift in accountability, a risk Balayn et al. (2023) identify as "checkbox culture." Organizations may perform reason-giving for aggregation rules or cue selection without meaningfully opening these choices to scrutiny or external challenge. The framework provides no enforcement mechanism against such cooptation. Second, by cataloguing the five obstacles as fundamental and potentially insurmountable, the paper risks discouraging practitioners from attempting trustworthiness alignment at all, particularly in resource-constrained public sector or civil society contexts where such alignment is arguably most urgent. The paper should be read as motivating structured deliberation, not as validating abandonment of the effort.

Third, the deliberative overhead introduced by the framework is not equally distributed. Large technology providers can absorb the costs of systematic disclosure and stakeholder consultation; smaller deployers, community health organizations, municipal governments, social enterprises may not. If the framework becomes a de facto regulatory expectation without corresponding support structures, it may inadvertently concentrate legitimate AI deployment among well-resourced actors. Fourth, the framework's evaluation rests on practitioners who are themselves producers of trustworthiness discourse. End users particularly those in low-rights situations such as welfare recipients, patients in under-resourced healthcare systems, or communities subject to predictive policing were not consulted. The framework's claims about user autonomy protection are therefore made without direct evidence from the populations whose autonomy is most at stake. Finally, the Cue Legitimation layer could, in practice, be wielded by providers to rationalize existing trustworthiness communications rather than interrogate them. Requiring that cues be "justified or made transparent as appropriate" leaves open who adjudicates adequacy, and regulators assigned this role may lack the technical capacity or political independence to do so meaningfully.

Social Impact

The DELIB-TRUST framework has the potential to produce meaningful positive effects in the governance of AI systems, particularly in high-stakes public sector deployments. By making explicit the value-laden and interpretative dimensions of trustworthiness assessment that are currently obscured behind seemingly objective metrics, the framework supports more contestable and democratically accountable AI governance. Citizens and civil society actors gain a vocabulary and a procedural expectation with which to challenge how AI trustworthiness is constructed and communicated, directly strengthening user epistemic autonomy. The framework is also timely with respect to emerging regulatory requirements.

The EU AI Act mandates transparency and human oversight for high-risk AI systems, but provides limited guidance on how trustworthiness is to be assessed or communicated. DELIB-TRUST's disclosure and evaluation harmonization layers could inform conformity assessment procedures and the work of notified bodies, offering a structured basis for audits that go beyond narrow technical performance benchmarks. By foregrounding multi-stakeholder deliberation, explicitly naming regulators, domain experts, deployers, and end users as co-participants in trustworthiness assessment, the framework opens a space for marginalized communities to participate in decisions that affect them. In healthcare, criminal justice, and social benefit contexts, this participatory orientation could shift trustworthiness assessment from a technical exercise conducted by providers toward a sociotechnical process that reflects the values and risk tolerances of affected populations. Over time, the meta-level deliberative layer (Layer 4) further supports iterative refinement, enabling trustworthiness governance to adapt as AI systems evolve and as new forms of harm become apparent.

References

Balayn, A.; Yurrita, M.; Yang, J.; and Gadiraju, U. 2023. "Fairness Toolkits, A Checkbox Culture?" On the Factors

that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 482–495. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1): 2–11.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Bartolucci, A.; Singh, K. P.; and Bae, S. 2016. Introduction to Statistical Analysis of Laboratory Data.

Beaudouin-Lafon, M. 2004. Designing interaction, not interfaces. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, 15–22. New York, NY, USA: Association for Computing Machinery. ISBN 1581138679.

Belanche, D.; Belk, R. W.; Casaló, L. V.; and Flavián, C. 2024. The dark side of artificial intelligence in services. *The Service Industries Journal*, 44(3-4): 149–172.

Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; Nachman, L.; Chunara, R.; Srikumar, M.; Weller, A.; and Xiang, A. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Bisconti, P.; Aquilino, L.; Marchetti, A.; and Nardi, D. 2024. A Formal Account of Trustworthiness: Connecting Intrinsic and Perceived Trustworthiness. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 131–140.

Braun, M.; Bleher, H.; and Hummel, P. 2021. A leap of faith: is there a formula for "trustworthy" AI? *Hastings Center Report*, 51(3): 17–22.

Brennen, A. 2020. What Do People Really Want When They Say They Want "Explainable AI?" We Asked 60 Stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7.

Busacca, B.; and Padula, G. 2005. Understanding the relationship between attribute performance and overall satisfaction: Theory, measurement and implications. *Marketing Intelligence and Planning*, 23(6): 543–561.

Cancro, G. J.; Pan, S.; and Foulds, J. 2022. Tell me something that will help me trust you: A survey of trust calibration in human-agent interaction. *arXiv preprint arXiv:2205.02987*.

Choudhary, T. 2025. Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude. *IEEE Access*, 13: 11341–11379.

Ciriello, R. F.; Chen, A. Y.; and Rubinsztein, Z. A. 2025. Compassionate AI Design, Governance, and Use. *IEEE Transactions on Technology and Society*, 1–6.

De Visser, E. J.; Peeters, M. M.; Jung, M. F.; Kohn, S.; Shaw, T. H.; Pak, R.; and Neerincx, M. A. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2): 459–478.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dobbe, R.; Krendl Gilbert, T.; and Mintz, Y. 2021. Hard choices in artificial intelligence. *Artificial Intelligence*, 300: 103555.

Endsley, M. R. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1): 32–64.

Ferrario, A. 2025. A Trustworthiness-based Metaphysics of Artificial Intelligence Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1360–1370. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

Fischer, R.; Liebig, T.; and Morik, K. 2024. Towards more sustainable and trustworthy reporting in machine learning. *Data Mining and Knowledge Discovery*, 38(4): 1909–1928.

Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. Version 1). arXiv.

Hoff, K. A.; and Bashir, M. 2015. Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust. *Human Factors*, 57(3): 407–434.

Holzinger, A.; Carrington, A.; and Müller, H. 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *KI - Künstliche Intelligenz*, 34(2): 193–198.

Huang, Y.; Gao, C.; Wu, S.; Wang, H.; Wang, X.; Zhou, Y.; Wang, Y.; Ye, J.; Shi, J.; Zhang, Q.; Li, Y.; Bao, H.; Liu, Z.; Guan, T.; Chen, D.; Chen, R.; Guo, K.; Zou, A.; Kuen-Yew, B. H.; Xiong, C.; Stengel-Eskin, E.; Zhang, H.; Yin, H.; Zhang, H.; Yao, H.; Yoon, J.; Zhang, J.; Shu, K.; Zhu, K.; Krishna, R.; Swayamdipta, S.; Shi, T.; Shi, W.; Li, X.; Li, Y.; Hao, Y.; Jia, Z.; Li, Z.; Chen, X.; Tu, Z.; Hu, X.; Zhou, T.; Zhao, J.; Sun, L.; Huang, F.; Sasson, O. C.; Sattigeri, P.; Reuel, A.; Lamparth, M.; Zhao, Y.; Dziri, N.; Su, Y.; Sun, H.; Ji, H.; Xiao, C.; Bansal, M.; Chawla, N. V.; Pei, J.; Gao, J.; Backes, M.; Yu, P. S.; Gong, N. Z.; Chen, P.-Y.; Li, B.; Song, D.; and Zhang, X. 2025. On the Trustworthiness of Generative Foundation Models: Guideline, Assessment, and Perspective. Version Number: 4.

Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. 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*, 624–635.

Kaur, D.; Uslu, S.; Rittichier, K.; and Durresi, A. 2023. Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, 55(2): 1–38.

Klingbeil, A.; Grützner, C.; and Schreck, P. 2024. Trust and reliance on AI An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160: 108352.

Koenig, P. D.; and Mehrotra, S. 2025. Emotional Responses to Artificial Intelligence Systems: A Scoping Review of Studied Applications, Key Concepts, and Central Associations. *International Journal of Human–Computer Interaction*, 0(0): 1–21.

Koriat, A. 1997. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4): 349–370.

Koyejo, O.; Natarajan, N.; Ravikumar, P.; and Dhillon, I. 2014. Consistent Binary Classification with Generalized Performance Metrics. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2, 2744–2752.

Krafft, T.; Zweig, K.; and König, P. 2020. How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation and Governance*. Online first).

Laux, J.; Wachter, S.; and Mittelstadt, B. 2024. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1): 3–32.

Lee, J.; and See, K. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1): 50–80.

Lee, M. K.; Psomas, A.; Procaccia, A. D.; Kusbit, D.; Kahng, A.; Kim, J. T.; Yuan, X.; Chan, A.; See, D.; Noothigattu, R.; and Lee, S. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–35.

Lekadir, K.; Frangi, A. F.; Porras, A. R.; Glocker, B.; Cintas, C.; Langlotz, C. P.; Weicken, E.; Asselbergs, F. W.; Prior, F.; Collins, G. S.; Kaissis, G.; Tsakou, G.; Buvat, I.; Kalpathy-Cramer, J.; Mongan, J.; Schnabel, J. A.; Kushibar, K.; Riklund, K.; Marias, K.; Amugongo, L. M.; Fromont, L. A.; Maier-Hein, L.; Cerdá-Alberich, L.; Martí-Bonmatí, L.; Cardoso, M. J.; Bobowicz, M.; Shabani, M.; Tsiknakis, M.; Zuluaga, M. A.; Fritzsche, M.-C.; Camacho, M.; Linguraru, M. G.; Wenzel, M.; De Bruijne, M.; Tolsgaard, M. G.; Goisauf, M.; Cano Abadía, M.; Papanikolaou, N.; Lazrak, N.; Pujol, O.; Osuala, R.; Napel, S.; Colantonio, S.; Joshi, S.; Klein, S.; Aussó, S.; Rogers, W. A.; Salahuddin, Z.; and Starmans, M. P. A. 2025. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388.

Lenat, D.; and Feigenbaum, E. 1991. On the thresholds of knowledge. *Artificial Intelligence*, 47(1–3): 185–250.

Leushuis, E.; Steeg, J.; Steures, P.; Bossuyt, P.; Eijkemans, M.; Veen, F.; Mol, B.; and Hompes, P. 2009. Prediction models in reproductive medicine: A critical appraisal†. *Human Reproduction Upyear*, 15(5): 537–552.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9): 1–46.

Liao, Q.; and Sundar, S. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness Accountability and Transparency*, 1257–1268.

Lucas, G.; Becerik-Gerber, B.; and Roll, S. 2024. Calibrating workers' trust in intelligent automated systems. *Patterns*, 5(9): 101045.

Lucas, G.; Stratou, G.; Lieblich, S.; and Gratch, J. 2016. Trust me: Multimodal signals of trustworthiness. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 5–12.

Markus, A.; Kors, J.; and Rijnbeek, P. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113: 103655.

Mayer, R.; Davis, J.; and Schoorman, F. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3): 709.

McDermott, P.; and Brink, R. 2019. Practical Guidance for Evaluating Calibrated Trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1): 362–366.

Mehrotra, S. 2021. Modelling Trust in Human-AI Interaction. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, 1826–1828. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

Mehrotra, S.; Degachi, C.; Vereschak, O.; Jonker, C. M.; and Tielman, M. L. 2024a. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM J. Responsib. Comput.*, 1(4).

Mehrotra, S.; Gadiraju, U.; Bittner, E.; van Delden, F.; M. Jonker, C.; and L. Tielman, M. 2025. "Even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '25, 51–62. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713132.

Mehrotra, S.; Jorge, C. C.; Jonker, C. M.; and Tielman, M. L. 2024b. Integrity-based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Trans. Interact. Intell. Syst.*, 14(1).

Miller, T. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267: 1–38.

Miró-Nicolau, M.; Jaume-i Capó, A.; and Moyà-Alcover, G. 2025. A comprehensive study on fidelity metrics for XAI. *Information Processing and Management*, 62(1): 103900.

Naiseh, M.; Al-Thani, D.; Jiang, N.; and Ali, R. 2021. Explainable recommendation: when design meets trust calibration. *World Wide Web*, 24(5): 1857–1884.

Okamura, K.; and Yamada, S. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one*, 15(2): e0229132.

Padhy, S.; Takkar, B.; Chawla, R.; and Kumar, A. 2019. Artificial intelligence in diabetic retinopathy: A natural step to the future. *Indian Journal of Ophthalmology*, 67(7): 1004–1009.

Park, P.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5): 1–16.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. Barcelona Spain: ACM. ISBN 978-1-4503-6936-7.

Reveilhac, M.; and Morselli, D. 2025. ChatGPT as a voting application in direct democracy. *Swiss Political Science Review*, spsr.12650.

Russell, S. 2020. *Human compatible: AI and the problem of control*. Penguin Books.

Scharowski, N.; Benk, M.; Kühne, S. J.; Wettstein, L.; and Brühlmann, F. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 248–260. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Schemmer, M.; Kuehl, N.; Benz, C.; Bartos, A.; and Satzger, G. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422. Sydney NSW Australia: ACM. ISBN 979-8-4007-0106-1.

Schlicker, N.; Baum, K.; Uhde, A.; Sterz, S.; Hirsch, M.; and Langer, M. 2025a. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM. *Computers in Human Behavior*, 170: 108671.

Schlicker, N.; Lechner, F.; Wehrle, K.; Greulich, B.; Hirsch, M. C.; and Langer, M. 2025b. Trustworthy enough? Examining trustworthiness assessments of large language model-based medical agents. *Technology, Mind, and Behavior*, 6(2): 1–29.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

Shailya, K.; Rajpal, S.; Krishnan, G.; and Ravindran, B. 2025. LExT: Towards Evaluating Trustworthiness of Natural Language Explanations. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1565–1587.

Steyvers, M.; and Kumar, A. 2024. Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, 19(5): 722–734. PMID: 37439761.

Tao, Y.; Viberg, O.; Baker, R.; and Kizilcec, R. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9): 346.

Terblanche, N.; Molyn, J.; Williams, K.; and Maritz, J. 2023. Performance matters: students' perceptions of Artificial Intelligence Coach adoption factors. *Coaching: An International Journal of Theory, Research and Practice*, 16(1): 100–114.

Thiebes, S.; Lins, S.; and Sunyaev, A. 2021. Trustworthy artificial intelligence. *Electronic Markets*, 31(2): 447–464.

Tomsett, R.; Preece, A.; Braines, D.; Cerutti, F.; Chakraborty, S.; Srivastava, M.; Pearson, G.; and Kaplan, L. 2020. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, 1(4): 100049.

Vereschak, O.; Bailly, G.; and Caramiaux, B. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Wan, Y.; Wang, W.; He, P.; Gu, J.; Bai, H.; and Lyu, M. R. 2023. BiasAsker: Measuring the Bias in Conversational AI System. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2023, 515–527. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703270.

Wang, Z. B.; Lin, P.; Cheng, K. Z.; and Vaughan, J. W. 2021. Explanations can be Manipulative: Learning to Better Trust Human Decisions with Machine Explanations. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Wischnewski, M.; Krämer, N.; Janiesch, C.; Müller, E.; Schnitzler, T.; and Newen, C. 2024. In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems. *Human-Machine Communication*, 8: 141–162.

Wischnewski, M.; Krämer, N.; and Müller, E. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.