Int. J. Human–Computer Studies 219 (2027) 103971



Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs



How communicating an AI's relational risk influences trust perceptions in AI-assisted decision-making

Sarah Schömbs*, Saumya Pareek, Jorge Goncalves, Wafa Johal

School of Computing and Information Systems, The University of Melbourne, Grattan Street, Parkville, Melbourne, 3010, VIC, Australia

HIGHLIGHTS

- Communicating an AI's relational risk enhances perceived transparency after errors across stakes.
- Relational risk disclosure fosters output scrutiny without reducing perceived AI utility.
- AI accuracy communication inappropriately increases trust perceptions across all dimensions after high-stakes errors.
- AI apology improves perceived utility but not system understanding.
- Decision stakes moderate how strategies influence AI trustworthiness perceptions.

ARTICLE INFO

Keywords:
Risk communication
AI trustworthiness
Trust
Transparency
AI-assisted decision-making

ABSTRACT

AI systems err; however, existing *Strategies* to influence a user's trust perception can sometimes fail to foster perceptions and assessments that reflect system limitations, with potentially detrimental real-world consequences. In this paper, we introduce a novel Relational Risk (RR) strategy derived from risk communication research to pre-emptively communicate the likelihood of AI error and its potential consequences, foregrounding the costs of AI error, and compare it with existing *Strategies* (AI accuracy, AI apology). Participants ($N = 208$) engaged in an AI-assisted decision-making task with varying *Stakes* and AI performance, and we measured how these influenced perceived AI trustworthiness across relevant perceptual dimensions (transparency, performance, utility). We find that RR enhances perceived AI transparency post-AI errors and encourages healthy scrutiny without diminishing or inflating perceived AI performance or utility. Conversely, AI accuracy boosts perceived AI transparency, performance, and utility under high decision-stakes, increasing system trustworthiness when unwarranted, while apology improves perceived AI utility but does not support system understanding. We contribute an in-depth investigation of how communication strategies shape AI trustworthiness perceptions, reveal important drawbacks of existing approaches, and demonstrate how integrating risk communication theory can enable more appropriate trustworthiness assessments.

1. Introduction

The recently published AI Incidents Database (McGregor, 2021; Slattery et al., 2025) has recorded thousands of real-world AI failures, most of which were encountered and reported by end-users during interactions with deployed systems (De Miguel Velázquez et al., 2024). A discrepancy between perceived and actual AI trustworthiness in AI systems can lead to suboptimal outcomes with potentially harmful consequences, depending on the AI's deployment context. Research across human–automation interaction, machine learning, and HCI is therefore investigating how best to support users in their assessment of system trustworthiness and their understanding of system characteristics, including capabilities and limitations (Schlicker et al., 2025a; Wang and Yin, 2021; Zhang et al., 2020; Cao et al., 2024) as a necessary premise for deploying AI decision support that is accessible to users across domains. Imperatively, this necessitates fostering scepticism and critical thinking among users, to recalibrate perceptions of AI trustworthiness to match different levels of system performance (Shen et al., 2024) and thus to decrease the gap between perceived and actual AI trustworthiness (Schlicker et al., 2025a,b). It also requires nudging users to assess AI out-

* Corresponding author.
Email address: s.schombs@unimelb.edu.au (S. Schömbs).

https://doi.org/10.1016/j.ijhcs.2026.103971
Received 17 February 2026; Received in revised form 20 October 2026; Accepted 7 November 2026
Available online 10 November 2026
1071-5819/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

puts more deliberately rather than applying broad heuristics (Buçinca et al., 2021; Park et al., 2019), especially in contexts where consequences are severe.

As AI-assisted decision-making becomes more widespread and accessible to users without expertise in AI, through applications such as AI-based dermatology assistance tools (Liu et al., 2020), the central challenge is to identify *how* to foster critical thinking and *how* to design intervention cues (Schlicker et al., 2025a,b) that communicate capabilities and limitations effectively in-situ, before negative consequences can manifest as a result of misplaced trust in AI systems.

Communicating AI accuracy as an a priori performance cue has been investigated as a potentially promising strategy in HCI to prime user expectations (Kocielnik et al., 2019), and to shape trust in AI assisted decision-making (e.g., (Yin et al., 2019; Rechkemmer and Yin, 2022; Papenmeier et al., 2019; He et al., 2023)). However, empirical findings on whether communicating AI accuracy translates into *appropriate* trust remain inconsistent. Prior work that focused predominantly on behavioural trust suggests that users can be insensitive to the accuracy value itself (Lai and Tan, 2019; Yin et al., 2019), which can have severe consequences in high-stakes applications. Yin et al. (2019) found that users' behavioural trust did not differ significantly across several adjacent stated accuracy levels, despite the magnitude of the numeric difference, which reflects concerns related to people's difficulties in interpreting numerical values (He et al., 2023). Similarly, Lai and Tan (2019) found that a stated accuracy of 50%, equivalent to chance performance, still significantly increased users' behavioural trust relative to providing no accuracy statement at all, which raises concerns that the presence of accuracy communication can inflate trust. While these findings shed some light on behavioural trust, an in-depth understanding of system trustworthiness perceptions remains scarce, particularly regarding how such strategies shape these perceptions in response to both correct and erroneous AI behaviour, despite its importance recently re-emphasised by Schlicker et al. (2025a).

In everyday life, *risk communication* is used to elicit reflective and contemplative thoughts, and to support people's understanding of probabilities and consequences associated with risks in decision-making (Schömbs et al., 2025; Visschers et al., 2009, 2012; Brown and Smith, 2007). Arguably, AI accuracy conveys the risk of relying on the AI implicitly. As a global performance cue, it does not communicate uncertainty or confidence at the level of a specific decision, but it conveys the likelihood dimension of risk by stating the probability that the AI will be correct, which implies its potential for error. However, from a risk-informed perspective, stated AI accuracy a) does not address the *consequence component* crucial to effective risk communication (Wilson et al., 2019; Dunlop et al., 2008) and b) deploys a *positive framing* of the AI's performance, a framing strategy leveraged in risk communication research to reduce risk awareness and promote risk-seeking decisions (McNeil et al., 1982; Peters et al., 2011). In response to the mixed results of AI accuracy, recent work has begun to examine alternative and complementary approaches, for example, explicitly disclosing AI weaknesses (Rieger et al., 2025), communicating system "incapabilities" in human-robot interaction (HRI) (Kwon et al., 2018; Schömbs et al., 2024), and explaining AI accuracy through everyday analogies (He et al., 2023), an approach inspired by risk communication research, with promising but unsatisfying results with respect to trust. Importantly, none of these approaches makes the *relational risks* of AI systems explicit to users, that is, the risks that stem from the AI system itself (Stuck et al., 2022). The communication of relational risk could in turn influence users' *perceived relational risk*, defined as a user's belief about the probability of negative consequences associated with the use of a specific AI system, based on, for example, the user's experience with that system (Stuck et al., 2022).

At the same time, growing work in HCI has begun to examine AI risks as a key topic in human–AI interaction; however, research has approached it predominantly from the perspective of AI practitioners. Examples include frameworks for documenting and reporting AI risks, e.g., RiskCards for language models (Derczynski et al., 2023), model-specific risk reports (Rao et al., 2025), and tools designed to support AI practitioners in ideating and anticipating risks during the development stage (Herdel et al., 2024; Wang et al., 2024; Buçinca et al., 2023). While these perspectives are essential for advancing human-AI interaction and stress the importance of considering risk in AI-assisted decision-making, they re-emphasise that research investigating *the communication* of *relational risks to users* remains largely unaddressed. Given the empirically evident and foreseeable risks associated with the deployment and use of AI-assisted decision-making (McGregor, 2021; Slattery et al., 2025) and the limited effectiveness of current approaches to effectively influence trust perceptions in a way that aligns with the AI's capabilities and limitations and to support users in their trustworthiness assessment, it is crucial to examine explicit risk communication as an alternative strategy for navigating trust perceptions in human–AI interaction. Such examination further serves to investigate how risk communication theory can be integrated into HCI to leverage natural synergies between the disciplines.

To address this research gap, we examine communicating Relational Risk (RR) in AI-assisted decision-making, which presents the likelihood of the AI erring and its consequences to foreground the cost of error. In detail, our RR design was informed by risk research conceptualising risk as a combination of the likelihood of a negative event and its potential consequence (Wilson et al., 2019; Dunlop et al., 2008). Whereas accuracy communication can be understood as an implicit operationalisation of this likelihood dimension, RR extends this by explicitly addressing both likelihood and consequence. We further ground our design in findings from risk communication research, which demonstrate that the way information is framed, particularly the valence of the frame, can systematically influence decision-making (Spiegelhalter, 2017; Gong et al., 2013). Accordingly, we adopt a negative framing for RR to potentially increase risk awareness, elicit critical attentiveness, and foster more risk-averse perceptions (McNeil et al., 1982; Peters et al., 2011).

Our work is guided by the following overarching research questions:

RQ1: How does communicating the Relational Risk (RR) of an AI influence users' perceived AI trustworthiness across varying AI performance when compared to established *Strategies*, i.e., AI accuracy and apology?

RQ2: How does communicating the Relational Risk (RR) of an AI influence users' expectations and understanding of the AI?

RQ3: How do the *Stakes* (low, high) inherent in the decision-making context affect users' perceptions of AI trustworthiness across varying AI performance (**RQ3.1**), and how does it interact with the investigated *Strategies* in influencing these perceptions (**RQ3.2**)?

We investigate the communication of our proposed Relational Risk (RR) approach as a proactive user-facing communication strategy in comparison to two established approaches in AI-assisted decision-making: stated AI accuracy (Acc), a similarly proactive a priori indicator of AI performance to influence trust and set appropriate expectations (Cao et al., 2024; Yin et al., 2019; Rechkemmer and Yin, 2022; Papenmeier et al., 2019), and AI apology (Apol), a reactive communication strategy to navigate trust dynamics particularly *post* erroneous AI behaviour (Kim and Song, 2021; Tolmeijer et al., 2020; Pareek et al., 2024). We assess how these *Strategies*, in interaction with *Stakes*, affect the trajectory of perceived AI trustworthiness and system understanding as users encounter both correct and erroneous AI behaviour. We thereby focus on an in-depth investigation into trust perceptions and their underlying constructs (perceived AI transparency, performance, and utility) as a critical foundational step (Schlicker et al., 2025a; Mayer et al., 1995; Roesler et al., 2022). This allows us to gain a fine-grained understanding of whether the investigated trustworthiness cues, RR in particular as a novel strategy, diminish system trustworthiness perceptions despite correct AI behaviour, for instance by undermining perceived utility, or render system trustworthiness perceptions unrecoverable after erroneous

AI behaviour. This offers empirical insight into cue utilisation, a core condition of the recently proposed Trustworthiness Assessment Model (TrAM) (Schlicker et al., 2025a,b), by examining whether and how RR and existing strategies (Acc, Apol) are utilised to form trustworthiness perceptions across varying levels of AI performance.

We present a 2 (*Stakes:* low vs. high) × 4 (*Communication Strategy:* RR, Acc, Apol, Baseline) between-subjects user study with $N = 208$ participants. The study was contextualised as an AI-assisted diagnostic task in which participants classified microscope images as infested or not (Fahnenstich et al., 2024). *Stakes* was operationalised through the decision-making context, with participants inspecting microscope images of either plant (low) or human tissue (high) for possible bacterial infestation. Participants interacted with the AI across 3 blocks with varying AI performance to gain an in-depth understanding of participants' perceptions of system trustworthiness in response to correct and erroneous AI identifications. We measured three core facets of AI trustworthiness after each block: perceived AI transparency, performance, and utility.

Our results show that communicating RR enhances perceptions of AI transparency after erroneous AI behaviour across low (plant tissue) and high (human tissue) stakes relative to the Baseline, while prompting users to anticipate failures, temper their expectations, and scrutinise AI outputs more carefully. We discuss how these findings position RR communication as a potential cognitive forcing function that may sustain caution and foster deliberate assessment without diminishing perceived AI performance or utility, while also raising questions about potential limitations in low-stakes contexts that warrant future investigation. Second, we find that Acc communication, although it made users aware of the AI's potential to make mistakes, also backfired: stated AI accuracy made users believe that the AI was highly accurate, elevating AI performance perceptions compared to the Baseline and RR in high-stakes contexts where misplaced perceptions of AI trustworthiness could be particularly problematic. Third, we find that Apol increases the perceived usefulness of the AI compared to the Baseline but fails to support system understanding, which demonstrates an important constraint of apology-based interventions.

Our work makes the following contributions:

- We propose RR as an important alternative approach to influencing trust perceptions in AI-assisted decision-making, informed by risk communication theory.
- We provide empirical evidence on the effects of RR compared to established *Strategies* across perceptions of AI transparency, performance, and utility, revealing the respective strengths and drawbacks of different *Strategies*. By examining these effects across correct and erroneous AI behaviour, we further provide rich empirical insight into *cue utilisation*, a factor central to the recently developed trustworthiness assessment model (TrAM) (Schlicker et al., 2025a), showing *how* RR and existing strategies are utilised to form trustworthiness perceptions.
- We demonstrate empirically that *Stakes* moderates the effectiveness and appropriateness of *Strategies* for perceived AI trustworthiness. Our findings imply that *Stakes* must be considered when designing and evaluating *Strategies* to shape trust perceptions in AI-assisted decision-making. We provide empirical support for recent work conducted by Schlicker et al. (2025b), which proposes that a risk–benefit evaluation occurs at the level of trustworthiness perceptions, rather than being confined to the behavioural stage as earlier trust models assumed (Mayer et al., 1995; Stuck et al., 2022).
- We offer implications for the design of *Strategies* to shape trust perceptions in AI-assisted decision-making that remain sensitive to the *Stakes* inherent in different decision-making contexts, which contributes to a critical topic in human–AI interaction.

Our work contributes to integrating risk communication theory in HCI, and shifts the conversation towards explicit risk communication as an alternative approach for navigating trust dynamics in AI-assisted decision-making. We position Relational Risk (RR) as an important strategy in human-AI interaction that warrants further investigation and argue that trust perceptions can be appropriately influenced via prior risk communication, without deterring AI use.

2. Related work

We review prior work on approaches to influencing trust perceptions in AI-assisted decision-making alongside risk communication theory to ground the manipulation examined in this study. This motivates the design of Relational Risk (*RR*) communication as an approach that builds on and extends existing approaches by integrating insights from risk communication theory. By critically comparing established approaches to influencing trust perceptions with principles from risk communication theory, we identify their limitations and clarify the rationale behind our study design.

2.1. Trust in human-AI interaction

In this work, we focus on users' perceptions of AI trustworthiness as a form of situational trust, a multi-dimensional construct in human–AI interaction (Hoff and Bashir, 2015; Mayer et al., 1995; Lee et al., 2022; Kohn et al., 2021; Roesler et al., 2022). Early works by Lee and See (2004) describe three determinants of perceived AI trustworthiness: performance, which reflects a user's assumptions about how well an AI performs, process, which denotes the extent to which a user comprehends how the AI operates, and purpose, which captures whether the AI is aligned with the user's objectives (perceived benefit). Recent work by Roesler et al. (2022) extends this framework in the context of AI-assisted decision-making. The authors conceptualise and validate perceived AI trustworthiness as a multi-dimensional construct defined by performance and utility, which load on the performance pillar of Lee and See (2004), purpose, which corresponds to the purpose pillar, and transparency, which maps onto the process pillar and constitutes an important extension in AI-assisted decision-making, as reflected by recent research echoing the need for more transparency in human-AI interaction (Bhatt et al., 2021; Prabhudesai et al., 2023).

Unless stated otherwise, throughout this paper we use the term trust to denote perceptions of AI trustworthiness.

2.1.1. Strategies to influence trust perceptions in AI-assisted decision-making

The alignment of a user's perceived AI trustworthiness with system capabilities and limitations, i.e., calibrated trust (Lee and See, 2004; Hoff and Bashir, 2015), is a key challenge in HCI given the increasingly ubiquitous deployment of AI systems in everyday settings and the potentially detrimental consequences that depend on the deployment context (Wang and Yin, 2022). Calibrated trust has therefore become a crucial "desideratum" of explanations in human-AI interaction (Wang and Yin, 2022) to foster a user's "appropriate use" of deployed AI systems (Lee and See, 2004). Calibrated trust, in detail, is defined as a user's trust that matches the system's capabilities, albeit avoiding both overtrust, which describes trust that "exceeds the system's capabilities", and distrust, which describes trust that "falls short of the system's capabilities" (Lee and See, 2004).

The recently developed trustworthiness assessment model (TrAM) (Schlicker et al., 2025a) emphasises the theoretical gap and the importance of investigating the discrepancy between *actual* and *perceived* AI trustworthiness, as a system's trustworthiness must be inferred through observable cues. To narrow this gulf between perceived and actual system trustworthiness, users require "observable cues" that are relevant and available from the system's side, and detectable and appropriately utilised from the user's side (Schlicker et al., 2025a,b). To support people's mental models of AI systems and to inform appropriate assessments of AI trustworthiness, different approaches have been investigated by the HCI community, spanning proactive cues introduced *prior to*

the interaction, instance-based approaches communicated *during* the interaction, and reactive strategies deployed *after* the AI has erred.

Global, a priori cues include the communication of AI accuracy as a performance indicator (He et al., 2023; Yin et al., 2019; Rechkemmer and Yin, 2022); local, in-the-moment explanations are communicated during interaction, often in conjunction with AI decision outcomes, such as AI confidence (Rechkemmer and Yin, 2022; Cao et al., 2024; Schömbs et al., 2024), counterfactuals (Wang and Yin, 2022), or locus of causality (Pareek et al., 2025; Kim and Song, 2023, 2021); and reactive strategies are deployed after the AI has erred, with apology as an established approach in both AI-assisted decision-making (Pareek et al., 2024; Kim and Song, 2021; Tolmeijer et al., 2020) and human-robot interaction (HRI) (Kox et al., 2021; Nesset et al., 2023), where its conceptual purpose is to prevent disuse or AI aversion following an error. Beyond apology, Pareek et al. (2024) adapted a broader set of trust repair strategies from HRI, among them denial and promise, alongside a novel model update strategy, for AI-assisted decision-making. Their findings position model update and apology as effective strategies to restore trustworthiness perceptions post AI error, while denial was largely counterproductive and instilled distrust in the system rather than repairing it. That said, model update presupposes that the underlying model has in fact been retrained or improved, which is rarely feasible within the short, single-session interactions typical of AI-assisted decision-making tasks. Interestingly, in the context of HRI, LeMasurier et al. (2024) observed that a robot system communicating failures proactively was perceived as more intelligent, understandable, and trustworthy than a reactive system that explained failures after they occurred. However, it remains largely unclear how these strategies are utilised by users in AI-assisted decision-making to form trustworthiness perceptions, particularly in response to correct and erroneous AI behaviour.

2.1.2. AI accuracy and priming expectations of AI performance

Among these approaches to shape trust perceptions of AI systems, AI accuracy has received particular attention as a simple a priori approach to guide expectations and trust development (Rechkemmer and Yin, 2022; Papenmeier et al., 2019; He et al., 2023), distinct from instance based AI confidence communication (Cao et al., 2024). Research shows that AI accuracy influences perceived trustworthiness more strongly than either AI confidence (Rechkemmer and Yin, 2022) or explanation fidelity, i.e., the degree to which explanations truthfully represent how the model works (Papenmeier et al., 2019). Interestingly, lay users seem to perceive accuracy as a factual property of the system, while interpreting confidence as an uncertain, volatile estimate (Rechkemmer and Yin, 2022). Conversely, He et al. (2023) compared AI accuracy, AI accuracy accompanied by an analogy explaining the accuracy value, and a no-information baseline, and found no significant differences in users' trustworthiness perceptions across these conditions. Their study, however, did not expose participants to correct or erroneous AI judgements, which raises the question of how accuracy communication shapes trustworthiness perceptions in response to correct AI judgements or AI failures.

At the same time, findings from trust research focusing on reliance as behavioural trust indicate that accuracy information needs to be treated with caution, as users' agreement with an AI does not sensitively reflect different levels of stated accuracy. Lai and Tan (2019) show that presented model accuracy information increases the degree to which users follow the model's judgement even when the communicated accuracy is solely 50%, relative to no accuracy statement at all. A surprising result, given that users should place considerably less trust in a system that communicates such low accuracy. While this may be unproblematic in contexts with little consequence, such misplaced perceptions could be harmful in deployment settings with high stakes (e.g., medical diagnostics). Similarly, Yin et al. (2019) report that participants' behavioural trust did not significantly differ between conditions with no accuracy information, and 60%, or 70% stated AI accuracy. Significant differences only emerged when contrasting this lower-accuracy batch (none/60%/70%) with the high-accuracy batch (90%/95%), with the latter eliciting greater behavioural trust in the AI's prediction.

In response to these mixed findings on accuracy communication, Rieger et al. (2025) shifted attention to explicitly exposing AI weaknesses as a global, a priori explanation. In two experiments, AI weaknesses were described as limitations in handling certain types of cases due to reduced training data, which has also been shown to influence how users attribute the locus of causality for AI failures (Pareek et al., 2025). AI weakness disclosures increased perceived transparency in the first experiment, but not in the second. The second experiment instead revealed a decline in initial trust relative to the group that received no information about the AI, which did not persist over the course of interaction, indicating no sustained erosion of trust perceptions through AI weakness disclosure. This result draws attention to an important consideration: AI interventions should serve to shape perceived AI trustworthiness appropriately, while not undermining the system's perceived utility, or inducing algorithmic aversion (Passi and Vorvoreanu, 2022), or discouraging users from benefiting from its support when appropriate, a concern that mirrors how apology strategies aim to prevent disuse following AI failure (Pareek et al., 2024).

In this paper, our objective is to examine how `RR` compared to established *Strategies* (`Acc`, `Apol`) affects perceptions of AI transparency, performance, and utility (RQ1) as crucial dimensions of system trustworthiness perceptions (Roesler et al., 2022). While `RR` may serve as an a priori approach to help users anticipate AI fallibility and thereby potentially support transparency, it is important to explore whether this comes at the expense of perceived utility and performance.

2.2. Risk communication and risks in human-AI interaction

Recent work specifies three types of risk that warrant consideration when assessing trust dynamics in human-AI interaction: relational risk, situational risk, and risk propensity (Stuck et al., 2022). Our work focuses on the first two constructs and does not address risk propensity in detail, which refers to an individual's general inclination towards risk aversion or risk-taking as an individual tendency.

2.2.1. Relational risk

Users navigate a relational form of risk associated with the AI system itself (Stuck et al., 2022). Perceived *relational risk* describes the risk related to the use of the technology itself, in our case an AI agent. It refers to the user's perception that using the AI agent may lead to positive or negative outcomes (Stuck et al., 2022). Importantly, risk itself is typically defined as a combination of the likelihood of a hazard and its potential consequence, which demonstrates that risk communication fundamentally involves conveying two determinants, i.e., probability *and* consequence (Wilson et al., 2019; Dunlop et al., 2008).

Human-AI interaction research has begun to incorporate knowledge derived from risk research to improve people's understanding of AI uncertainty. Presenting uncertainty in a frequency format has been shown to support more appropriate reliance and reduce confirmation bias (Cao et al., 2024; Zhang et al., 2020), which aligns with risk communication research showing that frequency formats increase people's risk perceptions (Peters et al., 2011; Spiegelhalter, 2017; Peters et al., 2011). Motivated by the risk communication literature, He et al. (2023) investigated the use of analogies from everyday contexts (e.g., weather forecasts) as an additional cue to explain stated accuracy values and to support numeracy. While the analogies were sometimes perceived as helpful for understanding AI accuracy, they had no significant effect on perceived AI trustworthiness and were also criticised as unfamiliar or unspecific for the task context. Notably, from a risk communication lens, these approaches are characterised by their consequence-agnosticism as they leave the consequence component crucial to risk communication theory unaddressed.

Beyond the probability and consequence dimensions of risk, *how* this information is communicated, specifically its framing, further shapes

risk perception and downstream decision-making. To guide people's risk perception and to elicit appropriate decision-making, traditional risk theory emphasises the importance of *frame valence* in risk communication, that is, whether risks are presented in a positive or negative framing (Spiegelhalter, 2017). For example, describing side-effect probabilities as the proportion of patients who remain unaffected reduces the perceived risk of a medication compared to emphasising the proportion who experience side effects (Peters et al., 2011). In treatment decisions, positive framing tends to increase acceptance of high-risk interventions such as surgery, while negative framing steers people's preferences towards non-invasive alternatives, reflecting more cautious, risk-averse decisions (McNeil et al., 1982; Armstrong et al., 2002; Gong et al., 2013). Exposing AI weaknesses (Rieger et al., 2025; Kwon et al., 2018; Schömbs et al., 2024) can be viewed as a loose form of negatively framed system capabilities, and our work extends this idea by designing a strategy informed through the lens of risk communication research. Informed by prior work, we design `RR` to explicitly communicate both the AI's error rate, conveyed in a frequency based format, and its potential consequences. Specifically, `RR` extends established approaches aimed at shaping system trustworthiness perceptions by adopting a negative frame valence and by explicitly incorporating the consequence dimension to foreground the costs of potential AI fallibility, with the aim of fostering more cautious perceptions of trustworthiness. Beyond examining how `RR` affects perceptions of AI transparency, performance, and utility (RQ1), we also investigate how it shapes users' expectations and understanding of the AI system (RQ2).

2.2.2. Situational risk — decision stakes

Situational risk is another critical type of risk in AI-assisted decision-making. It defines risk as a contextual factor, regardless of the use of a specific technology (Stuck et al., 2022). Situational risk is considered crucial for human–AI trust, as misplaced trust in AI can have real-world consequences with varying degrees of severity depending on the application context. Indeed, as Kohn et al. (2021) argue, "trust in a situation that has no potential for negative consequences is effectively meaningless". Nevertheless, empirical investigations that assess situational risk in human-AI research remain scarce (Hoesterey and Onnasch, 2023). Contextualised within AI assisted decision-making, prior research predominantly focused on either side of the coin - low situational risk (e.g., AI supporting the identification of animals or shapes (Pareek et al., 2024) or answering factual tasks (Kim et al., 2024)) or high situational risk (e.g., skin cancer detection (Cao et al., 2024), loan prediction (He et al., 2023) or the assessment of jail time in criminal law cases (Kahr et al., 2024)).

Some research compared the effect of situational risk on trust during AI assisted decision-making, with inconsistent findings showing no effect of situational risk on AI trustworthiness (Hoesterey and Onnasch, 2023; Fahnenstich et al., 2024; Yin et al., 2019), high situational risk significantly decreasing AI trustworthiness (Pareek et al., 2025; Ashoori and Weisz, 2019), and participants exhibiting greater convergence with AI support than with human support under high situational risk (Fahnenstich et al., 2024). Findings by Hoesterey and Onnasch (2023) show that participants' decision time significantly increased under high situational risk. In line, verification behaviour was elevated in high-risk contexts, more so when users were supported by decision automation (a system-generated recommendation) as opposed to information automation (system-highlighted cues). However, this heightened scrutiny gradually declined as interaction with the decision support system progressed, mirrored by a decline in risk perception, which the authors attributed to "habituation" to the situational risk.

In the quest to understand the effectiveness and appropriateness of *Strategies* relative to their application context (e.g., health), we first singularly study the effects of *Stakes* on AI trustworthiness across its facets (RQ3.1). By *Stakes*, we refer to situational risk, yet we use the term *Stakes* to more clearly distinguish it within our study. We then investigate how the *Stakes* inherent to the decision context interacts with *Strategy* in influencing participants' perceptions of AI trustworthiness under conditions of correct and incorrect AI performance (RQ3.2). We carefully operationalise the *Stakes* of the AI application through the context of the diagnostic task (human tissue vs. plant tissue samples) to provide a highly comparable way to study its effects.

3. Method

We conducted a mixed-methods user study with a 2 (*Stakes:* low, high) × 4 (*Strategy*: `Acc`, `RR`, `Apol`, Baseline) between-subjects factorial design to examine the impact of *Strategy* on perceptions of AI trustworthiness (RQ1) and system understanding (RQ2), and to investigate the role of *Stakes* in shaping trust perceptions during AI-assisted decision-making (RQ3).

3.1. Task

We required an AI-assisted decision-making task that allowed us to manipulate *Stakes* while remaining comparable across conditions (low vs. high). To do so, we contextualised our experiment as a diagnostic task that required participants to screen and classify different microscope samples. During the task, the microscope sample images showed varying bacterial density, see Fig. 1. We asked participants to perform a visual estimation for each microscope image, i.e., to estimate the percentage of bacteria-covered area and classify the sample as either infested or not, given a bacterial infestation threshold of 35% coverage. Samples showing more than 35% infestation were to be classified as 'infested'. A visual estimation task with a predefined cut-off boundary is a task type frequently used to assess trust dynamics in human-AI interaction, owing to their real-world applicability (Rieger et al., 2024; Fahnenstich et al., 2024). To accomplish the task, participants received decision support from an AI agent that presented its judgement about whether a sample was infested or not. We included a training block to support task familiarisation and comprehension.

3.2. Experimental conditions

3.2.1. Operationalising stakes

We derived a set of microscope images from the OSF[1] repository by Fahnenstich et al. (2024), which are microscope images of bacteria originating from the University Medical Centre Rotterdam. While Fahnenstich et al. (2024) showed no influence of stakes on a unidimensional trust measure, they argue that it does not undermine the suitability of the task stimuli as a manipulation of stakes, given their successful validation of the stimuli themselves. We contacted the authors to draw on their stimuli and scenario framing for our manipulation.

We selected a subset of nine microscope images that served as the task stimuli. *Stakes* was operationalised by varying the decision-making context, with participants either classifying microscope images that represented "human tissue" or "plant tissue" samples. Each stimulus was accompanied by a corresponding identifier (plant type/patient name; date), exemplified in Fig. 1. To contextualise *Stakes* further, participants received a short 'brief' at the beginning of the task, accompanied by the respective institution's logo to simulate a realistic decision-making context and thereby increase ecological validity. The brief framed the source of the microscope images and further established the *Stakes* associated with the classification decisions. In the low *Stakes* condition, participants were told to assess samples from a plant nursery called "Verdara Plant Centre", with samples taken from diverse plants, some of which exhibited signs of damage (e.g., root decay). In the high *Stakes* condition, the same task was framed within a medical context. The accompanying brief explained that the batch consisted of samples from different patients at the "Novaira Medical Centre", some of whom were experiencing serious symptoms (e.g., respiratory issues).

[1] https://osf.io/uw239/.

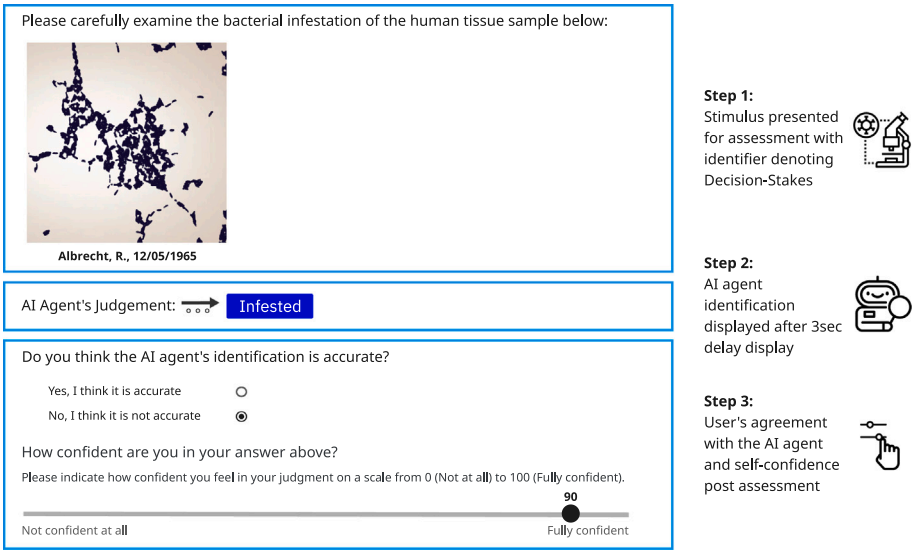


Fig. 1. Example diagnostic trial from the high-stakes condition (human tissue sample) showing an incorrect AI classification as 'Infested' during Block 2, despite bacterial coverage being < 35%.

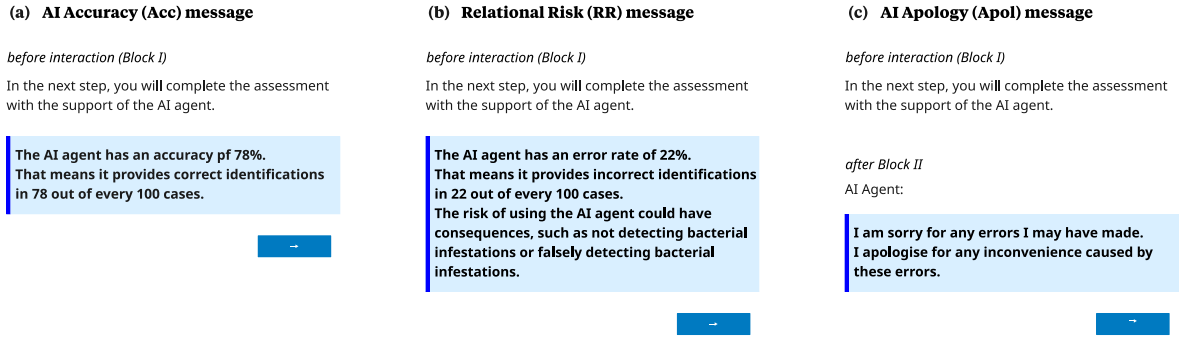


Fig. 2. *Strategy* manipulations are shown here as presented to participants in the online study. AI accuracy (a) and Relational risk (b) messages were presented before the interaction (prior to Block 1), reflecting their proactive design, while the AI apology message (c) was presented after Block 2, following the failure injection, reflecting its reactive design.

3.2.2. *Operationalising strategies to shape trustworthiness perceptions*

We carefully refined and implemented three *Strategy* conditions (see Section 2 for rationale), each of which provided participants with different "observable cues" during the AI-assisted decision-making task to provide information that influences perceptions of system trustworthiness (Schlicker et al., 2025a), shown in Fig. 2.

`Acc`: This condition communicates the likelihood of correctness as AI accuracy in a percentage format as a well established proactive *Strategy presented prior to interaction* following previous work (e.g., (Rechkemmer and Yin, 2022; Yin et al., 2019)), accompanied by a frequency based representation to support numeracy (Cao et al., 2024; Zhang et al., 2020), while adopting a positive frame valence by communicating the accuracy rate (i.e., number of correct identifications). *"The AI agent has an accuracy of 78%. That means it provides correct identifications in 78 out of every 100 cases."*

`RR`: This condition, similarly proactive and presented prior to the interaction, communicates the Relational Risk of using the AI agent itself by a) adopting a negative frame valence by communicating the likelihood of AI error, in line with risk theory on negative framing (Peters et al., 2011; McNeil et al., 1982; Armstrong et al., 2002; Spiegelhalter, 2017) (communicating the error rate; number of incorrect identifications), accompanied by a frequency based format to support numeracy (Cao et al., 2024; Zhang et al., 2020), and by b) explicitly including the consequence dimension of risk following risk theory (Wilson et al., 2019; Dunlop et al., 2008). *"The AI agent has an error rate of 22%. That means it provides incorrect identifications in 22 out of every 100 cases. The risk of using the AI agent could have consequences, such as not detecting bacterial infestations or falsely detecting bacterial infestations."*

`Apol`: In contrast to Acc and RR, this condition constitutes a reactive *Strategy*, presented solely after AI failure rather than prior to interaction, in line with its conceptual purpose. This condition communicates an apology from the AI as an established *Strategy* in AI-assisted decision-making (Kim and Song, 2021; Pareek et al., 2024) employed to navigate trust perceptions post AI failures (recovery) and over repeated interactions. A desideratum of an effective *Strategy* should be to sustain perceived usefulness while still inducing appropriate caution, which motivates the inclusion of this approach. We draw on the framing from Pareek et al. (2024). *"I am sorry for any errors I may have made. I apologise for any inconvenience caused by these errors."*

We further included a Baseline condition in which participants completed the study without any communicative strategy.

3.3. Repeated interaction with varying AI performance

In order to operationalise varying AI performance, we divided the main experiment into three blocks, with each block involving three microscope assessment tasks, which allowed us to examine how *Strategy* (`RR`, `Acc`, `Apol`) and *Stakes* (low, high) influence system trustworthiness perceptions across correct and erroneous AI judgements. In Block 1, the AI agent provided a correct identification across all three cases. In Block 2, we introduced two erroneous identifications in a controlled sequence, with a false negative in the first case and a false positive in the third, based on prior findings showing order effects of AI failures on trust (Kahr et al., 2024). In Block 3, the AI agent again produced correct decisions across all three cases. We adopted this tripartite structure as an established protocol in human–AI interaction to examine users' trust formation, erosion and recovery, respectively, over repeated interactions (Kahr et al., 2024; Pareek et al., 2024; Kim and Song, 2023; Tolmeijer et al., 2021).

For all three blocks, we deliberately selected stimuli that clearly deviated from the bacterial infestation cut-off for participants to be able to determine the AI agent's performance, which allowed us to inject detectable AI failures, consistent with our objective of examining trustworthiness perceptions in response to (in)correct AI judgements. Our design follows prior work that operationalised clearly identifiable errors to investigate failure experiences and trustworthiness in HCI (Kahr et al., 2024; Tabatabaei et al., 2026; Rieger et al., 2025). We simulated the AI and fixed the accuracy at 78% to ensure control over the frequency, position, and type of error across Block 2, consistent with prior studies in AI-assisted decision-making (Buçinca et al., 2021; He et al., 2023).

3.4. Measures

3.4.1. Trust in automation - propensity to trust - TiA-PtT

Trust in human–automation interaction can be categorised as dispositional and situational (Hoff and Bashir, 2015). Dispositional trust describes a user's propensity to trust automation determined by factors such as age, gender, or personality (Mayer et al., 1995; Hoff and Bashir, 2015). We measure participants' dispositional trust in automation as a covariate to account for individual differences, using the TiA-PtT questionnaire (Trust in Automation – Propensity to Trust subscale) (Körber et al., 2019). This measure has been shown to reliably capture dispositional trust and to be an important determinant of perceived AI trustworthiness in AI-assisted decision-making (Pareek et al., 2025).

3.4.2. Perceived AI trustworthiness - transparency, performance, utility

A user's situational trust in AI can be dynamic and shift depending on the specific interaction and type of AI system (Kohn et al., 2021) as it is influenced by variable contextual characteristics such as the type of AI system, its complexity, as well as the perceived risks associated with system use (Hoff and Bashir, 2015). We investigate participants' situational trust using the validated Multi-dimensional Trust Questionnaire (MTQ) (Roesler et al., 2022), and measure the three dimensions: perceived AI transparency (*"The way the system works is clear to me"*, *"I am well informed about how the system works"*, *"I understand how the system works"*), performance (*"The system works safely"*, *"The system works well"*, *"The system works accurately"*), and utility (*"The system makes my work more difficult"*, *"The system is useful for my work"*, *"I find that the system supports my work"*). Each three-item subscale is rated on a 4-point scale from *"Disagree"* to *"Agree"*.

3.4.3. Agreement and user confidence

We adopted the agreement and confidence measures from prior research (Kim et al., 2024; Pareek et al., 2024) and captured participants' agreement with the AI's judgement by asking "*Do you think the AI's judgement is accurate?*" with a binary response option. We further asked how confident participants were in their (dis)agreement with the AI. Participants' self-confidence was measured on a slider ranging from 1 (not confident at all) to 100 (fully confident) captured in 10-point increments. To avoid potential biases through anchoring effects (Sellers, 2013), the slider handle was only displayed after participants actively clicked on the slider. We measured both agreement and self-confidence after each sample assessment case, primarily to serve as a robustness check for our manipulation of varying AI performance across Block 1–3.

3.5. Procedure

We conducted several pilot studies to finalise the experimental flow of our user study; see Fig. 3 for an overview. Given the 2 × 4 between-subjects design, we randomly assigned participants to one of the *Stakes* and *Strategy* conditions.

Participants were first guided to read a plain language statement of the study. After giving consent, participants received a pre-task questionnaire to collect demographic information and participants' propensity to trust automation (Körber et al., 2019). Subsequently, participants received general instructions about the bacterial diagnostic task and how to assess the sample's infestation through visual estimation. The explanation was illustrated with a vignette, including an example case that displayed a microscope image alongside a corresponding image of either a human patient or a plant to set the scenario for the *Stakes* (human tissue vs. plant tissue sample). For task familiarisation, participants then completed a training block with three practice trials, distinguished from the main experiment by the different staining of the microscope images (Fahnenstich et al., 2024), and could only proceed once all training trials were completed successfully. We further administered an attention

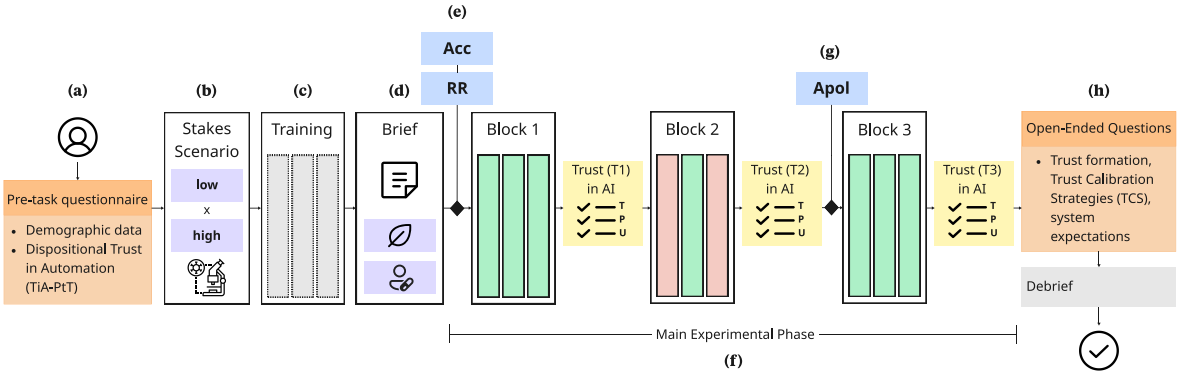


Fig. 3. The experimental flow. (a): Pre-task questionnaire. (b): Operationalisation of *Stakes* (plant tissue vs. human tissue, between-subjects). (c): Training Phase. (d): Brief (medical centre vs. plant nursery). (e) and (g): Positioning of the *Strategy* between-subjects manipulation. (e): A priori *Strategies*: AI accuracy (`Acc`); Relational risk (`RR`). (f): Block 1–3 with repeated measures of perceived AI trustworthiness (transparency, performance, utility). (g): Apology (`Apol`) as reactive post error *Strategy*. Participants in the Baseline condition received no *Strategy* at (e) or (g). (h): Post-task open-ended question.

check and a comprehension check, both of which we used as exclusion criteria to ensure data quality.

Prior to the main experimental phase, we informed participants that, for each assessment, they would be supported by an AI agent. They then received a 'brief', see Section 3.2.1 for more details, that situated the main experimental phase in a real-world context while emphasising the respective *Stakes*.

During the main experiment, participants completed three blocks with varying AI performance, as described in Section 3.3. In each block, they assessed three microscope samples, one by one: for each sample, participants viewed the image, and after a three-second delay with an animated loading display, received the AI's identification. We implemented the delay to increase the perceived authenticity of the AI model (Pareek et al., 2025). After each case, participants reported their agreement with the AI's classification and their self-confidence. Subsequent to each block, we captured participants' perceptions of AI trustworthiness (transparency, performance, utility) (Roesler et al., 2022).

The *Strategies* were manipulated between-subjects, and their positioning during the main experiment was in line with their conceptual purpose (see Fig. 3). Participants in the `Acc` and `RR` conditions received their messages before Block 1, as a global a priori strategy intended to prime users before the interaction. Participants in the `Apol` condition received their messages after Block 2, consistent with its role as a reactive *Strategy* to navigate trust after AI failure and in accordance with prior work (Kim and Song, 2021; Kim et al., 2024; Pareek et al., 2024). Participants in the Baseline condition received no explicit communication about the AI system.

After the main experiment, participants responded to qualitative open-ended questions that asked them to reflect on their evolution of trust and what caused it, specifically after the AI made incorrect suggestions, and how the respective *Strategy* they received (`Acc`, `RR`, `Apol`) influenced their expectations and understanding of the AI.

We ended the study with a debrief stating that the scenario and samples were fictional and that there were no real risks associated with the experiment.

3.6. Participants

We conducted an a priori power analysis to determine our sample size using the R package `InteractionPoweR` (Baranger et al., 2023) to account for interaction effects. The calculation was based on an $\alpha = .05$ and a power of.9, resulting in a targeted sample size of at least $N = 204$. We deployed our study on the crowdsourcing platform Prolific.[2] We applied several pre-screening criteria to ensure participants were proficient in English, aged ≥ 18, had a prior approval rate of ≥ 98%, and to maintain a balanced gender distribution. Further, to mitigate learning effects, each participant could take part only once. Participants were compensated at £9.00 per hour; with an estimated completion time of 12 min, this converted to average payments of around £1.80, rated "good" under the platform's payment standards.[3]

We aimed for a balanced sample of 52 participants per condition ($N = 208$ in total) to meet the power analysis requirements and recruited participants until this target was reached. We excluded participants who did not finish the survey or who failed the attention check or the comprehension check. The majority were excluded because they failed the comprehension check ($N = 27$), which tested whether participants understood and could recall the infestation threshold by applying it to an example case and classifying it as infested or not infested. The final sample comprised $N = 208$ participants with an average age of 39.06 ($SD = 13.35$).

[2] https://www.prolific.com/.

[3] https://researcher-help.prolific.com/en/article/9cd998.

4. Results

We employed a Cumulative Link Mixed Model (CLMM) to examine how different *Strategies* (`Acc`, `RR`, `Apol`) influence users' perceived AI trustworthiness across its constituent facets (transparency, performance, utility) and across varying AI performance (Block 1–3) (RQ1). We further analysed how these strategies affect users' expectations and understanding of the AI system (RQ2). Finally, we assessed how the *Stakes* inherent to the decision-making context affect perceptions of AI trustworthiness across varying AI performance (RQ3.1), and how they may interact with the investigated *Strategies* in shaping these perceptions (RQ3.2). We conducted post-hoc analyses of the interaction using Tukey-adjusted faceted contrasts of Estimated Marginal Means (EMMs) to correct for multiple comparisons within each level of the interacting variables. We report these comparisons alongside the corresponding EMMs. Effect sizes are reported as odds ratios, a suitable interpretation of logistic regression, which represent the change in odds that an outcome (perceived AI trustworthiness) will occur given a one-unit change in the predictor variables (*Strategy*, *Stakes*) (Field et al., 2014).

4.1. CLMM construction

To analyse our RQs, we employed Cumulative Link Mixed Models, which have been increasingly advocated for in HCI research as an appropriate frequentist approach for ordinal data, particularly in repeated measures designs where random effects must be accounted for (Victor Syiem and Velloso, 2026). Our data share these characteristics, given their ordinal nature, repeated-measures structure, and the need to account for random effects and covariates. We therefore modelled perceptions of AI trustworthiness as the dependent variable, operationalised across dimensions with TrustDim ∈ {transparency, performance, utility} and block $b \in \{1, 2, 3\}$ ($TrustDim_b$). The model included *Strategy* and *Stakes* as independent variables, as well as their interaction.

In our model, we incorporated propensity to trust automation ($TiA-PtT$) as a covariate to adjust for participants' predispositions (Hoff and Bashir, 2015). We further included $CentredAge$ as a covariate to control for any effects related to age, given an age range of 18–77 in our data. To account for individual differences and the repeated measures design, we included participants' IDs (PID) as a random effect. The Cronbach's α values for the respective dimensions transparency ($\alpha = 0.91$), performance ($\alpha = 0.76$), and perceived utility ($\alpha = 0.71$) indicate acceptable to excellent internal consistency among items. Nevertheless, to account for potential item-specific variance in the responses, we further included $Item$ as a random effect. For the Block 2 and Block 3 models, we included the trust score from the preceding block(s) $\sum_{j=1}^{b-1} TrustDim_{Block\,j}$ as an additional predictor, where $b \in \{2, 3\}$ denotes the current block and j indexes the preceding block(s), thereby modelling the temporal dependency of trust across blocks (Pareek et al., 2024). Given that `Apol` was introduced after Block 2 due to its conceptual purpose as a reactive trust recovery strategy post failure, responses from `Apol` and Baseline were aggregated for Blocks 1 and 2, as there were no experimental differences between these experimental blocks. We used the statistical R package `ordinal` [4] to build our CLMM, which was formulated as follows:

$$TrustDim_b \sim Strategy \times Stakes + TiA-PtT + CentredAge + \sum_{j=1}^{b-1} TrustDim_{Block\,j} + (1 \mid PID) + (1 \mid Item)$$

We further assessed the Variance Inflation Factor (VIF) across all employed models to ensure there was no multicollinearity among the independent variables (VIF < 5) (Rawlings et al., 1998). A detailed overview of the constructed models is provided in the Appendix (Table A1),

[4] https://cran.r-project.org/web/packages/ordinal/index.html.

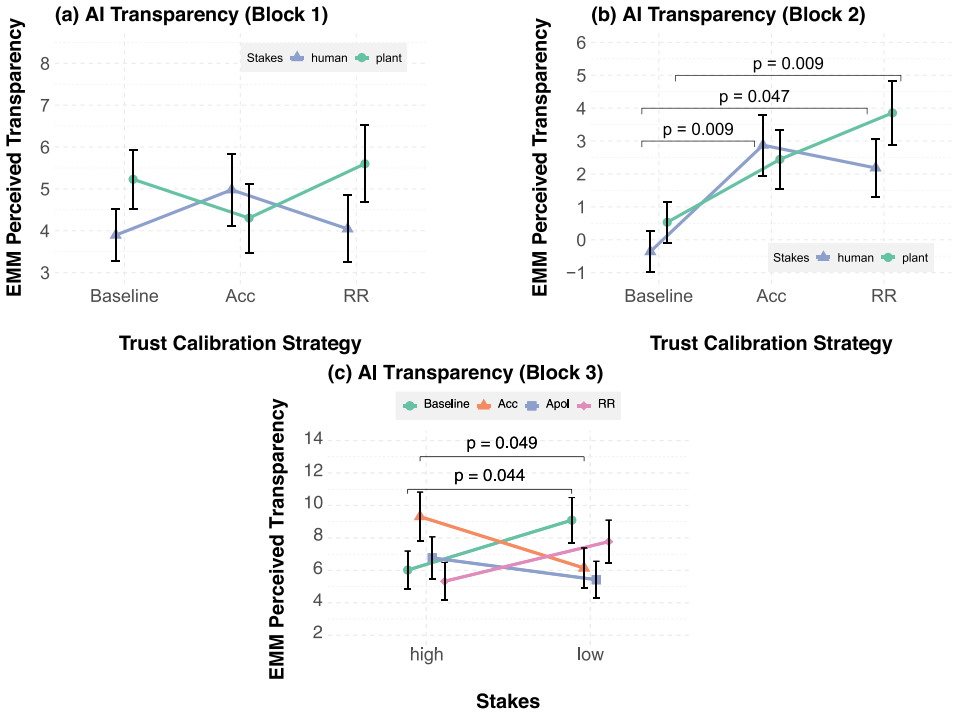


Fig. 4. The interaction of *Strategy* and *Stakes* on perceived AI transparency across (a) Block 1 (correct AI judgements), (b) Block 2 (incorrect AI judgements), and (c) Block 3 (correct AI judgements). Estimated marginal means (EMMs) of perceived AI transparency. In (b), brackets mark significant post hoc pairwise comparisons of *Strategy* within *Stakes*, and in (c) of *Stakes* within *Strategy*. No comparison was significant in (a). Error bars denote standard error (SE).

along with model summaries of the CLMM analysis for perceived AI transparency, performance, and utility (Tables A2–A4).

4.2. Quantitative results

4.2.1. Robustness check - agreement and confidence

We deliberately manipulated the AI's performance across the three blocks to examine how the communicated *Strategy* (`Acc`, `RR`, `Apol`) influences perceptions of AI trustworthiness under varying observable performance, resulting in a simulated overall AI accuracy of 78%. Participants' agreement with the AI agent's identification was analysed after each assessment case as a primary robustness check. In Block 1, the AI provided correct identifications in all three cases; in Block 2, it produced two erroneous identifications; and in Block 3, it again performed correctly across all three trials. Participants' agreement ratings reflected this manipulation, with 96.15% consensus in Block 1, 38.14% in Block 2, and 97.44% in Block 3. We further examined participants' accuracy in agreement (i.e., agreeing when the AI was correct, disagreeing when it was incorrect), which resulted in 24 errors out of 624 responses in Block 1 (96.15%), 66 errors in Block 2 (89.42%), and 16 errors in Block 3 (97.44%). Consistent with prior research (Fahnenstich et al., 2024; Pareek et al., 2024), we assessed participants' confidence in their own decisions as a secondary robustness check of the manipulation. Mean confidence scores were 92.53% ($SD = 11.13$) in Block 1, 84.05% ($SD = 15.96$) in Block 2, and 92.76% ($SD = 11.98$) in Block 3. These ratings confirm that the task operationalisation (observable AI performance) was perceived as intended, with overall high confidence across blocks and a slight but expected drop in Block 2 due to the presence of observable AI errors.

4.2.2. Perceived transparency

For **Block 1**, in which the AI consistently provided correct assessments, neither `Acc` ($\beta = 1.08$, $SE = 0.91$, $p = .233$) nor `RR` ($\beta = 0.15$, $SE = 0.88$, $p = .865$) differed significantly from the Baseline. Our results further show no significant interaction effects between *Stakes* and *Strategy* (all $p > .05$; see Fig. 4a).

In **Block 2**, as the AI exhibited erroneous identifications, our CLMM showed a significant effect of *Strategy* on perceived transparency, with `Acc` ($\beta = 3.24$, $SE = 1.10$, $p = .003$, $OR = 25.47$) and `RR` ($\beta = 2.55$, $SE = 1.07$, $p = .018$, $OR = 12.74$) significantly increasing perceived transparency relative to Baseline. We further observed that prior block-level transparency perceptions significantly predicted transparency perceptions in **Block 2** ($\beta = 1.24$, $SE = 0.26$, $p < .001$, $OR = 3.47$). Post-hoc faceted contrasts revealed that the effect of *Strategy* was moderated by *Stakes*. Imperatively, `RR` consistently led to higher perceived transparency relative to Baseline, across both *low*- (`RR`: $EMM = 3.86$, $SE = 0.96$ vs. Baseline: $EMM = 0.53$, $SE = 0.62$; $\beta = 3.33$, $SE = 1.12$, $p = .009$, $OR = 27.94$) and *high Stakes* contexts (`RR`: $EMM = 2.18$, $SE = 0.89$ vs. Baseline: $EMM = -0.36$, $SE = 0.62$; $\beta = 2.55$, $SE = 1.07$, $p = .047$, $OR = 12.81$), see Fig. 4(b). Conversely, `Acc` ($EMM = 2.88$, $SE = 0.92$) significantly increased transparency perceptions compared to Baseline ($EMM = -0.36$, $SE = 0.62$) specifically under *high Stakes*; $\beta = 3.24$, $SE = 1.10$, $p = .009$, $OR = 25.54$), with no significant effect under *low Stakes* ($\beta = 1.91$, $SE = 1.06$, $p = .168$).

In **Block 3**, in which the AI again provided correct assessments, the CLMM revealed a significant effect of *Strategy*; `Acc` led to significantly higher perceived transparency than Baseline ($\beta = 3.29$, $SE = 1.61$, $p = .041$, $OR = 26.88$). Importantly, *Strategy* moderated the effect of *Stakes* on perceived transparency. Specifically, under the Baseline condition, participants in the *high Stakes* context reported significantly lower transparency compared to those in the *low Stakes* context ($\beta = -3.07$, $SE = 1.53$, $p = .044$, $OR = 21.54$). By contrast, when AI accuracy was communicated (`Acc`), the pattern reversed, with participants in the *high Stakes* context reporting significantly higher transparency than those in the *low Stakes* context ($\beta = 3.17$, $SE = 1.61$, $p = .049$, $OR = 23.81$); see Fig. 4 (c). No moderating effect of *Strategy* emerged for `Apol` ($p = .380$)

or RR ($p = .107$). Preceding transparency perceptions from **Block 1** ($\beta = 2.12$, $SE = 0.34$, $p < .001$, $OR = 8.36$) and **Block 2** ($\beta = 1.53$, $SE = 0.31$, $p < .001$, $OR = 4.61$) significantly predicted transparency ratings in **Block 3**.

Our findings reveal no significant effect of *Stakes* on perception of AI transparency across **Block 1** ($\beta = 1.33$, $SE = 0.75$, $p = .076$) and **Block 2** ($\beta = 0.89$, $SE = 0.86$, $p = .299$). In **Block 3**, however, *low Stakes* is perceived as significantly more transparent than *high Stakes*; $\beta = 3.07$, $SE = 1.53$, $p = .044$, $OR = 21.63$.

4.2.3. Perceived AI performance

Our analysis showed no significant effect of *Strategy* on participants' perceptions of AI performance in **Block 1**, with neither Acc ($\beta = -0.65$, $SE = 0.99$, $p = .516$) nor RR ($\beta = -1.77$, $SE = 0.99$, $p = .074$) differing significantly from Baseline. Likewise, no significant interaction emerged between *Stakes* and *Strategy* (all $p > .05$); see Fig. 5(a). In **Block 1**, participants' dispositional trust in automation significantly predicted their performance perceptions of the AI agent ($\beta = 1.388$, $SE = 0.56$, $p = .013$, $OR = 4.01$).

After the AI exhibited failures in its assessment (**Block 2**), our analysis revealed a significant effect of *Strategy*, with Acc significantly increasing perceived AI performance compared to Baseline ($\beta = 2.63$, $SE = 1.06$, $p = .014$, $OR = 13.82$). No significant difference emerged for RR relative to Baseline ($p > .05$). Again, participants' dispositional trust in automation significantly predicted their performance perceptions ($\beta = 1.63$, $SE = 0.56$, $p = .004$, $OR = 5.19$). Notably, faceted post-hoc contrasts further demonstrated that this effect was moderated by *Stakes*. Specifically, Acc ($EMM = 1.15$, $SE = 0.94$) increased performance perceptions compared to Baseline ($EMM = -1.48$, $SE = 0.72$) under *high Stakes*; $\beta = 2.63$, $SE = 1.06$, $p = .036$, $OR = 13.82$), whereas we observed no significant effect under *low Stakes* ($\beta = 1.91$, $SE = 1.04$, $p = .160$); see Fig. 5(b). These results suggest that when the AI exhibited errors, AI accuracy communication prior to the interaction enhances perceived AI performance under *high Stakes*.

When the AI returned to provide correct assessments (**Block 3**), we found a significant effect of *Strategy*; Acc led to significantly higher perceptions of AI performance relative to Baseline ($\beta = 3.20$, $SE = 1.27$, $p = .012$, $OR = 24.42$). No significant differences were found for RR or Apol compared to Baseline ($p > .05$). Notably, faceted contrasts showed that Acc elicited significantly higher perceptions of AI performance compared to RR under *high Stakes* ($EMM = 8.80$, $SE = 1.24$ vs. $EMM = 4.08$, $SE = 0.91$; $\beta = 4.72$, $SE = 1.28$, $p = .001$, $OR = 112.17$), whereas no significant difference was observed under *low Stakes* ($\beta = -1.62$, $SE = 1.16$, $p = .505$); see Fig. 5(c). Alongside this result, we found that *Strategy* moderated the effect of *Stakes* on perceived AI performance. Specifically, when presented with Acc, participants in the *high Stakes* scenario ($EMM = 8.80$, $SE = 1.24$), i.e., human tissue samples, reported significantly higher performance perceptions than those in the *low Stakes* scenario, i.e., the inspection and classification of plant tissue samples ($EMM = 4.33$, $SE = 0.94$; $\beta = 4.47$, $SE = 1.28$, $p < .001$, $OR = 87.08$).

In **Block 3**, participants' dispositional trust in automation significantly predicted their performance perceptions ($\beta = 1.25$, $SE = 0.57$, $p = .028$, $OR = 3.50$), alongside the effect of preceding **Block 1** performance perceptions ($\beta = 1.74$, $SE = 0.34$, $p < .001$, $OR = 5.70$).

The *Stakes* inherent to the decision-making context did not significantly affect participants' perceptions of AI performance, either when the AI produced correct assessments in **Block 1** ($\beta = 1.06$, $SE = 0.85$, $p = .214$) and **Block 3** ($\beta = 1.73$, $SE = 1.24$, $p = .161$), or when the AI exhibited erroneous assessments in **Block 2** ($\beta = -0.79$, $SE = 0.85$, $p = .353$).

4.2.4. Perceived AI utility

Across the initial error-free Block (**Block 1**) and the subsequent erroneous Block (**Block 2**), we found no significant effect of *Strategy* on participants' perceptions of AI utility compared to Baseline (all $p > .05$). Moreover, no significant interactions with *Stakes* emerged; faceted contrasts confirmed the absence of any moderation effects (all $p > .05$). Our

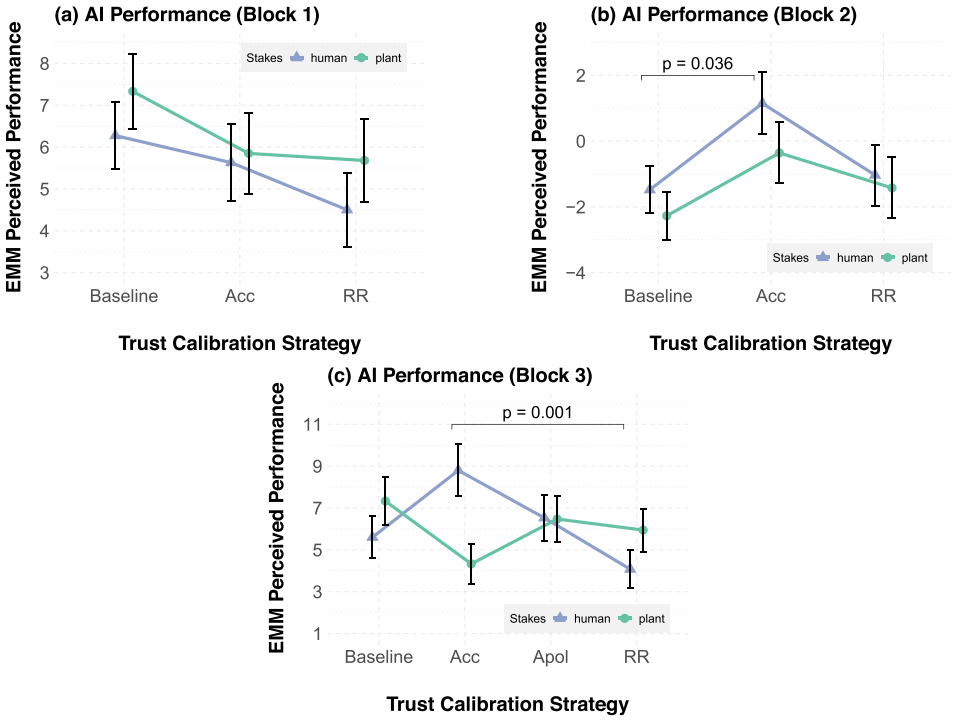


Fig. 5. The interaction of *Strategy* and *Stakes* on perceived AI performance across (a) Block 1 (correct AI judgements), (b) Block 2 (incorrect AI judgements), and (c) Block 3 (correct AI judgements). Estimated marginal means (EMMs) of perceived AI performance. In (b) and (c), brackets mark significant post hoc pairwise comparisons of *Strategy* within *Stakes*. No comparison was significant in (a). Error bars denote standard error (SE).

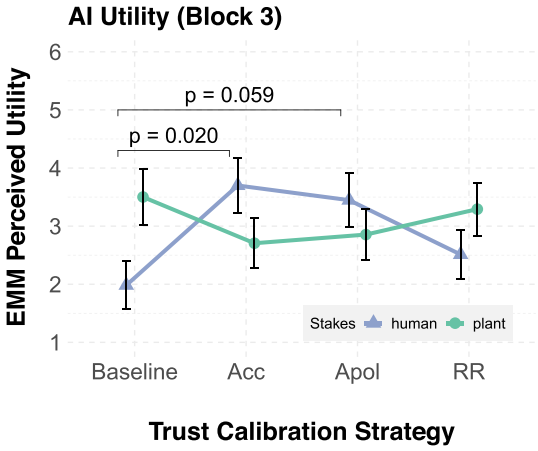


Fig. 6. Perceived AI utility in Block 3. Error bars denote standard error (SE).

findings suggest that perceived AI utility is resilient to both AI accuracy and Relational Risk signalling. Participants' dispositional trust in automation significantly predicted utility perceptions in both **Block 1** ($\beta = 0.78$, $SE = 0.30$, $p = .009$, $OR = 2.18$) and **Block 2** ($\beta = 0.95$, $SE = 0.39$, $p = .015$, $OR = 2.58$), while preceding block-level utility perceptions emerged as an additional significant predictor in **Block 2** ($\beta = 0.57$, $SE = 0.18$, $p = .002$, $OR = 1.78$).

In **Block 3**, when the AI returned to providing accurate assessments, our analysis revealed significant effects of *Strategy*, with `Acc` ($\beta = 1.72$, $SE = 0.59$, $p = .004$, $OR = 5.57$) and `Apol` ($\beta = 1.46$, $SE = 0.58$, $p = .012$, $OR = 4.33$) significantly increasing perceived AI utility compared to Baseline. Post-hoc faceted contrasts further demonstrate *Stakes* as a moderator; `Acc` ($EMM = 3.70$, $SE = 0.48$) led to greater perceived utility when the AI agent assisted under *high Stakes*, i.e., human tissue samples, compared to Baseline ($EMM = 1.98$, $SE = 0.41$); $\beta = 1.72$, $SE = 0.59$, $p = .020$, $OR = 5.57$; see Fig. 6), whereas in the *low Stakes* plant inspection contexts, the effect was directionally reversed but not significant ($\beta = -0.80$, $SE = 0.59$, $p = .526$). A comparable albeit weaker trend was observed for `Apol`: under *high Stakes*, `Apol` was associated with somewhat higher utility ratings than Baseline, though this difference did not reach significance ($p = .059$); under *low Stakes*, the effect was again reversed but non-significant. *Strategy* further moderated the effect of *Stakes* on perceived AI utility in **Block 3**. Specifically, under the Baseline condition, participants in the *high Stakes* context (human tissue samples) reported significantly lower AI utility perceptions compared to those in the *low Stakes* context (plant tissue samples); $\beta = -1.52$, $SE = 0.59$, $p = .010$, $OR = 4.57$. Both preceding **Block 1** ($\beta = 1.82$, $SE = 0.18$, $p < .000$, $OR = 6.18$) and **Block 2** ($\beta = 0.38$, $SE = 0.14$, $p = .006$, $OR = 1.46$) utility perceptions emerged as additional significant predictors in **Block 3**.

Similar to perceptions of AI transparency, our findings show no significant effect of *Stakes* on perceived AI utility across **Block 1** ($\beta = 0.242$, $SE = 0.46$ $p = .596$) and **Block 2** ($\beta = -0.20$, $SE = 0.59$, $p = .730$). Notably, *low Stakes* led to a significantly increased perception of AI utility compared to *high Stakes* in **Block 3**; $\beta = 1.52$, $SE = 0.59$, $p = .010$, $OR = 4.57$.

4.3. Qualitative data analysis

We analysed our qualitative data following Thomas' general inductive approach (Thomas, 2006) to gain an in depth understanding of our research objectives. Following the steps outlined in Thomas (2006), two researchers from our research team coded the open-ended responses provided in the post-task questionnaire. While guided by our evaluation objectives with respect to RQ1–3, as proposed by Thomas (2006), we specifically sought to gain an in-depth understanding of the mechanisms by which the *Strategies* operate in shaping trust, as well as RQ2, i.e., how the communication of Relational Risk (`RR`) compared to other *Strategies* shapes expectations and users' understanding of the AI agent under varying AI performance and *Stakes*. The first coder thoroughly familiarised themselves with the data and developed an initial set of preliminary categories from the raw open-ended responses, while the second coder independently developed a second set of categories. Subsequently, in four iterative rounds, the two coders reviewed, compared and discussed their categories to eliminate any ambiguities and discrepancies and to establish an accurate and shared interpretation of the data. The two coders refined and merged categories to then deductively code the qualitative responses until a satisfactory coding alignment was reached (Thomas, 2006; McHugh, 2012), with an inter-rater reliability of Cohen's $\kappa = 0.81$. We highlight the categories we identified following the general inductive approach in bold (Thomas, 2006), and present participants' responses to exemplify our findings while contextualising them with a 3-tuple (PID, *Stakes*, *Strategy*) to reflect their experimental condition. PID refers to the participant ID.

4.3.1. AI accuracy as a Strategy to shape perceptions of AI

Communicating AI accuracy (`Acc`) as an a priori performance indicator made a considerable number of participants **aware that the AI could make mistakes**; *"I was told that the AI had 78% accuracy. I kept that in mind when deciding if I agreed or didn't agree with it. So, I think while I trusted it, I knew it had some wiggle room for error."* – (P20, human, `Acc`). This awareness often translated into **slightly tempered expectations** of the AI agent; *"Due to the error margin, I expected some inaccurate suggestions by the AI agent, but not so much. Hence, the trust in the AI agent wasn't that high."* – (P115, plant, `Acc`)

Alongside these impressions, our qualitative analysis revealed that several participants reported being **attentive as a result of the stated AI accuracy**, though to varying degrees. For example, one participant described being *"cautiously optimistic,"* noting: *"I understood it was fairly reliable but not infallible, so I approached its responses with a balanced mix of trust and scepticism"* – (P123, plant, `Acc`). Others expressed a more apprehensive stance; *"I expected the AI to make correct guesses, but there was still room for mistakes. So I had to be very attentive"* – (P13, human, `Acc`).

Conversely, several participants articulated that the communicated AI accuracy **raised their performance expectations** and instilled confidence in the AI's performance; *"I expected it to work pretty well overall."* – (P4, human, `Acc`); *"It made me very confident in its abilities"* – (P105, plant, `Acc`). These statements align with our quantitative findings, which showed that `Acc` significantly enhanced perceptions of AI performance when the AI erred (Block 2) and returned to correct identifications (Block 3) compared to the Baseline, and compared to communicating Relational Risk (`RR`) in Block 3, all notably so under *high Stakes*.

4.3.2. Relational risk as a Strategy to shape perceptions of AI

In line with AI accuracy, the communication of Relational Risk (`RR`) prior to the interaction prompted participants to **anticipate error occurrences in the AI's outputs**; *"I knew it could and would make mistakes"* – (P80, human, `RR`); *"It made me aware that the AI may give incorrect information"* – (P194, plant, `RR`). The recognition of the possibility of errors led many to **recalibrate their expectations of the AI**; *"It prepared me mentally in that I was expecting some but not most of its answers to be wrong. I did not trust it 100%."* – (P79, human, `RR`); *"The message about the AI's 22% error rate set my expectations for occasional mistakes, prompting me to approach its recommendations with a critical and nuanced perspective."* – (P95, human, `RR`).

As hinted in the latter statement, the communication of Relational Risk **elicited a more vigilant stance**: *"Being challenged to be careful before making a choice based on the AI"* – (P189, plant, `RR`), and *"It makes you pay more attention, to verify, and not rely on the AI"* – (P205, plant, `RR`), and frequently **instilled a heightened sense of caution**: *"It made me more cautious about blindly accepting its recommendations"* – (P96, human, `RR`).

As a result, several participants reported to **scrutinising the AI more thoroughly**, which manifested in a more deliberate and diligent assessment of the AI's output as a behavioural consequence; *"I found myself scrutinising the microscope images much more closely, I'd do my own double-check. It added an extra layer of cognitive effort because I couldn't just trust it implicitly any more."* – (P96, human, RR)

Notably, the communicated error rate, as an inherent part of the Relational Risk approach to shaping trust perceptions and expectations, was perceived differently across participants. Some participants regarded the 22% error rate as high and responded with scepticism:*"It made me very cautious as that's a large margin of error"* – (P91, human, RR). Some participants, by contrast, considered the error rate as acceptable, emphasising that while it increased awareness of the AI's potential fallibility, it did **not erode the AI's overall utility**, which is in line with our quantitative findings (see Section 4.2.4); *"I began to have cautious faith in the AI agent because I was aware of its 22% error rate up front, which helped me realise that it's a useful but limited tool"* – (P203, plant, RR), and *"My expectations were that it would get some wrong but it still ended up being useful in the end"* – (P82, human, risk).

4.3.3. Apology as a Strategy to shape perceptions of AI

Our qualitative findings revealed mixed results regarding the effect and interpretation of AI apology (Apol) as a mechanism to influence trust perceptions. On the one hand, the Apol strategy was often interpreted as **the AI taking accountability for its errors** and as a promise of improved future performance, which contributed to restoring trust; *"That message showed accountability, which helped rebuild trust. It signalled the AI recognises mistakes and aims to improve"* – (P29, human, Apol). On the other hand, several participants regarded the Apol strategy as **ineffective in shaping trust perceptions or supporting system understanding**. *"This message did nothing for me. An AI apologising is pointless and means absolutely nothing."* – (P140, plant, Apol). Beyond outright dismissal, some participants stressed that an AI apology alone could not meaningfully influence their trust in the AI unless it was accompanied by evidence of improvement in the AI's behaviour; in this sense, trust was conditional on observable corrections following the apology. *"I was glad but it did not affect my expectations as I needed proof to see it was working well first. That was just an apology."* – (P45, human, Apol). A few participants further emphasised that its ineffectiveness also originated from the lack of explanatory value, as the message did not support their understanding of the AI system; *"It didn't affect my trust at all because it wasn't an explanation for why it made the errors."* – (P30, human, Apol).

4.3.4. Stakes

Our qualitative analysis revealed some **awareness of the *Stakes* inherent to the AI's decision-making context**, albeit limited. Notably, most of these reflections emerged from the *high Stakes* group, where participants emphasised the potentially severe consequences associated with the diagnostic task involving human tissue samples and the implications for patients' health and lives; *"People's lives depend on these judgements"* – (P26, human, Acc). When considering the assigned *Strategy*, the majority of expressions originated from participants in the *high Stakes* group who received Relational Risk (RR) as an intervention *Strategy* prior to the interaction; *"[…] my pick could harm real life patients."* – (P89, human, RR), and *"it caused me to pause and reassess how much I could rely on it. Even a single error in a clinical or diagnostic context can have significant consequences."* – (P86, human, RR)

5. Discussion

Real-world consequences due to misplaced trust in fallible AI systems (Cao et al., 2024; McGregor, 2021) warrant interventions to reduce the gap between actual and perceived trustworthiness in AI-assisted decision-making across applications (Schlicker et al., 2025a). Our study responds to this concern by examining the communication of Relational Risk, as defined by Stuck et al. (2022), as a *Strategy* to shape trust perceptions under varying AI performance, and by comparing it to established existing approaches, i.e., AI accuracy and apology. We find that Relational Risk, operationalised through communicating the likelihood of an AI error and its consequences (Wilson et al., 2019; Dunlop et al., 2008) while adopting a negative framing (McNeil et al., 1982; Peters et al., 2011), fosters transparency perceptions without undermining utility or inducing AI-averse perceptions. We also identify *Decision-Stakes* as an important determinant of the *effectiveness* of these *Strategies* to increase or decrease perceptions of system trustworthiness, and, in turn, of their *appropriateness*. For instance, Acc significantly increased perceived AI performance under *high Stakes* after the AI erred, which can be considered effective but inappropriate in a context in which misplaced trust carries severe consequences. Our work provides a fine-grained assessment of perceived AI trustworthiness across its dimensions of transparency, performance, and utility (RQ1), user understanding of the AI (RQ2), and situates these dimensions within varying *Stakes* (RQ3). We extend recent work on communicating AI (in)capabilities and limitations (Rieger et al., 2025; He et al., 2023; Kwon et al., 2018), and contribute to shifting the conversation in human-AI interaction to explicitly communicating risk to users in-situ, while integrating risk theory. In the following, we discuss our results and offer important implications for AI trustworthiness research in the context of AI-assisted decision-making.

5.1. Shaping trust perceptions and system understanding across varying AI performance

5.1.1. The crux of transparency

Recent years present a growing body of literature demonstrating transparency as a central requirement for trustworthy AI systems (Bhatt et al., 2021; Prabhudesai et al., 2023). Prior work emphasises that communicating uncertainty enhances transparency and supports the appropriate formation of trust between humans and AI agents (Bhatt et al., 2021), while withholding such information fosters a false sense of perfectly reliable systems (Prabhudesai et al., 2023). Our results extend this discussion. After erroneous AI behaviour, Relational Risk communication emerges as a consistent means of enhancing transparency perceptions across stakes. Our results support findings by Rieger et al. (2025), showing that communicating AI weaknesses can enhance perceptions of AI transparency. This finding is interesting, since approaches that reliably enhance transparency are desirable. Notably, AI accuracy information produces a context-selective and somewhat paradoxical effect: rather than leading users to question system transparency under high-stakes, it increased transparency perceptions solely in the high-stakes decision context after the AI exhibited erroneous identifications. This pattern warrants caution, as high-stakes contexts ideally should prompt greater scrutiny and information-seeking rather than elevated transparency perceptions. Our qualitative findings revealed that both communicating Relational Risk and AI accuracy made participants anticipate potential AI errors, which contribute to perceptions of AI transparency. While this may explain part of the effect, another may be that accuracy information, by acknowledging fallibility while adhering to a positive framing (Spiegelhalter, 2017), created a sense of (positively interpreted) predictability or controllability perceived as particularly relevant in the high-stakes context, which is known to decrease risk perceptions (Wilson et al., 2019; Walpole and Wilson, 2021), and foster risk-seeking behaviour (McNeil et al., 1982; Peters et al., 2011). When the AI then erred, participants' expectations were confirmed and may have reinforced their perceptions, similar to a confirmation bias (Passi and Vorvoreanu, 2022). The absence of consequence information could have strengthened such positive reinforcement, missing to nudge people's cognitive engagement with the negative consequence associated with the high-stakes context. Another explanation could be that, under low-stakes, accuracy information may be seen as simple background noise; whereas in high-stakes decisions, the same information becomes a valuable signal to shift transparency perceptions. Our results contribute to prior findings showing that users exhibit greater convergence with AI

support compared to human support under high stakes, and that AI advice attracts higher behavioural trust when stakes are high (Fahnenstich et al., 2024). However, more research is needed to understand such intervention strategies and their effects on AI transparency under varying decision-stakes.

Notably, following repeated interactions with both correct and erroneous AI identifications (Block 3), *Strategy* also moderated the effect of *Stakes* on perceived transparency. Whereas participants in the Baseline (no information) condition judged the high-stakes context as significantly less transparent than the low-stakes context, AI accuracy information led participants to perceive the high-stakes context as significantly more transparent. However, when Relational Risk is communicated, there is no such effect. Similar to explanations described beforehand, this could be due to the positive frame valence of the approach (accuracy rate vs. error rate), which reassures users in high-stakes contexts and inflates transparency perceptions, whereas Relational Risk communication keeps awareness of potential failure salient and relevant across stakes. Qualitative insights support this interpretation, as explicit references to consequences were concentrated predominantly in the high-stakes `RR` condition and appeared rarely in the `Acc` condition. Although the communicated consequence in `RR` concerned Relational Risk rather than situational stakes, it may still have shaped how participants interpreted the stakes, possibly through a more critical cognitive engagement. This pattern is concerning: users receiving AI accuracy information may be satisfied with minimal disclosures precisely where more comprehensive transparency is warranted. It demonstrates the danger of inappropriate intervention cues in human-AI interaction, which risk overstating the openness of AI systems in contexts where consequences are most severe, creating an illusion of transparency where users should request more information. **We argue that AI accuracy information should solely serve as a means to support people's perception of AI transparency in low-stakes contexts.** We further propose **Relational Risk as a candidate strategy for enhancing perceived AI transparency that warrants further investigation**, given that it maintains the anticipation of potential failures and encourages critical engagement across decision contexts, which we discuss in Section 5.1.4.

Introduced after erroneous AI outputs, apology neither significantly decreased nor enhanced perceptions of transparency. Qualitative insights suggest that while some participants perceived the apology as the AI taking accountability for its mistakes, many regarded it as ineffective, sometimes noting that it did not support their system understanding. This indicates that an apology offers little value for enhancing perceived transparency. These findings contribute to recent works showing that while an apology is effective for increasing trust post failure (Pareek et al., 2024), even more so than simple explanations (Kox et al., 2021), it might not sufficiently support transparency due to its nature as a reactive strategy (LeMasurier et al., 2024).

5.1.2. Performance perceptions: the double-edged role of accuracy

For users interacting with AI systems, understanding and anticipating an AI's performance, i.e., forming expectations about how well an AI system will perform, are essential for determining whether and how much to trust it (Roesler et al., 2022; Lee and See, 2004). Nevertheless, as Passi and Vorvoreanu (2022) argue, this performance calibration is often lacking, which constitutes one of the central challenges for appropriate trust calibration.

Our work provides several insights into how different *Strategies* influence people's performance perceptions. AI Accuracy and Relational Risk information both tempered expectations and encouraged more cautious and attentive engagement with the AI agent and its outcomes, as demonstrated in our qualitative findings. Risk communication induced a particularly vigilant stance, supporting results on exposing AI limitations prior to human-AI interaction (Rieger et al., 2025). Our quantitative findings reveal that stated AI accuracy significantly increased performance perceptions compared to the baseline after the AI exhibited erroneous identifications, and compared to communicating Relational Risk once the AI system returned to accurate behaviour in the final Block. This 'boosting' effect may prevent end users from turning away from the AI after the system exhibits failures. Such an effect could be advantageous in low-stakes applications or domains where expertise primarily lies with the AI, as well as in complementary Human–AI scenarios in which AI expertise is superior and reliance on AI support may be warranted (Bansal et al., 2021; Pareek et al., 2024). However, increased AI performance perceptions induced by the communication of AI accuracy were most pronounced in the high-stakes decision-making context, where elevated perceptions of AI performance may be problematic and where a certain degree of wariness may be necessary. Our qualitative insights confirm that AI accuracy instilled increased performance expectations and the belief that the AI would be performing well. When looking at this result through the lens of risk communication research, one explanation may lie in the positive nature of the communication itself. AI accuracy presents the AI in terms of successful outcomes, which has been shown to reduce risk perception and promote more risk-seeking decisions (McNeil et al., 1982; Peters et al., 2011; Spiegelhalter, 2017). Importantly, while this positive framing significantly boosted performance perceptions, the negatively-framed Relational Risk strategy did not produce the inverse effect: it did not significantly decrease performance perceptions compared to the baseline. Our findings contribute to explaining the findings by Lai and Tan (2019) that report increased trust even for small accuracy levels, and that accuracy is perceived as a stable (positive) 'trait' of the AI model (Rechkemmer and Yin, 2022).

Our work demonstrates the drawback of accuracy communication: **AI accuracy can inflate AI performance perceptions in the system precisely where scepticism is warranted, creating a risk of misplaced trust in critical contexts**. At the same time, it suggests that **accuracy communication may still be a viable strategy in low-stakes contexts, where heightened performance perceptions are desired and less consequential.**

5.1.3. Communicating relational risk does not diminish AI utility

During the initial error-free performance of the AI agent and the subsequent block of erroneous assessments, neither AI accuracy communication nor Relational Risk communication significantly altered perceptions of AI utility. This stability is noteworthy. Despite its negative framing, Relational Risk communication led the AI to be perceived as "a useful but limited tool", without inducing signs of algorithmic aversion (Passi and Vorvoreanu, 2022). Hence, **communicating the Relational Risk does not diminish perceived utility, which indicates that users continue to see value in the AI system even when its fallibility is explicitly foregrounded.** This is in line with the objectives of effective risk communication — to "alert the target audience", but "without arousing unnecessary anxiety" (Breakwell, 2000). Our finding contributes to the results by Rieger et al. (2025) which show that communicating AI limitations may reduce initial trust, but does not erode it over the course of interaction. Qualitative insights add nuance to this interpretation. Some participants emphasised the high error rate and described heightened risk perception, while others stated that the communicated error rate did not diminish their sense of utility. The emphasis on error rate raises the question of whether it is the likelihood of the AI erring or the consequence component of Relational Risk that constitutes the more prominent determinant of perceived AI trustworthiness. Notably, after participants observed varying AI performance during repeated interactions, both accuracy and apology significantly increased perceptions of AI utility compared to the Baseline (Block 3). Accuracy, again, significantly elevated the perceived usefulness of the AI in precisely the context where misplaced trust could be most problematic, i.e., high-stakes, which re-emphasises the drawback of AI accuracy described beforehand.

5.1.4. Relational risk as a cognitive forcing function?

The communication of Relational Risk elicited healthy scrutiny towards the AI which often manifested as a deliberate assessment of the AI's output, e.g., double-checking its identification. This effect resonates

with the idea of cognitive forcing functions as an approach to interrupt users' tendencies to operate in autopilot mode and to instead encourage more deliberate, System 2 thinking (Kahneman, 2011; Buçinca et al., 2021; Park et al., 2019). It also echoes the need to implement interventions that elicit human oversight (Passi and Vorvoreanu, 2022). If we aspire to make "users the last line of defense against AI failures" in real-world AI applications (Passi and Vorvoreanu, 2022), we must equip users with the ability to recognise and navigate both the capabilities and limitations of these systems to elicit appropriate perceptions of AI trustworthiness given the associated risks. This requires encouraging users to approach AI decisions and recommendations with careful assessment rather than applying heuristics, and promoting a more critical engagement with AI outputs in AI-assisted decision-making (Buçinca et al., 2021; Park et al., 2019; Shen et al., 2024). While established cognitive forcing functions typically withhold the AI's output temporarily to prompt deliberation (Buçinca et al., 2021; Park et al., 2019; Miller, 2023), RR introduces no such procedural friction as an a priori communication strategy. Nonetheless, our findings suggest that **Relational Risk communication may have the potential to serve as a function analogous to a cognitive forcing function**, by prompting deliberate engagement and instilling scrutiny through foregrounding the AI's potential fallibility and the associated consequences upfront. Future work should investigate its effectiveness regarding perceived trustworthiness and cognitive engagement compared to existing cognitive forcing functions such as "on demand", "update", or "waiting" features (Buçinca et al., 2021).

However, Relational Risk communication is not without drawbacks. While it can foster healthy scrutiny and deliberate assessment, in low-stakes contexts such vigilance may impose unnecessary cognitive load or time costs (Hoesterey and Onnasch, 2023). In these settings, AI accuracy information may be preferable, as it can induce attentiveness without prompting undue scrutiny. Prior work suggests that scrutiny and verification behaviour in high-risk automation contexts may decline over time due to habituation (Hoesterey and Onnasch, 2023); thus, **a key question is whether Relational Risk communication can sustain caution without leading users to become habituated to Relational Risk communication in human-AI interaction**.

Our findings indicate that AI accuracy and Relational Risk communication each offer strengths under different conditions. Recognising this complementarity lays the groundwork for future adaptive approaches that could investigate *RR* as an adaptive strategy following Buçinca et al. (2024)'s offline reinforcement learning approach for adaptive support policies that provide assistance based on e.g., individual differences and task context to determine when *RR* should be conveyed, which may also help address potential habituation effects. In line, we implemented Relational Risk as an a priori communication. Future research should examine its implementation alongside AI outputs 'in-the-moment' (Buçinca et al., 2024), akin to model confidence, and explore the balance between effectiveness and a potential "crying wolf" effect (Wickens et al., 2009) of Relational Risk communication. That is, investigating the potential risk that a high false-alarm rate and repetitive risk signalling could desensitise users in AI-assisted decision-making over repeated interactions (Wickens et al., 2009; Cañas, 2022; LeClerc and Joslyn, 2015).

5.2. The role of stakes in shaping trust perceptions in AI-assisted decision-making

There is no universally applicable rule for what constitutes appropriate trust, since it is intrinsically linked to context and application (Passi and Vorvoreanu, 2022). While following a false AI recommendation may be tolerable in a low-stakes context such as plant diagnostics, accepting an erroneous decision in a high-stakes context such as medical diagnosis can be detrimental. Any intervention designed to shape trust in AI must therefore remain sensitive to the inherent stakes of the decision-making context and foster a calibration of trust in the appropriate direction (Lee and See, 2004).

In our study, participants perceived the AI system as less transparent and less useful after experiencing varying AI performance when the stakes were high. More importantly, as we discussed above, decision-stakes moderated the influence of communication strategies on all three dimensions of trustworthiness; i.e., transparency, performance, and utility. We therefore argue that the *Strategy* to shape trust in AI systems cannot be examined in a vacuum, but must be contextualised within the stakes in which they are embedded. Importantly, our findings show that stakes influence perceived AI trustworthiness itself, rather than solely playing a role at the level of behavioural trust, where the Organizational Trust Model (Mayer et al., 1995) and Enhanced Model of Trust (Stuck et al., 2022) have situated the effect of stakes. Our results instead align with the recently proposed trustworthiness assessment model (Schlicker et al., 2025a), which posits an initial risk–benefit assessment (Schlicker et al., 2025b) during the formation of trustworthiness perceptions, and we provide empirical support for this theoretical proposition. For HCI research moving forward, this implies a critical shift: stakes are not merely a background variable but a decisive moderator that determines the effectiveness and, crucially, the appropriateness of strategies designed to influence trust in AI systems. Effectiveness captures whether a strategy successfully alters perceptions (e.g., boosting perceived AI performance), while appropriateness concerns whether such changes are desirable in light of the decision context. A strategy that inflates trust under high-stakes conditions may be effective but not appropriate, whereas in low-stakes settings the same intervention may be both effective and appropriate. **This implies that interventions to shape trust perceptions must be designed with decision-stakes in mind, acknowledging that the same *Strategy* can have different effects depending on the stakes of the task.**

5.3. Limitations

Our study defined risk communication as Relational Risk, that is, the risk of using the system itself, which is conceptually distinct from contextual factors (Stuck et al., 2022). We encourage future work to examine the communication of stakes as a consequence component to emphasise risks arising from the context in which the AI is deployed. While the visual estimation task of assessing microscope images is not an everyday scenario, it enabled a directly comparable operationalisation of stakes across contexts. We encourage future work to investigate the communication of Relational Risk across real-world AI-assisted decision-making scenarios and in the context of LLMs to complement current approaches to uncertainty communication such as uncertainty highlighting (Bo et al., 2025), uncertainty language framing (Kim et al., 2024), or in-situ disclaimers (Bo et al., 2025). In line, we acknowledge the small number of trials per block (three trials across three blocks), a design grounded in prior work establishing a detectable performance pattern with a comparatively small trial count (Kahr et al., 2024; Rieger et al., 2025; Tabatabaei et al., 2026). However, we are confident in our operationalisation of AI performance, as confirmed by our robustness check. Motivated by our promising findings, future work should investigate the effects of Relational Risk communication across a larger number of trials incorporating greater variance across instances to examine AI assisted decision-making contexts with higher ecological validity. We further acknowledge that Relational Risk conceptually consists of two distinct components and we further encourage future work to examine the relative contribution of each component. All *Strategies* were implemented in line with their conceptual purpose, which resulted in different points of presentation prior to the interaction or post failure block, which we addressed and accounted for by fitting statistical models for each block. To carefully assess participants' perceptions across varying performances, we deliberately designed observable AI errors for the purpose of this work (i.e., to investigate the effects of *Strategy* on participants perceptions of system trustworthiness in response to correct or erroneous AI judgements). This design choice follows previous work in HCI (Kahr et al., 2024; Tabatabaei et al., 2026) operationalising observable

errors to investigate failure experiences. While we acknowledge that observed AI performance is an important driver of trust in AI-assisted decision-making (Rechkemmer and Yin, 2022; Yin et al., 2019), we held it constant across conditions, and included a baseline condition which reflected observed accuracy only, which allowed us to attribute the observed effects to the investigated *Strategies*. This extends research conducted by Pareek et al. (2024), who observed that trust perception in AI-assisted decision-making is not solely attributable to observed AI accuracy, but can also be shaped by information cues presented to users post AI errors. Nevertheless, we recognise that AI errors may not always be evident in real-world scenarios or that users might not hold the expertise to determine the correctness of the AI's judgement. While we focused on perceptions of AI trustworthiness as a foundational first step to understand the mechanisms and effects of Relational Risk communication in the context of AI-assisted decision-making, we encourage future research to investigate how communicating Relational Risk in particular influences system trustworthiness perceptions under low human expertise and whether our findings extend to reliance behaviour as an important behavioural trust construct in human-AI interaction (Bansal et al., 2021; Buçinca et al., 2021).

6. Conclusion

AI systems and decision support applications remain imperfect when deployed and accessible to users (McGregor, 2021) while spanning domains that range from the seemingly lightweight, such as plant disease detection, to those with profound consequences for human health and life. This work introduces Relational Risk communication as an alternative *Strategy* in AI-assisted decision-making to shape perceptions of AI trustworthiness. We examined its effects alongside AI accuracy and AI apology across the dimensions of perceived AI transparency, performance, and utility, and in interaction with stakes. We further assessed their impact on system understanding. Our findings propose that the communication of Relational Risk may serve as a cognitive forcing function, which encourages conscientiousness, healthy scrutiny and a deliberate assessment of AI identifications, while also raising questions about potential costs in low-stakes contexts and risks of habituation that warrant further investigation. Imperatively, our results show that Relational Risk increases perceptions of transparency across low and high stakes when encountering AI failures and that the perceived AI utility remains resilient to risk signalling, despite the emphasis on the system's fallibility. Established strategies, i.e., AI accuracy and apology, each showed important strengths and drawbacks which we discuss. AI accuracy showed strengths in increasing and sustaining perceived transparency and performance, including post failure, but risked inflating trust in high-stakes settings where caution is warranted. AI apology improved perceived AI utility, but offered little support to enhance system understanding. Across all dimensions, stakes significantly moderated these effects, which we discuss as an important determinant for the effectiveness and appropriateness of *Strategies* to shape trust perceptions. Relational Risk therefore emerges as an important approach for AI-assisted decision-making that offers an alternative and promising intervention to elicit appropriate trust in AI systems. More broadly, our work points to risk communication research as a critical lens for developing strategies to navigate trust dynamics in human-AI interaction and thus to inform future HCI research.

CRediT authorship contribution statement

Sarah Schömbs: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Saumya Pareek:** Writing – review & editing, Formal analysis. **Jorge Goncalves:** Writing – review & editing, Supervision, Conceptualization. **Wafa Johal:** Writing – review & editing, Supervision, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT-5 large language model and Claude Sonnet 4.5 as coding support for the implementation of the task in the survey (JavaScript, HTML) and for proofreading selected parts of the manuscript. After using these tools, the authors of this work carefully reviewed and edited the content as needed. The authors take full responsibility for the content of the manuscript.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Wafa Johal reports that financial support was provided by the Australian Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Our work was supported by the Melbourne Research Scholarship funded by the University of Melbourne and the Rowden White Scholarship established with a donation from Sir Alfred Edward Rowden White. This research was partially supported by the Australian Research Council (Grant No. FT250100459, Grant No. CE260100108).

Appendix A. Model results

This section provides the detailed results of the quantitative analysis. Model specification: $TrustDim_b \sim Strategy \times Stakes + TiA - PtT + CentredAge + \sum_{j=1}^{b-1} TrustDim_{\text{Block } j} + (1 \mid PID) + (1 \mid Item)$. Here, TrustDim $\in$ {transparency, performance, utility} denotes the perceived AI trustworthiness dimensions, $b \in \{1, 2, 3\}$ indexes the experimental block, $Strategy$ refers to the intervention manipulation, and *Stakes* denotes situational risk, operationalised through the decision context.

Table A1
Block-specific CLMM formulas instantiated from the overarching model specification. All models include random intercepts for participants (PID) and items.

Block	Model formula
1	$\text{TrustDim}_{\text{Block 1}} \sim Strategy \times Stakes + TiA - PtT + CentredAge + (1 \mid PID) + (1 \mid Item)$
2	$\text{TrustDim}_{\text{Block 2}} \sim Strategy \times Stakes + TiA - PtT + CentredAge + \text{TrustDim}_{\text{Block 1}} + (1 \mid PID) + (1 \mid Item)$
3	$\text{TrustDim}_{\text{Block 3}} \sim Strategy \times Stakes + TiA - PtT + CentredAge + \text{TrustDim}_{\text{Block 1}} + \text{TrustDim}_{\text{Block 2}} + (1 \mid PID) + (1 \mid Item)$

Table A2
CLMM results for perceived AI **transparency** across Block 1–3.

Dim = Transparency	Block 1				Block 2				Block 3			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
Stakes (plant)	1.334	0.752	1.774	.076	0.893	0.861	1.037	.300	3.074	1.529	2.011	.044*
Strategy: Accuracy	1.083	0.907	1.194	.233	3.238	1.103	2.934	.003**	3.291	1.615	2.038	.042*
Strategy: Relational Risk	0.150	0.880	0.170	.865	2.545	1.074	2.370	.018*	−0.694	1.478	−0.469	.639
Strategy: Apology	—	—	—	—	—	—	—	—	0.750	1.545	0.485	.628
TiA-PtT	0.694	0.490	1.417	.156	0.719	0.576	1.248	.212	1.029	0.722	1.426	.154
Centred Age	−0.030	0.020	−1.534	.125	−0.005	0.023	−0.200	.842	−0.040	0.029	−1.394	.163
Trust_Block1	—	—	—	—	1.245	0.256	4.856	<.001***	2.124	0.343	6.184	<.001***
Trust_Block2	—	—	—	—	—	—	—	—	1.529	0.306	5.004	<.001***
Stakes × Accuracy	−2.010	1.290	−1.558	.119	−1.326	1.517	−0.874	.382	−6.241	2.304	−2.709	.007**
Stakes × Relational Risk	0.221	1.299	0.170	.865	0.780	1.530	0.510	.610	−0.632	2.080	−0.304	.761
Stakes × Apology	—	—	—	—	—	—	—	—	−4.420	2.191	−2.018	.043*
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	9.847		3.138		16.538		4.067		22.330		4.725	
Item (Intercept)	0.158		0.398		0.051		0.226		0.000		0.000	

* $p < .05$, ** $p < .01$, *** $p < .001$. Em dashes indicate predictors not estimated in that block.

Table A3
CLMM results for perceived AI **performance** across Block 1–3.

Dim = Performance	Block 1				Block 2				Block 3			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
Stakes (plant)	1.057	0.850	1.244	.214	−0.794	0.856	−0.927	.354	1.731	1.236	1.401	.161
Strategy: Accuracy	−0.645	0.994	−0.649	.516	2.626	1.063	2.470	.014*	3.195	1.272	2.512	.012*
Strategy: Relational Risk	−1.774	0.992	−1.788	.074.	0.444	1.039	0.427	.669	−1.525	1.150	−1.326	.185
Strategy: Apology	—	—	—	—	—	—	—	—	0.929	1.214	0.765	.444
TiA-PtT	1.388	0.558	2.488	.013*	1.647	0.565	2.916	.004**	1.252	0.569	2.199	.028*
Centred Age	0.009	0.022	0.392	.695	0.001	0.023	0.032	.974	−0.043	0.022	−1.917	.055.
Trust_Block1	—	—	—	—	0.136	0.295	0.461	.645	1.741	0.339	5.129	$<10^{-6}$***
Trust_Block2	—	—	—	—	—	—	—	—	0.172	0.253	0.681	.496
Stakes × Accuracy	−0.837	1.419	−0.590	.556	−0.712	1.478	−0.482	.630	−6.198	1.828	−3.391	<.001***
Stakes × Relational Risk	0.122	1.419	0.086	.932	0.406	1.466	0.277	.782	0.139	1.669	0.083	.934
Stakes × Apology	—	—	—	—	—	—	—	—	−1.789	1.744	−1.026	.305
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	11.745		3.427		16.433		4.054		13.193		3.632	
Item (Intercept)	0.018		0.135		0.426		0.653		0.270		0.520	

* $p < .05$, ** $p < .01$, *** $p < .001$. Em dashes indicate predictors not estimated in that block. A trailing dot denotes $p < .10$.

Table A4
CLMM results for perceived AI **utility** across Block 1–3.

Dim = Utility	Block 1				Block 2				Block 3			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
Stakes (plant)	0.242	0.456	0.530	.596	−0.205	0.593	−0.345	.730	1.519	0.587	2.587	.010**
Strategy: Accuracy	−0.216	0.546	−0.396	.692	1.323	0.736	1.798	.072.	1.717	0.594	2.892	.004**
Strategy: Relational Risk	−0.643	0.538	−1.197	.231	0.219	0.723	0.302	.762	0.529	0.561	0.944	.345
Strategy: Apology	—	—	—	—	—	—	—	—	1.465	0.585	2.505	.012*
TiA-PtT	0.779	0.299	2.602	.009**	0.946	0.390	2.427	.015*	0.478	0.270	1.773	.077.
Centred Age	0.009	0.012	0.782	.434	−0.013	0.016	−0.797	.425	−0.008	0.011	−0.773	.440
Trust_Block1	—	—	—	—	0.575	0.182	3.164	.002**	1.821	0.179	10.148	<.001***
Trust_Block2	—	—	—	—	—	—	—	—	0.377	0.136	2.770	.006**
Stakes × Accuracy	−0.007	0.779	−0.008	.993	−0.492	1.027	−0.479	.632	−2.512	0.841	−2.986	.003**
Stakes × Relational Risk	0.367	0.774	0.475	.635	−0.025	1.020	−0.024	.980	−0.738	0.816	−0.904	.366
Stakes × Apology	—	—	—	—	—	—	—	—	−2.111	0.832	−2.537	.011*
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	2.882		1.698		7.396		2.720		2.296		1.515	
Item (Intercept)	0.000		0.000		0.166		0.407		0.000		0.000	

* $p < .05$, ** $p < .01$, *** $p < .001$. Em dashes indicate predictors not estimated in that block. A trailing dot denotes $p < .10$.

Data availability

Data will be made available upon request.

References

Armstrong, K., Schwartz, J.S., Fitzgerald, G., Putt, M., Ubel, P.A., 2002. Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. Med. Decis. Mak. 22, 76–83. https://doi.org/10.1177/0272989X0202200108

Ashoori, M., Weisz, J.D., 2019. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. arxiv preprint arxiv:1912.02675.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–16, https://doi.org/10.1145/3411764.3445717, Yokohama Japan. https://dl.acm.org/doi/10.1145/3411764.3445717.

Baranger, D.A.A., Finsaas, M.C., Goldstein, B.L., Vize, C.E., Lynam, D.R., Olino, T.M., 2023. Tutorial: power analyses for interaction effects in cross-sectional regressions. Adv. Methods Pract. Psychol. Sci. 6, https://doi.org/10.1177/25152459231187531

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., 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. ACM, pp. 401–413, https://doi.org/10.1145/3461702.3462571, Virtual Event USA. https://dl.acm.org/doi/10.1145/3461702.3462571.

Bo, J.Y., Wan, S., Anderson, A., 2025. To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–23, https://doi.org/10.1145/3706598.3714097, Yokohama Japan. https://dl.acm.org/doi/10.1145/3706598.3714097.

Breakwell, G.M., 2000. Risk communication: fators affecting impact. Br. Med. Bull. 56, 110–120. https://doi.org/10.1258/0007142001902824

Brown, S.L., Smith, E.Z., 2007. The inhibitory effect of a distressing anti-smoking message on risk perceptions in smokers. Psychology & Health 22, 255–268. https://doi.org/10.1080/14768320600843127

Buçinca, Z., Malaya, M.B., Gajos, K.Z., 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 5, 1–21. https://doi.org/10.1145/3449287. https://dl.acm.org/doi/10.1145/3449287.

Buçinca, Z., Pham, C.M., Jakesch, M., Ribeiro, M.T., Olteanu, A., Amershi, S., 2023. AHA!: Facilitating AI impact assessment by generating examples of harms. arxiv preprint arxiv:2306.03280.

Buçinca, Z., Swaroop, S., Paluch, A.E., Murphy, S.A., Gajos, K.Z., 2024. Towards optimizing human-centric objectives in AI-assisted decision-making with offline reinforcement learning. arxiv preprint arxiv:2403.05911.

Cañas, J.J., 2022. AI and ethics when human beings collaborate with AI agents. Front. Psychol. 13, 836650. https://doi.org/10.3389/fpsyg.2022.836650. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.836650/full.

Cao, S., Liu, A., Huang, C.M., 2024. Designing for appropriate reliance: the roles of AI uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 8, 1–32. https://dl.acm.org/doi/10.1145/3637318.

De Miguel Velázquez, J., Šćepanović, S., Gvirtz, A., Quercia, D., 2024. Decoding real-world artificial intelligence incidents. Computer 57, 71–81. https://doi.org/10.1109/MC.2024.3432492. https://ieeexplore.ieee.org/document/10718657/.

Derczynski, L., Kirk, H.R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M.R., Mohammad, S., 2023. Assessing language model deployment with risk cards. arxiv prpeprint arxiv:2303.18190.

Dunlop, S., Wakefield, M., Kashima, Y., 2008. Can you feel it? Negative emotion, risk, and narrative in health communication. Media Psychology 11, 52–75. https://doi.org/10.1080/15213260701853112. http://www.tandfonline.com/doi/abs/10.1080/15213260701853112.

Fahnenstich, H., Rieger, T., Roesler, E., 2024. Trusting under risk – comparing human to AI decision support agents. Comput. Hum. Behav. 153, 108107. https://doi.org/10.1016/j.chb.2023.108107. https://linkinghub.elsevier.com/retrieve/pii/S0747563223004582.

Field, A., Miles, J., Field, Z. ed), 2014. Discovering Statistics Using R, Repr ed. Sage, Los Angeles, CA, USA.

Gong, J., Zhang, Y., Yang, Z., Huang, Y., Feng, J., Zhang, W., 2013. The framing effect in medical decision-making: a review of the literature. Psychology, Health & Medicine 18, 645–653. https://doi.org/10.1080/13548506.2013.766352. https://www.tandfonline.com/doi/full/10.1080/13548506.2013.766352.

He, G., Buijsman, S., Gadiraju, U., 2023. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. Proc. ACM Hum.-Comput. Interact. 7, 1–29. https://doi.org/10.1145/3610067. https://dl.acm.org/doi/10.1145/3610067.

Herdel, V., Šćepanović, S., Bogucka, E., Quercia, D., 2024. ExploreGen: large language models for envisioning the uses and risks of AI technologies. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7, 584–596. https://doi.org/10.1609/aies.v7i1.31660.

Hoesterey, S., Onnasch, L., 2023. The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. Cogn. Technol. Work 25, 15–29. https://doi.org/10.1007/s10111-022-00718-y. https://link.springer.com/10.1007/s10111-022-00718-y.

Hoff, K.A., Bashir, M., 2015. Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors: J. Hum. Factors Ergon. Soc. 57, 407–434. https://doi.org/10.1177/0018720814547570. https://journals.sagepub.com/doi/10.1177/0018720814547570.

Kahneman, D., 2011. Thinking, Fast and Slow. Allen Lane and Penguin Books, New York.

Kahr, P.K., Rooks, G., Snijders, C., Willemsen, M.C., 2024. The trust recovery Journey. The effect of timing of errors on the willingness to follow AI advice. In: Proceedings of the 29th International Conference on Intelligent User Interfaces. ACM, pp. 609–622, https://doi.org/10.1145/3640543.3645167, Greenville SC USA. https://dl.acm.org/doi/10.1145/3640543.3645167.

Kim, T., Song, H., 2021. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. Telemat. Inform. 61, 101595. https://doi.org/10.1016/j.tele.2021.101595. https://linkinghub.elsevier.com/retrieve/pii/S0736585321000344.

Kim, T., Song, H., 2023. “I believe AI can learn from the Error. Or can it not?”: The effects of implicit theories on trust repair of the intelligent agent. Int. J. Soc. Robot. 15, 115–128. https://doi.org/10.1007/s12369-022-00951-5. https://link.springer.com/10.1007/s12369-022-00951-5.

Kim, S.S.Y., Liao, Q.V., Vorvoreanu, M., Ballard, S., Vaughan, J.W., 2024. I’m not sure, But..": Examining the impact of large language models’ uncertainty expression on user reliance and trust. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA. pp. 822–835, https://doi.org/10.1145/3630106.3658941

Kocielnik, R., Amershi, S., Bennett, P.N., 2019. Will you accept an imperfect AI?: Exploring designs for adjusting end-user expectations of AI systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk. pp. 1–14, https://doi.org/10.1145/3290605.3300641

Kohn, S.C., De Visser, E.J., Wiese, E., Lee, Y.C., Shaw, T.H., 2021. Measurement of trust in automation: a narrative review and reference guide. Front. Psychol. 12, 604977. https://doi.org/10.3389/fpsyg.2021.604977. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.604977/full.

Körber, M., 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). Springer International Publishing, Cham, pp. 13–30, https://doi.org/10.1007/978-3-319-96074-6_2

Kox, E.S., Kerstholt, J.H., Hueting, T.F., De Vries, P.W., 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Auton. Agents Multi-Agent Syst. 35, 30. https://doi.org/10.1007/s10458-021-09515-9. https://link.springer.com/10.1007/s10458-021-09515-9.

Kwon, M., Huang, S.H., Dragan, A.D., 2018. Expressing robot incapability. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. ACM, Chicago IL USA. pp. 87–95, https://doi.org/10.1145/3171221.3171276. https://dl.acm.org/doi/10.1145/3171221.3171276.

Lai, V., Tan, C., 2019. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, pp. 29–38, https://doi.org/10.1145/3287560.3287590, Atlanta GA USA.

LeClerc, J., Joslyn, S., 2015. The cry wolf effect and weather-related decision making. Risk Anal. 35, 385–395. https://doi.org/10.1111/risa.12336

Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. Hum. Factors 46, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, B., Cordeil, M., Prouzeau, A., Jenny, B., Dwyer, T., 2022. A design space for data visualisation transformations between 2D and 3D in mixed-reality environments. In: CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–14, https://doi.org/10.1145/3491102.3501859, New Orleans LA USA. https://dl.acm.org/doi/10.1145/3491102.3501859.

LeMasurier, G., Gautam, A., Han, Z., Crandall, J.W., Yanco, H.A., 2024. Reactive or proactive? How robots should explain failures. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp. 413–422, https://doi.org/10.1145/3610977.3634963, Boulder CO USA.

Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G.S., Peng, L.H., Webster, D.R., Ai, D., Huang, S.J., Liu, Y., Dunn, R.C., Coz, D., 2020. A deep learning system for differential diagnosis of skin diseases. Nat. Med. 26, 900–908. https://doi.org/10.1038/s41591-020-0842-3

Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An integrative model of organizational trust. Acad. Manag. Rev. https://doi.org/10.5465/amr.1995.9508080335

McGregor, S., 2021. Preventing repeated real world AI failures by cataloging incidents: the AI incident database. Proc. AAAI Conf. Artif. Intell. 35, 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817. number: 17.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 22, 276–282. https://doi.org/10.11613/BM.2012.031

McNeil, B.J., Pauker, S.G., Sox, H.C., Tversky, A., 1982. On the elicitation of preferences for alternative therapies. N. Engl. J. Med. 306, 1259–1262. https://doi.org/10.1056/NEJM198205273062103

Miller, T., 2023. Explainable AI is dead, long live explainable AI!: Hypothesis-driven decision support using evaluative AI. In: 2023 ACM Conference on Fairness Accountability and Transparency. ACM, pp. 333–342, https://doi.org/10.1145/3593013.3594001, Chicago IL USA. https://dl.acm.org/doi/10.1145/3593013.3594001.

Nesset, B., Romeo, M., Rajendran, G., Hastie, H., 2023. Robot broken promise? Repair strategies for mitigating loss of trust for repeated failures. In: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, pp. 1389–1395, https://doi.org/10.1109/RO-MAN57019.2023.10309558, Busan, Korea, Republic of. https://ieeexplore.ieee.org/document/10309558/.

Papenmeier, A., Englebienne, G., Seifert, C., 2019. How model accuracy and explanation fidelity influence user trust. arxiv preprint arxiv:1907.12652.

Pareek, S., Velloso, E., Goncalves, J., 2024. Trust development and repair in AI-assisted decision-making during complementary expertise. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. ACM, pp. 546–561, https://doi.org/10.1145/3630106.3658924, Rio de Janeiro Brazil.

Pareek, S., Schömbs, S., Velloso, E., Goncalves, J., 2025. ”It’s not the AI’s fault because it relies purely on data”: how causal attributions of AI decisions shape trust in AI systems. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp. 1–18, https://doi.org/10.1145/3706598.3713468, New York, NY, USA. https://dl.acm.org/doi/10.1145/3706598.3713468.

Park, J.S., Barber, R., Kirlik, A., Karahalios, K., 2019. A slow algorithm improves users’ assessments of the algorithm’s accuracy. Proc. ACM Hum.-Comput. Interact. 3, 1–15. https://doi.org/10.1145/3359204

Passi, S., Vorvoreanu, M., 2022. Overreliance on AI: Literature Review. Technical Report MSR-TR-2022-12. Microsoft. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/.

Peters, E., Hart, P.S., Fraenkel, L., 2011. Informing patients: the influence of numeracy, framing, and format of side effect information on risk perceptions. Med. Decis. Mak. 31, 432–436. https://doi.org/10.1177/0272989X10391672

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q.V., Banovic, N., 2023. Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in Human-AI decision making. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. ACM, pp. 379–396, https://doi.org/10.1145/3581641.3584033, Sydney NSW Australia.

Rao, P.S.B., Šćepanović, S., Zhou, K., Bogucka, E.P., Quercia, D., 2025. RiskRAG: a data-driven solution for improved AI model risk reporting. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–26, https://doi.org/10.1145/3706598.3713979, Yokohama Japan.

Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998. Applied Regression Analysis: A Research Tool, second ed. Springer Texts in Statistics, Springer, New York.

Rechkemmer, A., Yin, M., 2022. When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models. In: CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–14, https://doi.org/10.1145/3491102.3501967, New Orleans LA USA. https://dl.acm.org/doi/10.1145/3491102.3501967.

Rieger, T., Kugler, L., Manzey, D., Roesler, E., 2024. The (im)perfect automation schema: who is trusted more, automated or human decision support? Hum. Factors: J. Hum. Factors Ergon. Soc. 66, 1995–2007. https://doi.org/10.1177/00187208231197347

Rieger, T., Schindler, H., Onnasch, L., Roesler, E., 2025. Explaining AI weaknesses improves human–AI performance in a dynamic control task. Int. J. Hum.-Comput. Stud. 199, 103505. https://doi.org/10.1016/j.ijhcs.2025.103505

Roesler, E., Rieger, T., Manzey, D., 2022. Trust towards human vs. automated agents: using a multidimensional trust questionnaire to assess the role of performance, utility, purpose, and transparency. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 66, 2047–2051. https://doi.org/10.1177/1071181322661065

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M.C., Langer, M., 2025a. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). Comput. Hum. Behav. 170, 108671. https://doi.org/10.1016/j.chb.2025.108671

Schlicker, N., Lechner, F., Wehrle, K., Greulich, B., Hirsch, M.C., Langer, M., 2025b. Trustworthy enough? Examining trustworthiness assessments of large language model-based medical agents. Technology, Mind, and Behavior 6, 144–172. https://doi.org/10.1037/tmb0000164

Schömbs, S., Pareek, S., Goncalves, J., Johal, W., 2024. Robot-assisted decision-making: unveiling the role of uncertainty visualisation and embodiment. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp. 1–16, https://doi.org/10.1145/3613904.3642911, New York, NY, USA. https://dl.acm.org/doi/10.1145/3613904.3642911.

Schömbs, S., Goncalves, J., Johal, W., 2025. "I can feel the risks by looking at the robot face": communicating risk through a physical agent. In: Proceedings of the 2025 ACM Designing Interactive Systems Conference. Association for Computing Machinery, pp. 236–252, https://doi.org/10.1145/3715336.3735759, New York, NY, USA. https://dl.acm.org/doi/10.1145/3715336.3735759.

Sellers, R., 2013. How sliders bias survey data. MRA's Alert 53, 56–57.

Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y.H., Qiwei, L., Rakshit, S., Si, C., Xie, Y., Bigham, J.P., Bentley, F., Chai, J., Lipton, Z., Mei, Q., Mihalcea, R., Terry, M., Yang, D., Morris, M.R., Resnick, P., Jurgens, D., 2024. Towards bidirectional Human-AI alignment: a systematic review for clarifications, framework, and future directions. arxiv preprint arxiv:2406.09264.

Slattery, P., Saeri, A.K., Grundy, E.A.C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., Thompson, N., 2025. The AI risk repository: a comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. arxiv preprint arxiv:2408.12622.

Spiegelhalter, D., 2017. Risk and uncertainty communication. Annu. Rev. Stat. Appl. 4, 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148

Stuck, R.E., Tomlinson, B.J., Walker, B.N., 2022. The importance of incorporating risk into human-automation trust. Theor. Issues Ergon. Sci. 23, 500–516. https://doi.org/10.1080/1463922X.2021.1975170

Tabatabaei, R., Kostakos, V., Johal, W., 2026. Oops, I did it again (but I know it): robot failure consistency and awareness in human-robot collaboration. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–18, https://doi.org/10.1145/3772318.3790336, Barcelona Spain. https://dl.acm.org/doi/10.1145/3772318.3790336.

Thomas, D.R., 2006. A general inductive approach for analyzing qualitative evaluation data. Am. J. Eval. 27, 237–246. https://doi.org/10.1177/1098214005283748

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T.M., Dixon, C., Tielman, M.L., 2020. Taxonomy of trust-relevant failures and mitigation strategies. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp. 3–12, https://doi.org/10.1145/3319502.3374793, Cambridge United Kingdom.

Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., Bernstein, A., 2021. Second chance for a first impression? Trust development in intelligent system interaction. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. ACM, pp. 77–87, https://doi.org/10.1145/3450613.3456817, Utrecht Netherlands. https://dl.acm.org/doi/10.1145/3450613.3456817.

Victor Syiem, B., Velloso, E., 2026. Better assumptions, stronger conclusions: the case for ordinal regression in HCI. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–21, https://doi.org/10.1145/3772318.3790821, Barcelona Spain. https://dl.acm.org/doi/10.1145/3772318.3790821.

Visschers, V.H.M., Meertens, R.M., Passchier, W.W.F., De Vries, N.N.K., 2009. Probability information in risk communication: a review of the research literature. Risk Anal. 29, 267–287. https://doi.org/10.1111/j.1539-6924.2008.01137.x

Visschers, V., Wiedemann, P., Gutscher, H., Kurzenhäuser, S., Seidl, R., Jardine, C., Timmermans, D., 2012. Affect-inducing risk communication: current knowledge and future directions. J. Risk Res. 15, 257–271. https://doi.org/10.1080/13669877.2011.634521

Walpole, H.D., Wilson, R.S., 2021. Extending a broadly applicable measure of risk perception: the case for susceptibility. J. Risk Res. 24, 135–147. https://doi.org/10.1080/13669877.2020.1749874

Wang, X., Yin, M., 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In: 26th International Conference on Intelligent User Interfaces. ACM, pp. 318–328, https://doi.org/10.1145/3397481.3450650, College Station TX USA. https://dl.acm.org/doi/10.1145/3397481.3450650.

Wang, X., Yin, M., 2022. Effects of explanations in AI-assisted decision making: principles and comparisons. ACM Trans. Interact. Intell. Syst. 12, 1–36. https://doi.org/10.1145/3519266

Wang, Z.J., Kulkarni, C., Wilcox, L., Terry, M., Madaio, M., 2024. Farsight: fostering responsible AI awareness during AI application prototyping. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp. 1–40, https://doi.org/10.1145/3613904.3642335, New York, NY, USA. https://dl.acm.org/doi/10.1145/3613904.3642335.

Wickens, C.D., Rice, S., Keller, D., Hutchins, S., Hughes, J., Clayton, K., 2009. False alerts in air traffic control conflict alerting system: is there a "Cry wolf" effect? Hum. Factors: J. Hum. Factors Ergon. Soc. 51, 446–462. https://doi.org/10.1177/0018720809344720

Wilson, R.S., Zwickle, A., Walpole, H., 2019. Developing a broadly applicable measure of risk perception. Risk Anal. 39, 777–791. https://doi.org/10.1111/risa.13207

Yin, M., Wortman Vaughan, J., Wallach, H., 2019. Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–12, https://doi.org/10.1145/3290605.3300509, Glasgow Scotland Uk. https://dl.acm.org/doi/10.1145/3290605.3300509.

Zhang, Y., Liao, Q.V., Bellamy, R.K.E., 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. Association for Computing Machinery, pp. 295–305, https://doi.org/10.1145/3351095.3372852, New York, NY, USA.