

Asymmetric Trust Effects of Corrective AI in Expert Advisory Workflows under Epistemic Dependence

Dennis Kim, Roya Daneshi, Bruce Draper, Sarath Sreedharan

Colorado State University

d.kim@colostate.edu, roya.daneshi@colostate.edu, Bruce.Draper@colostate.edu, sarath.sreedharan@colostate.edu

Abstract

The increasing integration of AI-powered tools into expert workflows, such as medicine, law, and finance, raises a critical question: how does AI involvement influence a user's trust in the human expert, the AI system, and the human-AI team? This question is especially important in expert advisory settings where users are epistemically dependent on human-AI systems: they are recipients of guidance produced by an expert using AI support, but often lack the domain knowledge needed to independently verify the recommendation. We investigated these dynamics through a user study ($N = 157$) using a simulated course-planning task. Our design varied advisor performance and the structure of AI involvement, including whether AI support was present and, when present, whether it was invoked by the advisor or automatically monitored the interaction. Across all conditions, workflows ultimately produced correct schedules. Results show an asymmetric trust effect: advisor errors significantly reduce trust in the human advisor, but visible AI correction does not produce a corresponding increase in trust toward the AI assistant. Trust judgments remain anchored to the advisor across multiple trust measures, and changing the visible structure of AI involvement does not substantially redistribute trust toward the AI assistant or the human-AI team. These findings suggest a limit of corrective AI as a governance mechanism: making AI oversight visible may improve workflow recoverability, but it does not necessarily redistribute trust or responsibility away from the human-facing expert. In epistemically dependent settings, correctness alone may be insufficient for trustworthy AI integration, as users may continue to assign responsibility to the human expert even when AI assistance visibly shapes the final outcome.

1 Introduction

Across domains such as healthcare, financial advising, legal consultation, and others, people depend on human experts to help them navigate complex decisions (Agrawal et al. 2023; Han and Ko 2025; Mayer, Davis, and Schoorman 1995; Kaye and Hamilton 2004). These interactions depend on trust in experts (Hendriks, Kienhues, and Bromme 2015): an evaluative judgment that an advisor will help achieve a user's goals under conditions of uncertainty (Lee and See 2004), which in practice often manifests in reliance behaviors such as reuse (Liao and Sundar 2022). Understanding

how trust is shaped within human-AI workflows is therefore critical for interpreting user perceptions and anticipating downstream reliance behavior.

As AI systems increasingly enter expert advisory workflows, users may attribute guidance to the human advisor, the AI system, or the human-AI workflow as a whole. This creates a puzzle for corrective AI: when AI helps identify or resolve an expert error, does it earn trust as a reliable contributor, or does its intervention primarily expose the human expert's fallibility? Even before a specific intervention occurs, the presence of AI may also shape how users interpret credibility within the advisory process (Nakano et al. 2025). The implications of this shift in advisory contexts remain unclear, and understanding how responsibility and epistemic authority are allocated within such human-AI advisory workflows is central to broader questions of accountability and governance in sociotechnical systems. The "user" refers to the recipient of guidance (e.g., a patient), who does not directly interact with the AI but evaluates the outcomes of a human-AI workflow. As such, the user serves as the trustor, forming judgments about the advisor, the AI system, and their combined behavior.

Prior work has examined trust in algorithmic systems (Lee and See 2004; Lee 2018) and human experts (Hendriks, Kienhues, and Bromme 2015), but less is known about how users distribute trust across human and AI contributors in precision-dependent tasks where both shape the final guidance. This gap is especially important when AI visibly corrects or safeguards expert recommendations, because such interventions may change how users assign credibility and responsibility within the advisory workflow.

From an epistemic perspective, trust in experts can be decomposed into expertise, integrity, and benevolence (Hendriks, Kienhues, and Bromme 2015), with advisor performance informing perceived expertise (Tenney, Spellman, and MacCoun 2008) and AI oversight potentially shaping integrity- and benevolence-related inferences. This multidimensional approach remains well-justified in trust measurement (Kim 2024) and is especially relevant for understanding trust in expert AI contexts (Asan, Bayrak, and Choudhury 2020), where AI integration is accelerating (Kaffee, Atanasova, and Rogers 2025; Moëll, Aronsson, and Akbar 2025). In this work, we experimentally manipulate two factors: advisor performance (whether the advisor fails to de-

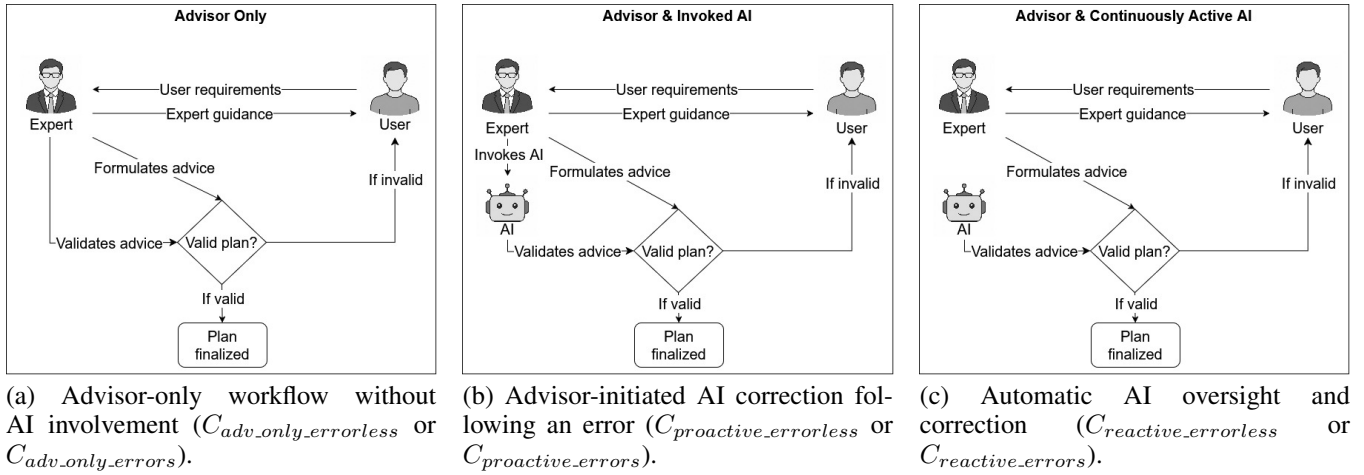


Figure 1: Illustrative advising workflows studied, highlighting differences in advisor performance and AI intervention modality.

tect an invalid course selection) and AI oversight modality (advisor-only with no AI involvement, reactive AI oversight that automatically monitors and flags issues, or proactive AI oversight invoked at the human advisor’s discretion). We examine whether and how these factors influence multidimensional trust judgments, entity-specific trust, and downstream behavioral intentions in expert-guided interactions. As we show, these effects are not uniformly distributed across system components, but instead are concentrated in trust judgments directed at the human advisor.

We leveraged an online academic advising simulation as a simplified, representative task to study this phenomenon. Throughout this paper, we refer to the human expert as the advisor (or simply the expert), following the academic advising scenario used in our study. Figure 1 shows the interaction structures considered in our study. Academic advising offers a useful environment for studying these dynamics, since advisors are expected to have expertise and specialized knowledge about courses as well as departmental and university degree requirements (EDUTECHLoft 2025). At the same time, while perceived risk is a necessary condition of trust (Mayer, Davis, and Schoorman 1995), the stakes of advising are moderate enough to limit confounding factors present in higher-risk domains such as medicine or finance, enabling a more controlled examination of trust dynamics. We therefore frame this study, in part, as a mechanism-focused investigation of trust attribution under AI oversight rather than as a direct claim about trust dynamics in all expert domains. Moreover, the scenario naturally supports AI assistants that can run in both a proactive and a reactive mode (Section 4.1 provides a full overview of the experiment setup). Each measurement draws on previously validated trust instruments, as described in Section 4.3.

Our results show that advisor performance significantly shapes trust-related evaluations: error-present conditions were evaluated less favorably across measures of epistemic trustworthiness, entity-specific trust, general trust, and likelihood of advisor reuse, even though errors were ultimately corrected in all conditions, either by the AI system or the

advisor. In contrast, AI intervention modality did not produce significant effects, suggesting that visible AI involvement did not substantially redistribute trust away from the human advisor. Descriptively, performance-related declines were largest in advisor-only conditions, followed by proactive AI use, and smallest under reactive AI oversight. Trust in the AI system remained comparatively stable, with mixed results for the human-AI team.

Taken together, our results suggest that corrective AI may change how an advisory workflow reaches a valid outcome without changing who users treat as epistemically responsible for that outcome.

1.1 Contributions

To summarize, the paper makes three contributions:

1. We present an interactive experimental paradigm for studying trust attribution in AI-assisted advisory workflows under controlled variations in advisor performance and AI integration structure.
2. Using this academic advising setup, we show that observed advisor errors reduced trust in the human advisor, while visible AI correction did not increase trust in the AI assistant despite its corrective role. AI oversight also did not significantly redistribute trust consequences across the advisor, AI assistant, and advisor-AI team, suggesting that epistemic responsibility remained anchored to the human-facing expert. Descriptive patterns suggest that AI involvement may attenuate trust penalties, with heterogeneous responses to advisor mistakes.
3. We derive design implications for human-AI advisory workflows, emphasizing the importance of advisor competence, real-world constraints on error visibility, the positioning of AI assistance in error detection, and the potential for complementary signals to support adaptive responses to user trust.

2 Hypotheses

We test four hypotheses motivated by prior work on trust in automation (Lee and See 2004; Dzindolet et al. 2003; Hoff and Bashir 2015) and our directional expectations about how expert missteps and recovery influence trust. Conditions are defined by AI modality (none, proactive, reactive) and advisor performance (errorless, error-present), yielding six conditions: $C_{adv_only_errorless}$, $C_{adv_only_errors}$, $C_{proactive_errorless}$, $C_{proactive_errors}$, $C_{reactive_errorless}$, and $C_{reactive_errors}$ (see Section 4.1).

In all conditions, the student (participant) is given a course scheduling task (i.e., preparing a course schedule for a new semester), specifies desired courses, and receives a valid schedule. In the no-AI condition, the advisor completes the task unaided (Figure 1(a)). In proactive conditions, the advisor invokes the AI to verify the scheduling choice, and errors, if present, are identified during verification (Figure 1(b)). In reactive conditions, the AI automatically intervenes and immediately flags errors (Figure 1(c)). When no error occurs, the interaction proceeds without an error-correction event. An advisor error refers to the advisor's initial failure to detect an invalid course selection, all such errors were corrected before the final schedule was produced.

Dependent variables captured four trust-related outcomes: epistemic trustworthiness assessed via the Hendriks, Kienhues, and Bromme (2015) questionnaire; entity-level trust in the advisor, AI assistant, and human-AI team adapted from Riedl, Hogeterp, and Reuter (2024); global trust in the advisor on a 100-point scale; and willingness to reuse the advisor on a 7-point Likert scale ("Extremely unlikely"–"Extremely likely"). All hypotheses concern how advisor performance and AI intervention modality influence these evaluations.

Prior work on trust in automation suggests that observed failures prompt trust recalibration (Dzindolet et al. 2003). Accordingly, we expect advisor performance (error-present vs errorless) to influence trust-related evaluations of the advisor (H1). We further expect that AI intervention modality (none, proactive, reactive) will influence how trust-related evaluations are allocated across the advisor, AI assistant, and advisor-AI team (H2), consistent with frameworks emphasizing how agency and decision authority are distributed in human-automation workflows (Parasuraman, Sheridan, and Wickens 2000). In this framework, differences in who initiates action (e.g., monitoring and verification) reflect different allocations of decision authority, which may influence how observers interpret an advisor's epistemic responsibility even when task outcomes are identical. Because users can evaluate human, AI, and hybrid trust targets separately (Riedl, Hogeterp, and Reuter 2024), we expected visible AI verification to shift some trust credit from the advisor toward the AI assistant or advisor-AI team. This shift may be strongest under reactive oversight, where monitoring occurs independently rather than through advisor invocation. We also expected AI modality to shape the trust penalty associated with advisor errors (H3), because communicated feedback and recovery can affect trust responses (Dzindolet et al. 2003; Hoff and Bashir 2015). Whereas an unaided error primarily signals advisor fallibility, AI correction may signal that the workflow contains safeguards. Advisor-invoked

verification may indicate responsible tool use, while automatic monitoring may indicate recovery independent of the advisor recognizing the need for assistance. We therefore expected AI oversight, particularly reactive oversight, to attenuate the error-related trust decline relative to advisor-only workflows. Finally, we expected advisor errors to reduce willingness to reuse the advisor (H4).

2.1 Hypotheses List

To summarize, the hypotheses we tested are as follows:

- H1: Advisor errors will significantly negatively affect trust evaluations of the advisor, including epistemic trustworthiness measured using the Hendriks, Kienhues, and Bromme (2015) instrument and entity-specific trust adapted from Riedl, Hogeterp, and Reuter (2024).
- H2: AI intervention modality (none, proactive, reactive) will significantly affect how trust-related evaluations are distributed across available trust targets, with AI-supported workflows expected to produce higher evaluations of the AI assistant and advisor-AI team when those targets are present.
- H3: The trust penalty from advisor errors will differ significantly depending on the AI intervention modality.
- H4: Participants will be significantly less likely to reuse advisors who made errors than those who did not, consistent with prior work linking trust to downstream reliance.

As reported below and in Section 5, H1 and H4 were supported: advisor performance significantly influenced advisor-directed trust evaluations across multiple measures, including epistemic trustworthiness (Hendriks, Kienhues, and Bromme 2015), entity-specific trust (Riedl, Hogeterp, and Reuter 2024), general trust (Lee and See 2004), and willingness to reuse (Liao and Sundar 2022). H2 and H3 were not supported: AI intervention modality did not produce significant main effects, nor did it significantly moderate the trust penalty associated with advisor errors.

We treat these deviations from the a priori expectations as theoretically informative, suggesting that observed advisor error played a more dominant role than AI involvement strategy in shaping trust and reuse judgments.

3 Relevant Work

Research on trust in automation, human-AI interaction, and social epistemology provides complementary perspectives motivating our study design and trust measurement strategy. We organize prior work around three questions central to our investigation: (i) how trust responds dynamically to observed performance and error, (ii) how procedural structure and responsibility cues shape trust in AI-supported expert workflows, and (iii) how trust should be measured when multiple potential trustees are present in a single interaction.

3.1 Trust Calibration Under Epistemic Dependence

A central theme in the trust-in-automation literature is trust recalibration: users update their trust judgments based on observed system behavior and outcomes (Lee and See

2004). Here, trust recalibration refers to differences in reported trust between error-present and errorless conditions following a single between-subjects interaction. This process is especially important under epistemic dependence, where users cannot fully understand or independently verify system correctness and must instead rely on observable cues such as performance, communication, and interaction dynamics to calibrate trust (Lee and See 2004). Recent work in human-AI interaction further formalizes trust in terms of user vulnerability and expectations over system behavior (Jacovi et al. 2021), while work on trust dynamics emphasizes that trust judgments emerge over interaction rather than from isolated assessments (Han and Ko 2025).

Empirical evidence reinforces that calibration operates at the level of individual interactions and outputs. For example, Kim et al. (2023) show that users do not rely on AI systems with a fixed level of trust, but instead selectively accept or reject outputs after engaging in verification behaviors, highlighting the dynamic and context-sensitive nature of trust. Prior syntheses similarly emphasize that trust judgments depend on contextual factors such as task demands, perceived risk, and user expectations, and may diverge from objective performance metrics (Hoff and Bashir 2015; Wong et al. 2025). Work on AI-assisted ethical decision-making further shows that reliance on AI advice can vary with properties of the recommendation itself, including whether it aligns with the decision-maker's values (Narayanan et al. 2023). Across domains, observed errors play a particularly important role in this process, as they provide salient evidence that can prompt users to update their beliefs about system reliability and competence. For example, work on algorithmic aversion shows that exposure to system errors can reduce subsequent reliance after users observe imperfections (Dietvorst, Simmons, and Massey 2015).

Importantly, trust responses depend not only on error occurrence but also on how errors are communicated and contextualized. Dzindolet et al. (2003) show that feedback about system limitations and error characteristics enables more appropriate calibration, helping users avoid both overreliance and disuse. Work on AI communication similarly shows that factuality-expression formats can affect user trust in AI-generated answers (Do and Geyer 2025), while uncertainty-expression cues can modulate reliance by signaling limitations in system outputs (Kim et al. 2024). Unlike settings where an AI system directly communicates reliability or uncertainty information to users, our study examines AI assistance embedded within a human-facing expert workflow. In this setting, corrective AI may provide evidence of workflow-level recovery while also exposing the advisor's need for correction, without necessarily making the AI assistant itself the primary target of trust. This motivates examining outcomes alongside how performance information is surfaced, interpreted, and attributed within an interaction.

Recent studies in AI-assisted advisory contexts further characterize trust as path-dependent and sensitive to sequences of interaction events. Han and Ko (2025) show that salient errors can produce sharp declines in trust that may be partially repaired through subsequent corrective interactions, while Pareek, Velloso, and Goncalves (2024) demon-

strate that trust can propagate across stages of a workflow when users are able to evaluate some components but not others. These findings highlight that trust judgments are formed over trajectories of interaction, where both errors and corrective signals contribute to how responsibility and reliability are inferred over time.

Together, these studies position trust calibration as an interaction-level process shaped by observed outcomes, communicated performance information, corrective signals, and the user's ability to evaluate what happened. This is especially important in epistemically dependent advisory settings. Building on this literature, our work examines whether trust judgments are shaped not only by final outcomes, but by whether the advisory workflow visibly encountered and resolved problems along the way, and by how AI support is incorporated into that sequence.

3.2 AI Oversight, Workflow Governance, and Responsibility Attribution

Beyond performance outcomes, trust judgments are also shaped by how users interpret epistemic roles and responsibility within a workflow. In human-AI settings, identical outcomes may support different inferences depending on who appears to monitor, verify, and authorize decisions. Yang and Ma (2025) argue that users' evaluations of AI systems and AI-assisted decision-making contexts involve context-dependent assessments of epistemic authority, responsibility, system behavior, and underlying processes. A parallel distinction appears in formal work on AI trustworthiness: Bisconti et al. (2024) distinguish the characteristics a system possesses from the characteristics observers infer from displayed behavior, and argue that perceived trustworthiness can depend on cues such as transparency, agency locus, and human oversight. In our setting, these process-based assessments are reflected through procedural cues in the interaction (e.g., how AI intervention occurs).

Classic models of human-automation interaction similarly emphasize the importance of procedural visibility. Parasuraman, Sheridan, and Wickens (2000) note that higher levels of automation can obscure ongoing human involvement, reducing users' ability to distinguish between human and system contributions. Such shifts do not inherently undermine trust, but they may alter how users attribute responsibility for errors and corrections.

How AI involvement is disclosed may also shape attribution. Progressive disclosure reveals explanations gradually and on demand, with their order, modality, and complexity affecting perceived transparency and understanding (Muralidhar et al. 2025). Thus, trust attribution may depend not only on whether AI correction is visible, but also on when and how its process is explained.

Social epistemology suggests that trust is supported by mechanisms that evaluate an informant's competence and responsiveness to error rather than by passive acceptance alone (Sperber et al. 2010; Origi 2014). Observable signals of monitoring and correction can therefore function as procedural evidence of epistemic responsibility. Related work on calibration and credibility judgments shows that errors are particularly damaging when interpreted as evidence of

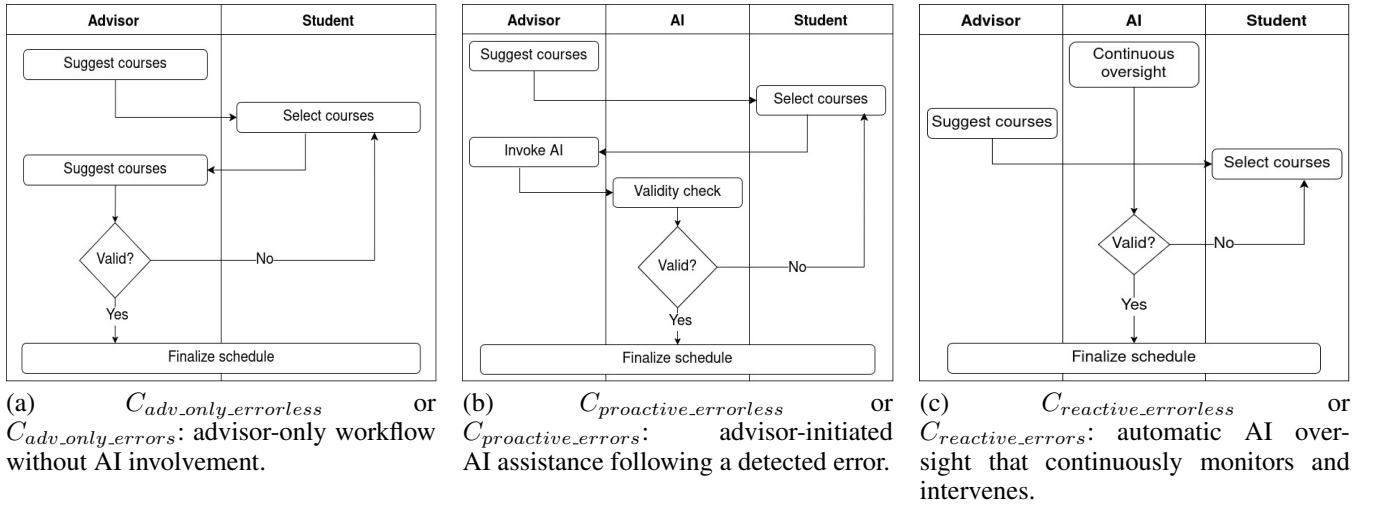


Figure 2: Interaction workflows for the academic advising paradigms examined in this study.

poor self-assessment rather than as isolated outcomes (Kim et al. 2004; Tenney, Spellman, and MacCoun 2008; Griffin and Tversky 1992). Together, these perspectives motivate treating AI oversight as part of advisory workflow governance rather than as a purely technical aid. Building on this, we examine whether visible AI correction in a human-facing advisory workflow changes how trust consequences are allocated across the advisor, the AI assistant, and their combined operation, even when the final recommendation is correct.

3.3 Trust, Reliance, and Accountability Across Multiple Targets

A methodological challenge in AI-supported advisory systems is that a single interaction may involve multiple plausible trustees. Users may evaluate the human advisor, the AI system, and their combined operation differently, and outcome evidence may not map cleanly onto a single target.

Hendriks, Kienhues, and Bromme (2015) conceptualize epistemic trustworthiness as a multidimensional construct comprising expertise, integrity, and benevolence, providing a validated framework for assessing trust in experts under epistemic uncertainty. However, this alone misses how trust distributes across multiple entities when AI is involved. Prior work (Riedl, Hogeterp, and Reuter 2024) demonstrates that users can meaningfully provide entity-specific trust judgments for human professionals, AI systems, and hybrid configurations using parallel item structures.

This multi-target measurement strategy is central to our study. By separately measuring trust in the advisor, the AI assistant, and the advisor-AI team, we can test whether trust penalties and credit are allocated across specific components of the workflow or remain concentrated on the human expert despite visible AI involvement. We use trust consequences to mean condition-related differences in trust and the entities toward which they are directed. Advisor-directed consequences were assessed through epistemic trustworthiness, advisor-specific trust, and general trust; AI- and team-directed consequences through entity-specific ratings in AI-

supported conditions. Advisor reuse was analyzed separately as a behavioral proxy for anticipated reliance.

4 Methodology

4.1 Study Design

We examine how trust judgments form in response to realistic advising sequences and AI involvement. Each participant received a distinct, randomly-assigned workflow varying along two factors: advisor performance and AI intervention modality, to assess their effects and interaction.

Participants engaged in a simulated academic advising interaction (Fig. 2) presented as a sequence of screenshots and system messages within a web-based interface (Qualtrics 2025) (see the technical appendix of the extended version (Kim et al. 2026)). They were informed that they would complete a course scheduling task (with or without AI) and that the study examined perceptions of this interaction. The AI, when present, operated in an oversight tool. When errors occurred, they were observable: with AI, they were surfaced through messages and visual text feedback indicating invalid course selections; without AI, the advisor identified them by cross-referencing a rule book and addressed them in the chat. In all conditions, participants assumed the role of the student and were instructed to follow the guidance. They were never provided means to verify their choices independently.

A between-subjects design was used to avoid carryover and learning effects, as trust judgments recalibrate based on prior performance (Lee and See 2004). Each participant completed the task once. Post-task, participants evaluated the advisor, while those in AI-supported conditions also evaluated the AI assistant and advisor-AI team.

An a priori G*Power analysis of the advisor-performance $\times$ AI-modality interaction showed that the 3×2 design required $N = 158$ to detect $f = .25$ with $\alpha = .05$ and 80% power; the final $N = 157$ provided approximately 79.9% power. Under the same assumptions, the restricted 2×2

analyses required $N = 128$; the available AI-supported subsample of $N = 109$ provided approximately 73.4% power.

The experimental interactions were as follows (illustrative workflows available in Fig. 2):

C_{adv_only_errorless}: Advisor-only, errorless condition. The student selects courses that satisfy the scheduling requirements, and the advising workflow proceeds without AI involvement.

C_{adv_only_errors}: Advisor-only, error-present condition. The student selects an invalid course, which the advisor detects only after checking the rule book and corrects without AI involvement.

C_{proactive_errorless}: Advisor-invoked AI, errorless condition. The student selects courses that satisfy the scheduling requirements within a workflow where AI support is available through advisor invocation.

C_{proactive_errors}: Advisor-invoked AI, error-present condition. The student selects an invalid course that the advisor does not initially detect; advisor-invoked AI assistance identifies the error so it can be corrected before the final schedule is produced.

C_{reactive_errorless}: Automatic AI oversight, errorless condition. The student selects courses that satisfy the scheduling requirements within a workflow involving automatic AI oversight.

C_{reactive_errors}: Automatic AI oversight, error-present condition. The student selects an invalid course that the advisor does not initially detect; automatic AI oversight flags the error so it can be corrected before the final schedule is produced.

4.2 Participants

Participants were recruited through Prolific (Prolific 2025) and completed the study remotely. Repeat participation was systematically restricted using Prolific's built-in eligibility controls to maintain data integrity. Participation was restricted to desktop or laptop users to ensure consistent interaction with the advising interface. In addition, participants were excluded from analysis if they failed the attention-check questions embedded throughout the study, including within the interactive simulation and the final questionnaire. No other exclusion mechanisms were applied.

A total of 160 participants completed the study. After accounting for those failing the attention-checks, the final analytic sample was $N = 157$. Each was compensated \$4.50 for their time upon completion of the study. All study procedures were approved by the institutional review board, and informed consent was obtained from all participants. Participants were assigned to experimental interactions via Prolific's randomization, and demographic information is summarized in the appendix. The study took approximately 17 minutes to complete, including the advising simulation and post-task questionnaire.

4.3 Measures

We intentionally employed multiple trust measures to capture distinct but complementary aspects of trust in human-AI advisory interactions. The Hendriks, Kienhues, and

Bromme (2015) instrument was used to assess epistemic trustworthiness in the advisor, while items adapted from Riedl, Hogeterp, and Reuter (2024) captured broader entity-specific trust judgments toward the advisor, AI assistant, and advisor-AI team. Consistent with Lee and See (2004), these items capture trust as an evaluative attitude toward system entities. Finally, the single-item trust measure provided a holistic assessment of trust that was not decomposed into specific dimensions, allowing us to examine whether broader trust judgments aligned with or diverged from measures of specific characteristics.

We adapted items from (Hendriks, Kienhues, and Bromme 2015; Riedl, Hogeterp, and Reuter 2024) for the academic advising scenario through minor wording changes when necessary (e.g., replacing domain-specific referents with "advisor," "AI assistant," or "advisor-AI team") while preserving the original response formats and construct intent. Unless otherwise noted, all multi-item measures were assessed using a 7-point Likert scale ranging from 1 ("Strongly [disagree/negative construct]") to 7 ("Strongly [agree/affirmative construct]"). Questionnaire items from Riedl, Hogeterp, and Reuter (2024) and Hendriks, Kienhues, and Bromme (2015) used in the study are reported in the appendices.

For multi-item constructs, we aggregate items within each instrument's intended structure and report results at the composite level when appropriate. This approach reduces item-level noise and aligns inference with the theoretical level at which the constructs are defined (e.g., Hendriks, Kienhues, and Bromme (2015) expertise, integrity, and benevolence), while preserving the original response scale as originally structured. In addition, aggregating items according to validated instrument structures (Hendriks, Kienhues, and Bromme 2015) reduces the number of statistical tests performed, thereby limiting family-wise error rate (FWER) inflation and the risk of spurious significance.

Epistemic Trustworthiness (METI) To assess participants' epistemic trust in the advisor, we administered items from the Muenster Epistemic Trustworthiness Inventory (METI) (Hendriks, Kienhues, and Bromme 2015). METI operationalizes epistemic trustworthiness along three dimensions: perceived expertise, integrity, and benevolence, reflecting perceived competence, adherence to professional standards (e.g., honesty and fairness), and goodwill toward others, respectively. Following the instrument's dimensional structure, we computed dimension scores by averaging item responses within each dimension (higher scores indicate stronger perceived epistemic trustworthiness). Item counts and internal consistency for each composite are reported in the appendices as is the full METI item set and dimension groupings used in this study.

Trust in the Advisor, AI System, and Team In addition to epistemic trustworthiness, we measured entity trust using three items adapted from Riedl, Hogeterp, and Reuter (2024). Items were adapted to the advising scenario. All participants rated the advisor with those in AI-supported conditions also rating the AI assistant and advisor-AI team. Advisor-only participants did not rate absent entities. An ex-

ample item is: “I can trust the information presented by the advisor.” All items demonstrated acceptable internal consistency as measures of entity trust (see technical appendices).

General Trust To capture a global assessment of trust that was not tied to a specific Likert instrument, participants completed a single-item general trust rating: “From 1–100, how trustworthy would you rate your advisor? (1 being the least trustworthy and 100 being the most)”. This measure served as a holistic outcome capturing overall trust in the advisor, unconstrained by specific item framing. A continuous scale allowed for finer-grained variation in global trust judgments than would be possible with a coarse Likert scale.

Likelihood of Advisor Reuse To assess behavioral intention, participants reported their likelihood of reusing the advisor in the future. This single-item outcome was analyzed separately from the multi-item trust composites.

We treat reuse intention as a behavioral proxy rather than a direct trust measure. Whereas trust instruments capture evaluative judgments (e.g., perceived expertise or trustworthiness), reuse intention reflects whether participants would act on those judgments by relying on the advisor again. Reuse intention is particularly relevant in advisory settings, where the practical consequence of trust is often continued reliance over time. This measure therefore assesses whether differences in epistemic or entity-level trust correspond to differences in anticipated reliance on the advisor.

4.4 Limitations

We took care in designing this simulation to approximate a realistic, though necessarily simplified, interaction involving AI-supported decision-making. Nevertheless, several limitations should be acknowledged. First, the use of a between-subjects design prevents individual participants from directly comparing different advising interactions or outcomes, which may limit insight into how trust recalibrates across repeated interactions.

As with all controlled experimental studies, the findings may also be limited in their generalizability beyond the simplified advising task, interface, and population studied. While we sought to enhance situational validity through a realistic advising scenario and grounded operationalizations of advisor performance and AI involvement, participants’ perceptions may differ in real-world advising contexts and especially in higher-stakes expert domains such as medicine, legal consultation, or finance. We therefore interpret the present results as evidence about trust attribution under conditions of epistemic dependence in a controlled proxy setting, rather than as a direct estimate of how strongly the same effects would appear in safety-critical environments.

A further limitation is that the AI was always correct when it intervened, allowing us to isolate advisor performance and intervention modality but not AI uncertainty or failure. This may have limited variation in AI-directed trust and contributed to the null AI-trust effects. Future work should vary AI reliability and examine how advisor and AI errors jointly shape trust in the advisor, AI assistant, and human-AI team.

Additionally, individual differences such as prior experience with academic advising, familiarity with AI systems,

or general attitudes toward automation may have influenced trust judgments. These factors were not explicitly measured and therefore fall outside the scope of the present study. Future work could examine how such individual characteristics interact with advisor performance and AI involvement to shape trust, including whether they moderate which aspects of trust are affected and which remain stable.

Although our results suggest that responsibility attribution remained anchored to the human advisor, we infer this pattern from the distribution of trust judgments across the advisor, AI assistant, and advisor-AI team rather than from a direct measure of perceived responsibility or accountability. Future work should include explicit responsibility-attribution measures to test this interpretation more directly.

5 Results

5.1 Overview of Analytical Approach

Advisor-directed outcomes were analyzed using 3×2 factorial ANOVAs crossing advisor performance (error vs. errorless) with AI modality (none, proactive, reactive). Because AI and advisor-AI-team ratings were collected only in AI-supported conditions, those outcomes used 2×2 factorial ANOVAs crossing performance with proactive versus reactive modality. These analyses evaluated the hypotheses: advisor performance effects tested H1 and H4, AI modality main effects tested H2, and advisor performance by modality interactions tested H3. Outcomes included epistemic trustworthiness composites (expertise, integrity, benevolence (Hendriks, Kienhues, and Bromme 2015)), entity-level trust adapted from Riedl, Hogeterp, and Reuter (2024), a 1 to 100 general trust rating, and likelihood of advisor reuse. Tukey’s HSD pairwise comparisons were conducted as exploratory analyses across all condition combinations. Effect sizes are reported as η^2 and Cohen’s d , with test statistics in standard ANOVA form and exact degrees of freedom. Variance differences between error-present and errorless conditions were assessed using Levene’s test.

5.2 Omnibus Tests Across Trust Constructs

We first report the full 3×2 factorial analyses, which provide the primary basis for evaluating the hypotheses. Additional omnibus statistics are reported in the appendices.

Advisor-directed trust (Riedl, Hogeterp, and Reuter 2024). Advisor performance significantly affected perceived trustworthiness, $F(1, 151) = 13.0, p < .001, \eta^2 = .08$; affective comfort, $F(1, 151) = 22.5, p < .001, \eta^2 = .13$; and information trustworthiness, $F(1, 151) = 19.6, p < .001, \eta^2 = .11$, with lower evaluations in error-present conditions. General trust, $F(1, 151) = 21.14, p < .001, \eta^2 = .12$, and likelihood of advisor reuse, $F(1, 151) = 26.29, p < .001, \eta^2 = .15$, showed the same pattern. These results support H1 and H4.

Epistemic trustworthiness (Hendriks, Kienhues, and Bromme 2015). Advisor performance significantly affected perceived expertise, $F(1, 151) = 21.89, p < .001, \eta^2 = .13$; integrity, $F(1, 151) = 8.80, p = .004, \eta^2 = .06$; and benevolence, $F(1, 151) = 6.92, p = .009, \eta^2 = .04$, with

higher evaluations in errorless than error-present conditions. These results further support H1.

AI-supported conditions (2×2 analysis) (Riedl, Hogeterp, and Reuter 2024). Restricting analyses to AI-supported conditions showed the same performance-driven pattern for advisor-directed trust measures (all $p \leq .003$, $\eta^2 = .08-.11$). Advisor-AI team evaluations showed significant performance effects for perceived trustworthiness, $F(1, 105) = 4.16$, $p = .044$, $\eta^2 = .04$, and information trustworthiness, $F(1, 105) = 7.07$, $p = .009$, $\eta^2 = .06$, but not affective comfort, $F(1, 105) = 3.75$, $p = .056$, $\eta^2 = .03$. AI assistant evaluations alone were not significantly affected by advisor performance (all $p > .05$).

No significant main effects of AI intervention modality were observed across outcomes in either the full 3×2 analyses (all $F(2, 151) \leq 1.81$, all $p \geq .168$, $\eta^2 \leq .023$) or the restricted 2×2 analyses (all $F(1, 105) \leq 1.10$, all $p \geq .296$, $\eta^2 \leq .010$), so H2 was not supported. No significant advisor performance by modality interactions were observed in either the 3×2 analyses (all $F(2, 151) \leq 2.55$, all $p \geq .082$, $\eta^2 \leq .033$) or the restricted 2×2 analyses (all $F(1, 105) \leq 1.83$, all $p \geq .179$, $\eta^2 \leq .017$), so H3 was not supported. To further evaluate these null effects, we conducted Bayesian ANOVAs on AI-supported trust evaluations using JASP’s default JZS Bayes factor approach with Cauchy priors. These analyses provided evidence favoring exclusion of AI modality effects ($BF_{\text{excl}} = 3.66-6.74$) and modality × performance interactions ($BF_{\text{excl}} = 3.62-14.53$) across AI assistant, advisor-AI team, and advisor-directed trust measures (refer to appendices).

5.3 Descriptive Statistics

Descriptive statistics for all trust constructs are reported in the appendix. Error-present conditions were evaluated less favorably across advisor measures, consistent with H1 and H4. AI modality differences showed no consistent pattern, consistent with unsupported H2 and H3. Error-present conditions also exhibited greater variability for several measures (e.g., expertise). Significant Levene’s test results were observed only for advisor-directed measures, with no significant variance differences for the AI assistant or combined team (refer to appendices). These exploratory variance analyses were not primary tests of the hypotheses.

5.4 Pairwise Condition Comparisons

Exploratory Tukey HSD comparisons, with factorial ANOVAs serving as the primary basis for inference, are reported with full statistics in the appendices. Epistemic expertise was lower in $C_{\text{adv_only_errors}}$ than in $C_{\text{adv_only_errorless}}$, and in $C_{\text{proactive_errors}}$ than in $C_{\text{proactive_errorless}}$ and $C_{\text{adv_only_errorless}}$. Affective comfort was lower in $C_{\text{adv_only_errors}}$ than $C_{\text{adv_only_errorless}}$, $C_{\text{reactive_errorless}}$, and $C_{\text{proactive_errorless}}$, while $C_{\text{proactive_errors}}$ was lower than $C_{\text{reactive_errorless}}$ and, in the AI-supported analysis, $C_{\text{proactive_errorless}}$. Information trustworthiness was lower in $C_{\text{adv_only_errors}}$ than in $C_{\text{reactive_errorless}}$, and in $C_{\text{proactive_errors}}$ than in $C_{\text{adv_only_errorless}}$, $C_{\text{reactive_errorless}}$, and, in the AI-supported analysis, $C_{\text{proactive_errorless}}$. Advisor reuse

and general trust were each lower in $C_{\text{adv_only_errors}}$ than in $C_{\text{adv_only_errorless}}$, $C_{\text{reactive_errorless}}$, and $C_{\text{proactive_errorless}}$, and lower in $C_{\text{proactive_errors}}$ than in $C_{\text{adv_only_errorless}}$ and $C_{\text{proactive_errorless}}$.

Overall, significant contrasts primarily reflected lower evaluations in error-present than in errorless conditions, supporting H1 and H4. They did not reveal a consistent AI-modality effect or advisor-performance-by-modality interaction supporting H2 or H3.

5.5 Summary of Results

Across analyses, advisor errors reduced advisor-directed trust evaluations, including epistemic trustworthiness, entity-specific trust, general trust, and advisor reuse, supporting H1 and H4. AI modality did not significantly change trust evaluations of the advisor, AI assistant, or advisor-AI team, and advisor performance effects did not differ across modalities. H2 and H3 were not supported by the frequentist analyses. Bayesian ANOVAs further provided evidence favoring exclusion of AI modality effects and modality × performance interactions across AI-supported trust evaluations (reported in appendices). The 3×2 and restricted 2×2 interaction analyses had approximately 79.9% and 73.4% power, respectively, to detect $f = .25$; smaller interactions cannot be ruled out. Pairwise comparisons primarily reflected error-present versus errorless contrasts.

6 Discussion

This study examines how advisor performance and AI involvement shape trust judgments in an interactive academic advising task. The design allows us to distinguish performance-driven trust calibration from attributional effects associated with AI intervention modality.

6.1 Trust Calibration to Performance and Modality

Although all conditions produced a correct schedule, participants responded meaningfully depending on whether the advisor failed to detect an invalidating course choice. AI modality did not produce reliable factor-level effects, suggesting that perceptions of advisor performance were the most salient factor shaping trust evaluations. The observed pattern aligns with prior work on trust calibration (Dzindolet et al. 2003), but in this setting these effects were concentrated on the human advisor rather than the AI assistant.

The reuse-intention results extend this pattern from evaluative trust judgments to anticipated reliance behavior. Although we do not directly model the relationship between entity-specific trust and reuse, the parallel decline in reuse intention suggests that performance-related trust penalties may have downstream implications for continued reliance.

Descriptive mean patterns suggest that AI involvement may shape evaluations in limited ways when errors are present, as performance-related declines followed a consistent ordering across modalities and multiple dependent variables, though these effects were not statistically significant at the factor level. Comparing average evaluation

drops between errorless and error-present conditions generally showed the largest penalties in advisor-only workflows, followed by advisor-invoked AI workflows, and the smallest penalties under automatic AI oversight (found in appendices). These descriptive patterns suggest a possible avenue for future work: modality differences may become more salient in higher-stakes contexts, where advisor-invoked AI could signal uncertainty differently than automatic AI oversight.

Beyond mean differences, failure conditions were associated with greater variability in trust responses, with standard deviations consistently higher in error conditions than in their errorless counterparts, particularly for advisor-directed measures where Levene's tests indicated significant differences. This pattern is detailed in the appendix. This may reflect differences in how participants tolerate failure, the extent to which these errors violated expectations, or potential compounding effects from repeated errors. However, identifying the underlying driver of this pattern was not part of our experimental design and is left for future work.

6.2 Trust Attribution and the Effects of AI Involvement

A central pattern in our results is that corrective AI changed the structure of the advisory workflow without significantly changing the target of trust consequences. Observed errors produced an attributional form of trust calibration, with trust consequences remaining anchored to the human advisor rather than being redistributed across the human-AI advisory system. Even when the AI assistant visibly detected and corrected errors, participants did not assign corresponding trust credit to the AI assistant. One interpretation is that AI correction creates a double signal: it demonstrates that the advisory workflow can recover from errors, but it also reveals that the human-facing advisor required correction. Under epistemic dependence, users may therefore treat the advisor as the primary bearer of responsibility, rather than crediting the AI assistant that contributes to the correction. In our setting, AI intervention modality defines how errors are detected and corrected, effectively acting as a governance mechanism over the advisory workflow. Across these modalities, verification authority is distributed differently. In advisor-only workflows, authority appears concentrated in the advisor; in proactive AI workflows, the advisor retains authority over when AI verification is invoked; and in reactive AI workflows, the AI assistant has more independent monitoring authority. However, changing this visible structure did not substantially redistribute trust among the advisor, the combined team, or the AI assistant alone. This interpretation is supported by the distribution of effects across trust targets in the measures adapted from Riedl, Hogeterp, and Reuter (2024) (see Section 5.2).

One possible explanation is that users anchor trust judgments to the human interacting with the AI, consistent with work on algorithmic aversion showing that people may prefer human judgment over algorithmic judgment after observing imperfections, even when the algorithm performs well (Dietvorst, Simmons, and Massey 2015). In our setting, however, the stronger effect was not distrust of the AI assis-

tant but the absence of corresponding trust credit for its corrective role. Another potential explanation is that the AI corrected the error without providing indications of its own certainty. Unlike systems that explicitly communicate certainty levels (Kim 2024), our AI presented its correction without signaling how confident it was in that contribution. Participants may therefore have interpreted the AI's role as procedural rather than judgment-based, leaving little basis for evaluating the AI's independent trustworthiness. This may help explain why trust in the AI did not significantly change.

Participants had limited evidence to evaluate the AI independently: it was always correct, and they saw neither failure nor the basis for its correction. Stable AI-directed trust may thus not generalize to variable reliability or greater transparency. Future work should vary transparency depth - from a correction notice to progressively disclosed rules, reasoning, or confidence - to test if disclosure shifts trust toward the AI assistant or advisor-AI team (Muralidhar et al. 2025).

These attributional patterns do not rule out a supportive role for AI in trust recovery. As discussed above, descriptive trends suggest that AI-supported workflows may modestly attenuate trust declines in error-present conditions (found in appendices). Although this pattern was not statistically significant, it warrants further study in higher-stakes settings or contexts where error transparency is nonnegotiable.

Taken together, these findings suggest a limit of procedural governance in human-AI advisory workflows. Altering who monitors, verifies, or corrects the workflow may not be sufficient to redistribute accountability when users rely on procedural cues under epistemic dependence. Instead, responsibility attribution remains concentrated on the human advisor, even when valuable contributions are more distributed across system components. This asymmetry may create an accountability burden for human experts in AI-assisted workflows: users may continue to assign responsibility to the human-facing advisor even when error detection and correction are distributed across the human-AI system.

6.3 Implications for Human-AI Advisory Design

Results indicate that trust and reliance show similar patterns, with both general trust and reuse intention declining in error-present conditions (refer to appendix), suggesting that sustaining user trust is critical for promoting continued reliance on human-AI advisory systems. Because trust penalties were directed primarily at the human advisor, maintaining trust in the advisor should be a central focus for system design. The AI assistant received little to no direct trust credit, despite meaningfully contributing to error detection and correction.

For designers and institutions, this suggests that AI oversight should not be treated as automatically redistributing trust: users may continue to evaluate the human-facing expert as the primary responsible party even when AI support visibly shapes the final outcome.

Advisor performance emerged as the dominant factor shaping trust evaluations, with failures causing substantial declines in trust. Although an intuitive approach could include hiding errors, this is not always ethically or practically feasible, particularly in domains requiring nonnegotiable transparency. Thus, system design should assume that

advisor errors may be observed and focus on preserving perceptions of advisor competence and managing their impact.

For design, the descriptive ordering suggests AI involvement may be most useful as a supportive safeguard rather than a substitute for advisor competence. Even when AI assistance attenuates trust loss, it may not eliminate the penalty or shift responsibility from the advisor. Automatic oversight may be especially useful in settings where advisor expertise is limited, but this possibility needs further empirical testing.

The greater variability observed after failures also has design implications: users may differ in how they interpret and tolerate advisor errors. One possible direction is to incorporate adaptive mechanisms that respond to user states of (dis)trust in real time, for example by adjusting the volume or quality of feedback provided following errors.

Together, these findings suggest that human-AI advisory design should prioritize advisor reliability while carefully structuring how AI participates in error detection and response. AI may play a supportive role in mitigating failure impact, but it should not be treated as a substitute for advisor competence or a mechanism for redistributing responsibility.

6.4 Generalizability and Broader Considerations

Although situated in academic advising, our study reflects a broader class of expert-guided decision settings in which users lack the domain knowledge needed to independently verify correctness. Similar dynamics arise in financial advising, legal consultation, medical triage, and technical support (Agrawal et al. 2023; Han and Ko 2025; Mayer, Davis, and Schoorman 1995; Kaye and Hamilton 2004), where users evaluate experts based on observed performance.

The importance of these cues is likely domain-dependent. Higher-stakes, time-critical, longer-term, or more formally accountable settings may shift how users interpret failures, assign responsibility, and distribute trust across entities.

Although AI intervention modality did not show reliable effects here, it may matter in other contexts. The findings suggest that such effects are likely secondary to observed expert performance, especially when users witness failures.

More broadly, these results suggest that the experimental setup can capture meaningful variation in trust judgments and responsibility attribution within human-AI advisory workflows. This is reflected in entity-specific trust evaluations, including statistically significant performance-driven effects and measurable, though not statistically significant, differences across AI intervention modalities.

7 Conclusion

This work contributes to the literature on trust in human-AI advisory systems by showing that trust judgments were more strongly shaped by advisor performance than by the structure of AI support. In an interactive academic advising context, trust depended on whether the advisor successfully identified and managed invalidating choices during the interaction, with error-present conditions leading to consistently lower evaluations across measures.

Across conditions, trust-related consequences were concentrated on the human advisor, with limited evidence that

variations in AI intervention modality systematically influenced evaluations. Even when AI visibly detected and corrected errors, trust did not shift favorably toward the AI, indicating that responsibility attribution remains anchored to the human advisor. While descriptive patterns suggest that AI involvement may attenuate trust penalties when invalid choices are present, these effects were not statistically significant. Together, these findings suggest that changes in governance structure, as operationalized through AI intervention modality, do not necessarily produce corresponding shifts in accountability within human-AI advisory workflows. These results indicate that improving advisor performance is likely more critical for sustaining trust than modifying the structure of AI integration, though the latter may become more important in higher-stakes settings.

By isolating advisor performance and AI intervention structure within a controlled advising scenario, this study provides evidence that users distinguish between the human advisor, the AI assistant, and the advisor-AI team when forming trust judgments. The results also suggest that responsibility attribution in human-AI advisory workflows may remain anchored to the human-facing expert, even when AI contributes to error detection and correction. Future work can extend this design to less controlled, real-world advisory environments, where additional contextual factors may further shape how trust and responsibility are formed. The broader implication is that AI oversight can make expert workflows more recoverable without necessarily making responsibility appear more distributed to users.

7.1 Future Work

Future work could examine these dynamics in repeated or longitudinal interactions as emphasized by Chen et al. (2025), where users engage with the same advisor or system over time. This would help assess how trust recalibrates with accumulated experience and whether the patterns observed here persist or change across interactions.

Another direction is to investigate how the amount of feedback provided during interaction influences trust. Varying the level of feedback or intervention may clarify how over- or under-communication affects trust calibration, particularly in scenarios involving advisor missteps.

Future work could also explore the use of physiological signals to better understand and potentially modulate trust in real time. Measures such as heart rate variability, eye tracking, etc. may provide insight into user state and enable adaptive systems that respond to shifts in trust or uncertainty.

Finally, extending this design to higher-stakes advisory contexts (e.g., finance, law, medicine) would assess the generality of these findings, where consequences may alter how users interpret advisor performance and AI involvement.

Acknowledgments

Dr. Sreedharan's research is supported by NSF through grant 2303019. This material is based in part upon work supported by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program. The views

and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

- Agrawal, V.; Kandul, S.; Kneer, M.; and Christen, M. 2023. From OECD to India: Exploring cross-cultural differences in perceived trust, responsibility and reliance of AI and human experts. *arXiv preprint arXiv:2307.15452*.
- Asan, O.; Bayrak, A. E.; and Choudhury, A. 2020. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 22(6): e15154.
- Bisconti, P.; Aquilino, L.; Marchetti, A.; and Nardi, D. 2024. A Formal Account of Trustworthiness: Connecting Intrinsic and Perceived Trustworthiness. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 131–140.
- Chen, R.; Wang, R.; Sadeh, N.; and Fang, F. 2025. Missing Pieces: How Do Designs that Expose Uncertainty Longitudinally Impact Trust in AI Decision Aids? An In Situ Study of Gig Drivers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 790–817. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1): 114–126.
- Do, H. J.; and Geyer, W. 2025. Hide or Highlight: Understanding the Impact of Factuality Expression on User Trust. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 785–797.
- Dzindolet, M. T.; Peterson, S. A.; Pomranky, R. A.; Pierce, L. G.; and Beck, H. P. 2003. The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies*, 58(6): 697–718.
- EDUTECHLoft. 2025. Using AI to Enhance Student Advising, Enrollment, and Communication in Higher Education. Accessed: 2025-12-06.
- Griffin, D.; and Tversky, A. 1992. The Weighing of Evidence and the Determinants of Confidence. *Cognitive Psychology*, 24(3): 411–435.
- Han, J.; and Ko, D. 2025. Trust Formation, Error Impact, and Repair in Human–AI Financial Advisory: A Dynamic Behavioral Analysis. *Behavioral Sciences*, 15(10): 1370.
- Hendriks, F.; Kienhues, D.; and Bromme, R. 2015. Measuring Laypeople’s Trust in Experts in a Digital Age: The Muenster Epistemic Trustworthiness Inventory (METI). *PLOS ONE*, 10(10).
- Hoff, K. A.; and Bashir, M. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3): 407–434.
- Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 624–635. ACM.
- Kaffee, L.-A.; Atanasova, P.; and Rogers, A. 2025. Local Differences, Global Lessons: Insights from Organisation Policies for International Legislation. *ArXiv:2503.05737*, arXiv:2503.05737.
- Kaye, K.; and Hamilton, S. 2004. Roles of Trust in Consulting to Financial Families. *Family Business Review*, 17(2): 151–163.
- Kim, D.; Daneshi, R.; Draper, B.; and Sreedharan, S. 2026. Asymmetric Trust Effects of Corrective AI in Expert Advisory Workflows under Epistemic Dependence. Extended version with technical appendices, arXiv:2602.11522.
- Kim, P. H.; Ferrin, D. L.; Cooper, C. D.; and Dirks, K. T. 2004. Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- Versus Integrity-Based Trust Violations. *Journal of Applied Psychology*, 89(1): 104–118.
- Kim, S. S. Y. 2024. Establishing Appropriate Trust in AI Through Transparency and Explainability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. ACM.
- Kim, S. S. Y.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 2024. "I'm Not Sure, But...": Examining the Impact of LLMs' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*.
- Kim, S. S. Y.; Watkins, E. A.; Russakovsky, O.; Fong, R.; and Monroy-Hernández, A. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1): 50–80.
- Lee, M. K. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1).
- Liao, Q. V.; and Sundar, S. S. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. New York, NY, USA: Association for Computing Machinery.
- Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review*, 20(3): 709–734.
- Moëll, B.; Aronsson, F. S.; and Akbar, S. 2025. Medical Reasoning in LLMs: An In-Depth Analysis of DeepSeek R1. *ArXiv:2504.00016*, arXiv:2504.00016.
- Muralidhar, D.; Belloum, R.; Marçal de Oliveira, K.; Ashok, A.; and Mohammad, P. B. 2025. The Effect of Progressive Disclosure in the Transparency of Large Language Models. In Plácido da Silva, H.; and Ciproso, P., eds., *Computer-Human Interaction Research and Applications*, volume 2370 of *Communications in Computer and Information Science*, 269–288. Springer Nature Switzerland.

- Nakano, H.; Takezawa, J.; Matulic, F.; Yang, C.-L.; and Yatani, K. 2025. Understanding Reader Perception Shifts upon Disclosure of AI Authorship. ArXiv preprint, arXiv:2510.24011.
- Narayanan, S.; Yu, G.; Ho, C.-J.; and Yin, M. 2023. How Does Value Similarity Affect Human Reliance in AI-Assisted Ethical Decision Making? In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 49–57. New York, NY, USA: Association for Computing Machinery.
- Origgi, G. 2014. Epistemic Trust. In Capet, P.; and Delavalade, T., eds., *Information Evaluation*, chapter 2, 35–54. ISTE Ltd and John Wiley & Sons, Inc.
- Parasuraman, R.; Sheridan, T. B.; and Wickens, C. D. 2000. A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 30(3): 286–297.
- Pareek, S.; Velloso, E.; and Goncalves, J. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 16 pages. New York, NY, USA: ACM.
- Prolific. 2025. Prolific: Online Platform for Participant recruitment. Online platform. Version used for data collection; accessed 2025-12-30.
- Qualtrics. 2025. Qualtrics XM. Software platform. Version used for data collection; accessed 2025-12-30.
- Riedl, R.; Hogeterp, S. A.; and Reuter, M. 2024. Do Patients Prefer a Human Doctor, Artificial Intelligence, or a Blend, and Is This Preference Dependent on Medical Discipline? Empirical Evidence and Implications for Medical Practice. *Frontiers in Psychology*, 15.
- Sperber, D.; Clément, F.; Heintz, C.; Mascaro, O.; Mercier, H.; Origgi, G.; and Wilson, D. 2010. Epistemic Vigilance. *Mind & Language*, 25(4): 359–393.
- Tenney, E. R.; Spellman, B. A.; and MacCoun, R. J. 2008. The Benefits of Knowing What You Know (and What You Don't): How Calibration Affects Credibility. *Journal of Experimental Social Psychology*, 44: 1368–1375.
- Wong, K. K. L.; Han, Y.; Cai, Y.; Ouyang, W.; Du, H.; and Liu, C. 2025. From Trust in Automation to Trust in AI in Healthcare: A 30-Year Longitudinal Review and an Interdisciplinary Framework. *Bioengineering*, 12: 1070.
- Yang, S.; and Ma, R. 2025. Classifying Epistemic Relationships in Human–AI Interaction: An Exploratory Approach. ArXiv preprint, arXiv:2508.03673.