# Adjust for Trust:
# Mitigating Trust-Induced Inappropriate Reliance on AI Assistance

**Tejas Srinivasan** and **Jesse Thomason**
University of Southern California
{tejas.srinivasan,jessetho}@usc.edu

## Abstract

Trust biases how users rely on AI recommendations in AI-assisted decision-making tasks, with low and high levels of trust resulting in increased under- and over-reliance, respectively. We propose that AI assistants should adapt their behavior through *trust-adaptive interventions* to mitigate such inappropriate reliance. For instance, when user trust is low, providing an explanation can elicit more careful consideration of the assistant's advice by the user. In two decision-making scenarios—laypeople answering science questions and doctors making medical diagnoses—we find that providing supporting and counter-explanations during moments of low and high trust, respectively, yields up to 38% reduction in inappropriate reliance and 20% improvement in decision accuracy. We are similarly able to reduce over-reliance by adaptively inserting forced pauses to promote deliberation. Our results highlight how AI adaptation to user trust facilitates appropriate reliance, presenting exciting avenues for improving human-AI collaboration.

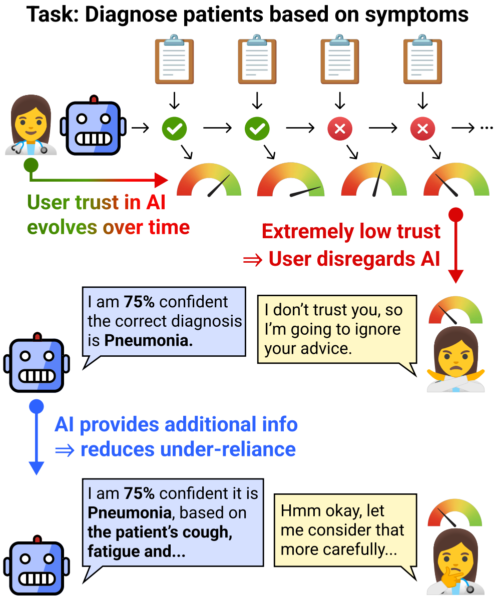

Task: Diagnose patients based on symptoms
User trust in AI evolves over time
Extremely low trust ⇒ User disregards AI
I am **75%** confident the correct diagnosis is **Pneumonia.**
I don't trust you, so I'm going to ignore your advice.
AI provides additional info ⇒ reduces under-reliance
I am **75%** confident it is **Pneumonia**, based on **the patient's cough, fatigue and...**
Hmm okay, let me consider that more carefully...

Figure 1: User trust in AI systems evolves over a series of decision-making interactions, impacting how carefully the user considers future AI recommendations. To mitigate the effects of extreme trust and encourage critical reliance, AI systems should adapt their behavior to users' trust levels. For instance, when trust is low, providing explanations reduces under-reliance.

## 1 Introduction

AI systems are being deployed to assist humans in a wide range of decision-making tasks (Cai et al., 2019; Chiang et al., 2023; Che et al., 2024). AI-assisted decision-making (Lai et al., 2023) typically consists of an AI system providing a recommendation for the user's consideration. A key factor modulating how users incorporate AI advice is *user trust*, which is the user's belief that the AI will help them achieve their goals in situations of uncertainty and vulnerability (Lee and See, 2004). Having higher trust makes users more likely to accept the AI's recommendation, all else being equal (Dzindolet et al., 2003). Moreover, trust is not a static belief, but instead continuously evolves as the user interacts with the AI and observes decision outcomes (Dhuliawala et al., 2023).

User trust does not always align with AI assistant trustworthiness, i.e. its true capability to help the user (Wright, 2010). Miscalibrated trust (Jacovi et al., 2021) may develop due to recency bias, the user's internal biases towards AI, or the assistant's inability to communicate its reasoning or limitations. Miscalibrated trust acts as a cognitive bias (Lee, 2024) and hinders critical evaluation of AI recommendations, resulting in *inappropriate reliance* (Parasuraman and Riley, 1997). For example, we find that doctors mistakenly accept 26% of AI misdiagnoses when their trust is high, compared to 8% when trust is lower, indicating over-reliance. Conversely, when user trust is low, doctors reject correct AI diagnoses 68% of the time, up from 40% otherwise, indicating a bias towards under-reliance.

We posit that AI assistants should adapt their behavior in response to users' trust levels in order to mitigate inappropriate reliance caused by extreme (low or high) trust. For instance, when trust is low, the assistant can reduce risk of disuse by providing the user with additional reasoning to support its recommendation (Figure 1). Similarly, when users are too trusting, the assistant can highlight reasons its recommendation may be incorrect, or can simply slow down the interaction. We hypothesize that strategically introducing these *trust-adaptive interventions* will prompt users to engage more carefully with AI advice, rather than accepting or rejecting advice without due consideration.

We examine the effect of trust-adaptive AI interventions at mitigating inappropriate reliance on two decision-making tasks: answering science trivia questions and making medical diagnoses based on patient symptoms. We first validate our premise that, when interacting with the AI assistant over a sequence of decision-making problems, users' trust level at the start of a given interaction affects their reliance behavior and decision-making performance (§3.2), with extreme levels of user trust (too high or too low) resulting in increased inappropriate reliance. Through controlled studies, we find that strategically providing supporting explanations when user trust is low reduces under-reliance and improves decision-making accuracy (§4.1). Similarly, providing counter-explanations reduces over-reliance when trust is high (§4.2). Combining these interventions to mitigate under- and over-reliance yields complementary improvements in decision-making accuracy and inappropriate reliance (§4.3). We also evaluate the utility of intervening by decelerating the interaction, finding that it helps reduce over-reliance but not under-reliance (§4.4).

Our findings highlight the utility of adapting to user trust in AI-assisted decision-making and present exciting avenues for facilitating appropriate reliance and improving human-AI collaboration.

## 2 Related Work

We draw on a vast literature on measuring user trust in AI systems and evaluate how decision aids can be used for mitigating inappropriate reliance in moments of low or high trust.

**Trust in Human-AI Interactions.** Much work has explored the nature of human trust in AI systems (Lai et al., 2023), particularly in situations characterized by risk and uncertainty (Jacovi et al., 2021). Lee and See (2004) provide the most commonly accepted definition of trust, as a person's attitude that an agent will help them achieve their goals. User trust is considered to be calibrated (Alizadeh et al., 2022) when it aligns with the AI system's true capabilities (Wright, 2010), thus reducing AI misuse and disuse (Alizadeh et al., 2022). While trust in AI has typically been attributed to socio-economic and individual factors (Bach et al., 2024), recent work (Dhuliawala et al., 2023; Pareek et al., 2024) examines how trust develops as users interact with AI systems over multiple timesteps.

**Measuring Trust.** Bach et al. (2024) identify a variety of mechanisms for measuring user trust, such as questionnaires (Schaffer et al., 2019), qualitative interviews (Barda et al., 2020), surveys (Lin et al., 2019), and point scales (Gulati et al., 2019). In AI-assisted decision-making, early works measured trust by observing user reliance behavior (Yin et al., 2019; Zhang et al., 2020); however, de Fine Licht and Brülde (2021) distinguish reliance, an observable behavior, from trust, a subjective belief. Self-reported trust levels from users, where users report their confidence in the AI's accuracy for a question, are a reasonable proxy for trust, albeit at a local, interaction level (Pareek et al., 2024; Dhuliawala et al., 2023). Instead, we adopt a more global lens for eliciting trust scores by asking users to report their belief in the AI's helpfulness on a scale of 0 to 10.

**Mitigating Inappropriate Reliance.** Inappropriate reliance, where users mistakenly accept incorrect AI predictions or reject correct ones (Parasuraman and Riley, 1997), is highly undesirable in high-stakes domains, such as healthcare and law (Schemmer et al., 2023). Appropriate reliance can be fostered through various decision aids, such as model confidences (Zhang et al., 2020; Vodrahalli et al., 2022), explanations (Wang and Yin, 2021; Bansal et al., 2021), uncertainty expressions (Zhou et al., 2024; Kim et al., 2024), and providing sources (Feng and Boyd-Graber, 2019). Cognitive forcing functions (Buçinca et al., 2021), which insert friction (Chen and Schmidt, 2024; İnan et al., 2025) and promote deliberation (Park et al., 2019; Rastogi et al., 2022; Ma et al., 2024a), are effective at mitigating over-reliance. We demonstrate how strategically providing these decision aids to users during moments of low or high trust can mitigate trust-induced inappropriate reliance.

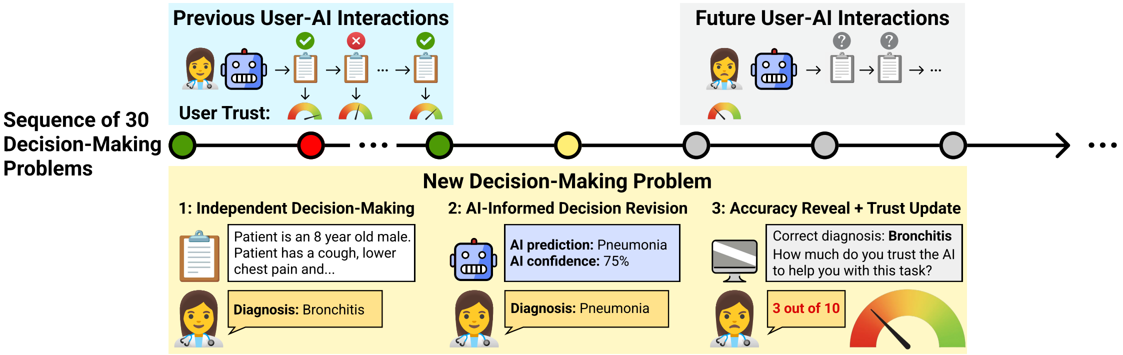

Previous User-AI Interactions
User Trust:
Future User-AI Interactions
Sequence of 30 Decision-Making Problems
New Decision-Making Problem
1: Independent Decision-Making
2: AI-Informed Decision Revision
3: Accuracy Reveal + Trust Update
Patient is an 8 year old male. Patient has a cough, lower chest pain and...
Diagnosis: Bronchitis
AI prediction: Pneumonia
AI confidence: 75%
Diagnosis: Pneumonia
Correct diagnosis: Bronchitis
How much do you trust the AI to help you with this task?
3 out of 10

Figure 2: In our user study, each user interacts with an AI for a sequence of 30 decision-making problems. In each problem, the user first makes a decision by themselves, and then receives advice from the AI which they use to make a final decision. The user is then told what the correct decision is, and reports their trust in the AI (out of 10).

## 3 How does Trust Impact Reliance on AI?

We study how user trust impacts reliance on AI advice over a sequence of decision-making problems (§3.1). User studies reveal that trust being too high or too low increases inappropriate reliance (§3.2).

### 3.1 Sequential AI-Assisted Decision-Making

We consider a setting where a human user interacts with an AI assistant on a sequence of $N$ decision-making problems, where each problem belongs to the same task, such as making medical diagnoses based on symptoms. Problems are tuples of input $x$, categorical choices $\mathcal{Y}$, and correct choice $y^* \in \mathcal{Y}$. The user solves each problem in three stages: **independent decision-making** based on their own knowledge, **decision revision** after viewing the AI recommendation, and **trust update** in the AI after observing decision accuracy (Figure 2).[1]

**1. Independent Decision-Making:** For the $i^{th}$ decision-making problem with input $x_i$, the user initially makes a decision $y_i^{u,\text{init}} \in \mathcal{Y}$.

**2. AI-Informed Decision Revision:** The user then views the AI prediction $y_i^{AI}$ and confidence estimate $c_i^{AI}$, and makes a final decision $y_i^{u,\text{fin}}$.

**3. Trust Update:** After the user makes their final decision $y_i^{u,\text{fin}}$, the interface informs the user of the accuracy of their decision and the AI prediction $y_i^{AI}$. Observing this feedback may alter the user's trust in the AI's ability to help them make better decisions. For instance, the AI misleading the user into making a wrong decision is likely to decay trust. After showing this feedback, we ask users to report how much they trust the AI to help them with the decision-making task based on all user-AI interactions so far, as an integer between 0 and 10.

Our trust operationalization captures a global belief in the AI's helpfulness, which is likely to influence how the user relies on AI advice in subsequent decision-making problems.

**Evaluating Appropriate Reliance.** We only evaluate interactions where the user's initial decision differs from the AI prediction, i.e. $y_i^{u,\text{init}} \neq y_i^{AI}$. We capture the degree of users' reliance on AI assistance using **Switch Rate** (Yin et al., 2019; Zhang et al., 2020), the fraction of interactions where the user switched their decision to the AI prediction. Following Ma et al. (2024b), we capture the degree of *appropriate* reliance on AI assistance using two metrics: **Over-Reliance** represents the fraction of interactions where the user switches to the AI's prediction when the AI is incorrect, while **Under-Reliance** represents the fraction of interactions where the user does not switch to the AI's prediction when the AI was in fact correct.

We hypothesize that very high and very low values of trust will increase the probability of inappropriate reliance in the *next* user-AI interaction. Specifically, we hypothesize that low user trust biases users towards rejecting correct AI advice, i.e. higher Under-Reliance. Similarly, high trust leads to users accepting incorrect advice more frequently, i.e. higher Over-Reliance.

### 3.2 Experiments

We evaluate the impact of users' trust level[2] on their reliance behavior and decision-making performance in subsequent user-AI interactions.

[1] Our setup assumes that both the user and AI can observe the ground-truth decision after each problem.

[2] Henceforth, "user trust" refers to the user's trust level at the start of an interaction, i.e. the trust score reported by the user at the end of the previous interaction.

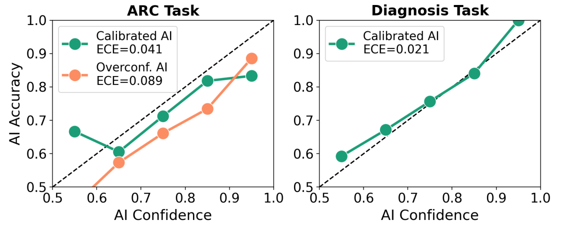


Figure 3: Calibration curves and Expected Calibration Error (ECE) of our simulated AI assistants.

**Decision-Making Tasks.** We conduct user studies on two decision-making tasks. The **ARC** task consists of answering science questions from the ARC dataset (Clark et al., 2018). Each problem consists of a question and two options. The **Diagnosis** task involves making medical diagnoses based on patient intake forms, sourced from the DDXPlus dataset (Fansi Tchango et al., 2022). Users must select from four possible diagnoses. Appendix A contains more details about problem selection.

**Simulated AI.** Our user studies use a simulated AI that provides a recommendation $R_i = (y_i^{AI}, c_i^{AI})$ for each decision-making problem. We experiment with two types of AI assistants: one perfectly calibrated and one overconfident.

For the calibrated AI assistant, for the $i^{th}$ problem, we first sample a confidence score $c_i^{AI} \sim \text{Uniform}(0.5, 0.95)$. We then decide if the AI prediction $y_i^{AI}$ will be the correct decision $y_i^*$ by sampling with probability $c_i^{AI}$:

$$y_i^{AI} = \begin{cases} y_i^* & \text{w.p. } c_i^{AI}, \\ \sim \text{Uniform}(\mathcal{Y} \setminus \{y_i^*\}) & \text{w.p. } 1 - c_i^{AI} \end{cases}$$

For the overconfident AI assistant, we sample the AI confidence $c_i^{AI}$ as above, and then sample another parameter $c_i'$ from the triangular distribution $\texttt{Tri}(0.5, c_i^{AI}, c_i^{AI})$. The AI prediction is sampled as before, but with a correctness probability $c_i'$ lower than the confidence $c_i^{AI}$ shown to the user.

This sampling procedure generates AI predictions and confidence scores for each decision-making problem. Figure 3 shows calibration curves for the calibrated and overconfident AI predictions.

**Experiment Setup.** We perform user studies on the Prolific platform, on three task settings: the ARC task with a calibrated AI (`ArcC`), the ARC task with an overconfident AI (`ArcO`), and the Diagnosis task with a calibrated AI (`DiagC`).

For the ARC task, we recruit users with at least an undergraduate degree. We recruit two groups of 30 users, one each for the `ArcC` and `ArcO` settings. Users achieve 67% accuracy on this task without AI assistance, whereas the calibrated and overconfident AI achieve 71% and 64% accuracy, respectively. Users are paid \$1.0, plus a \$0.10 bonus for every correct final decision.

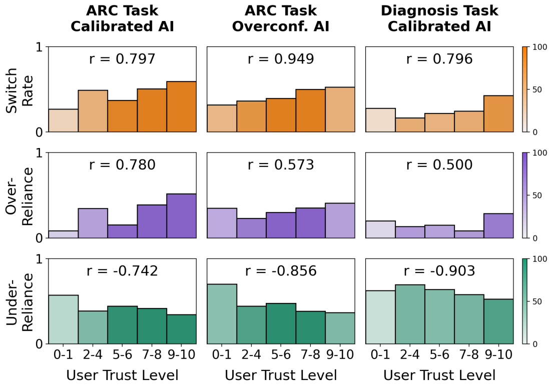


Figure 4: Reliance metrics at different levels of user trust. In each plot, $r$ represents the weighted Pearson correlation coefficient. All correlations are statistically significant, with $p < 0.001$. Bar shades correspond to number of user interactions at each trust level.

For the Diagnosis task (`DiagC` task setting), we conduct studies with professional doctors, who achieve 74% task accuracy without the AI. Due to the lower number of qualified participants on Prolific, we recruit 20 users. Users are paid \$2.0, plus a \$0.10 bonus for every correct final decision.

For each task setting, we sample 10 sequences $S_i = \{P_1^i, P_2^i, ..., P_{30}^i\}$ of 30 decision-making problems $P_j^i = \{(x_j, y_j^*), R_j\}$. Users in each task setting are randomly assigned to a sequence $S_i$ upon starting the study. Appendix A contains additional details about the user study setup.

**Findings.** Figure 4 shows the relationship between user trust and reliance, aggregated across all users in the same task setting. We highlight a few key takeaways (**T***). **T1**: `Switch Rate` is strongly correlated with user trust, which suggests that users' internal trust influences how likely they are to accept AI advice. **T2**: User trust has moderate to strong correlation with `Over-Reliance`. We further observe that `Over-Reliance` is highest at *high* values of user trust (9–10). **T3**: User trust has a strong negative correlation with `Under-Reliance`. At lower values of trust ($< 5$), users exhibit the most `Under-Reliance`. **These findings suggest that extreme values of user trust act as a cognitive bias, resulting in higher inappropriate reliance in subsequent human-AI interactions.**

## 4 Trust-Adaptive AI Interventions Mitigate Inappropriate Reliance

We hypothesize that AI systems can counterbalance the cognitive bias caused by trust by adapting their behavior according to the user's trust level.[3] We introduce *trust-adaptive interventions* for reducing inappropriate reliance. Trust-adaptive interventions are designed to correct for trust-induced cognitive bias, and are only applied when the user's trust level is either above or below a certain threshold. We hypothesize that uniformly applying these interventions, rather than only when trust is low or high, will worsen inappropriate reliance.

**Experiment Setup.** We evaluate the utility of adaptive interventions through a between-subjects study.[4] Each user is assigned to one of three experimental conditions: a **No Intervention** baseline where the intervention is never applied, an **Intervention Always** baseline where the intervention is applied whenever the AI's prediction differs from the user's initial decision, and the **Trust-Adaptive Intervention** condition where the intervention is only applied when the user's trust lies above or below a specified threshold. Based on the relationship we observe between reported user trus and reliance (Figure 4), we select a threshold trust of 5 out of 10 for mitigating under-reliance, and 8 out of 10 for mitigating over-reliance. We conduct studies with 30 users assigned to each experimental condition in the `ArcC` and `ArcO` task settings, and 20 users for each condition in the `DiagC` task setting.

**Evaluating Interventions.** We evaluate intervention conditions on `Under-Reliance` or `Over-Reliance`, depending on the type of inappropriate reliance the intervention is intended to mitigate. We also evaluate **`Total Inappropriate Reliance`**, which is the sum of `Under-Reliance` and `Over-Reliance`, to check if mitigating one type of inappropriate reliance exacerbates the other type to the same degree. Finally, **`Final Decision Accuracy`** captures the effect of interventions on decision-making performance.

We compute metrics across all user interactions where the user and AI disagree ($y_i^{u,\text{init}} \neq y_i^{AI}$) and also analyze subsets of these interactions based on user trust levels. Aggregating over interactions rather than users may skew user representation, for example, a user with generally low trust will represent more low trust interactions than other users. On the other hand, macro-aggregation by averaging per-user metrics results in high inter-user variance, since some users have very few interactions meeting the specified criteria. Macro-aggregation results are presented in Appendix C.

### 4.1 Mitigating Under-reliance with Supporting Explanations

AI explanations have been widely studied as a decision aid (Bussone et al., 2015; Wang and Yin, 2021; Poursabzi-Sangdeh et al., 2021). Prior work has shown that natural language explanations supporting the AI prediction cause over-reliance (Si et al., 2024; Sieker et al., 2024; Hashemi Chaleshtori et al., 2024). We hypothesize that providing natural language supporting explanations when user trust is low ($< 5$) can mitigate under-reliance.

We generate supporting explanations for all problems by prompting GPT-4o (Hurst et al., 2024) to provide a 3–4 sentence explanation $E_i^s$ for each option $y_i \in \mathcal{Y}$ of each problem $P$ (prompts in Table 5). Explanations were manually reviewed by an author to ensure they entailed the corresponding prediction. When the intervention was applied during a user-AI interaction, the AI would provide an augmented recommendation $R_i' = R_i + E_i^s$. To encourage users to read the explanation, they are allowed to make their final decision only 15 seconds after the AI advice is shown.

We separately evaluated the interventions across interactions where the user's initial decision disagrees with the AI prediction, and across the "low trust" subset of interactions. Figure 5 shows that **providing supporting explanations *adaptively* mitigates under-reliance when user trust is low**, especially when AI confidence is miscalibrated.

**Trust-Adaptive Explanations Help.** In the Trust-Adaptive condition, users exhibit lower `Under-Reliance`, lower inappropriate reliance, and higher decision accuracy *across all task settings*. These improvements are particularly notable when user trust is low and an explanation is provided in the Trust-Adaptive Intervention condition but not in the No Intervention condition. Providing explanations when trust is low results in 13–31% reduction in `Under-Reliance`, 9–38% reduction in `Total Inappropriate Reliance`, and 10–19% improvement in `Final Decision Accuracy`.

[3] Our setup assumes that the AI can observe the user's last-reported trust level at the start of every interaction.

[4] We do not conduct a within-subjects study because the same user cannot be subjected to multiple conditions without introducing interaction effects.

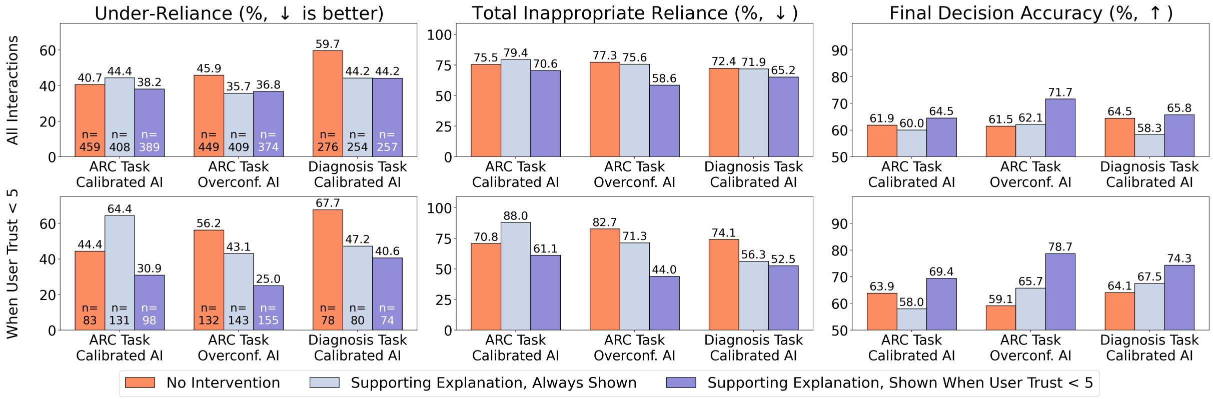


Figure 5: Reliance metrics and decision accuracy for users, evaluating the utility of supporting explanations at mitigating under-reliance. $n$ represents the number of user-AI interactions that we aggregate over for the corresponding condition. Showing explanations adaptively reduces `Under-Reliance` and `Total Inappropriate Reliance` while boosting `Final Decision Accuracy` across all task settings, particularly when user trust is low.

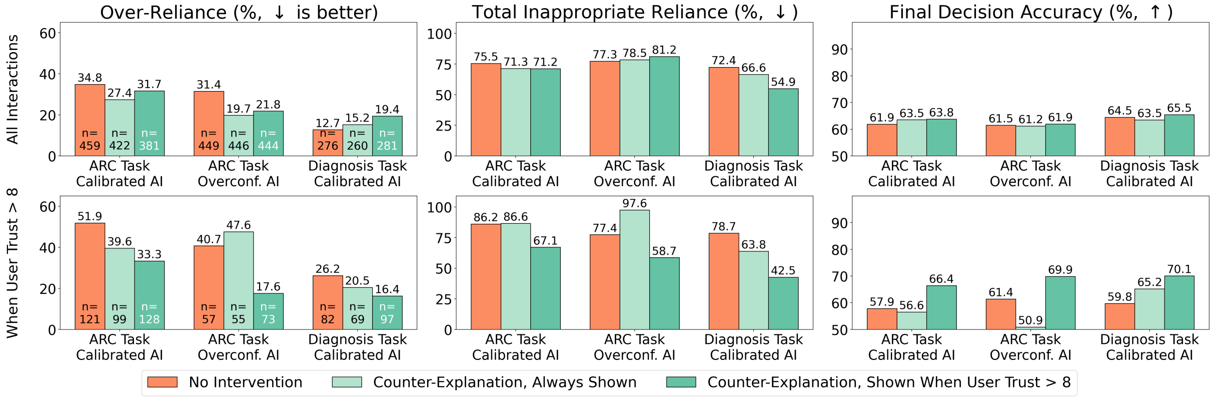


Figure 6: Reliance metrics and decision accuracy for users, evaluating the utility of counter-explanations at mitigating over-reliance. Showing counter-explanations adaptively reduces `Over-Reliance` and `Total Inappropriate Reliance` while boosting `Final Decision Accuracy` across almost all settings, particularly when trust is high.

**Explanations Offset AI Miscalibration.** The largest improvements occur when the AI system is over-confident, i.e. the `ArcO` task setting, which had the largest number of interactions where user trust was low. This finding indicates that explanations are effective for mitigating AI disuse when AI confidences are miscalibrated.

**Persistent Explanations Can Hurt.** The Intervention Always condition does not yield similar improvements and worsens decision accuracy in the Diagnosis task. In the `ArcC` task setting, showing explanations uniformly *increases* `Under-Reliance` when trust is low. We suspect that showing explanations always rather than adaptively exposes the user to many misleading explanations, resulting in an overall loss of trust in the explanations' trustworthiness. In `ArcO` and `DiagC`, explanations reduce inappropriate reliance when trust is low, but not across all user interactions.

## 4.2 Mitigating Over-reliance with Counter-Explanations

Similar to how providing supporting explanations are an effective intervention when user trust is low, we investigate whether providing reasons for why the model prediction might be incorrect can counter-balance the cognitive effect of high trust ($> 8$). Such *counter-explanations* have been shown to reduce over-reliance compared to regular supporting explanations (Si et al., 2024).

Similar to the supporting explanations, we generate natural language counter-explanations for each option by prompting GPT-4o to list 1–2 reasons why that option *might* be incorrect, while not completely rejecting that option (e.g. "I believe Bronchitis is the correct diagnosis due to ..., but it is possible that..."). The counter-explanations frequently include expressions of uncertainty ("could potentially", "it may be that"), alternative possibilities, and specific circumstances under which the

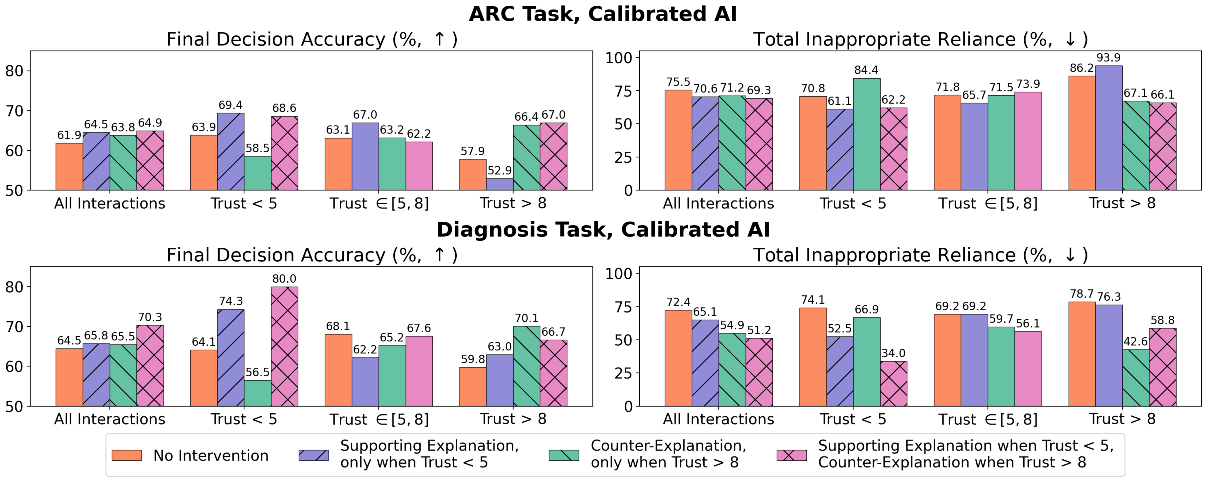


Figure 7: Effect of providing supporting explanations (/) and counter-explanations (\), depending on user trust, within the same user session yields complementary benefits in inappropriate reliance and decision-making accuracy.

model's prediction may be incorrect. Table 6 contains examples of generated counter-explanations.

Figure 6 shows that **providing counter-explanations *adaptively* mitigates over-reliance when user trust is high**.

**Trust-Adaptive Counter-Explanations Help.** Counter-explanations are effective at reducing over-reliance when user trust is high, across *all* task settings, with 10–23% reduction in `Over-Reliance`, 19–36% reduction in `Total Inappropriate Reliance`, and 8–20% improvement in `Final Decision Accuracy`. When considering all interactions where the user and AI have different prediction, improvements are also observed in almost all cases but to a lesser extent.

**Persistent Explanations Are Not As Helpful.** 2: The effects become less pronounced in the Intervention Always condition. In the `ArcO` setting, users perform uniformly worse than in the No Intervention condition when trust is high. Users may be less inclined to closely evaluate counter-explanations when they are always shown.

### 4.3 Mitigating Under- and Over-Reliance Simultaneously

We now investigate whether showing supporting explanations when trust is low ($< 5$) and counter-explanations when trust is high ($> 8$) in the same user-AI study yields complementary improvements in decision accuracy and inappropriate reliance.

Figure 7 shows that **providing different types of explanations based on user trust yields complementary performance improvements** on both the ARC and Diagnosis tasks. The benefits observed by using supporting explanations during low trust mirror those previously observed in §4.1, while the benefits of using counter-explanation are similar to those observed in §4.2.

### 4.4 Intervention through Deceleration

We have demonstrated that supporting and counter-explanations are useful for mitigating under- and over-reliance, respectively. We now investigate the utility of another type of intervention: slowing down the interaction to promote deliberation, without providing additional information to the user.

To mitigate under-reliance by decelerating the interaction, we display a "The AI is thinking..." message for 10 seconds before the AI prediction is revealed to the user. To mitigate over-reliance, we reveal the AI prediction and ask the user to carefully consider the AI advice; the user is made to wait 10 seconds before making their final decision.

**Findings.** Figure 8 shows the effect of the above decelerating interventions at mitigating inappropriate reliance in the `ArcC` task setting. We see that telling users that the AI is thinking is not particularly effective at reducing `Total Inappropriate Reliance`, even when trust is low. On the other hand, forcing users to consider the AI advice closely when trust is high improves `Total Inappropriate Reliance` and `Final Decision Accuracy`. Our findings extend those of Buçinca et al. (2021), showing that cognitive forcing is particularly useful for reducing over-reliance when user trust is high.

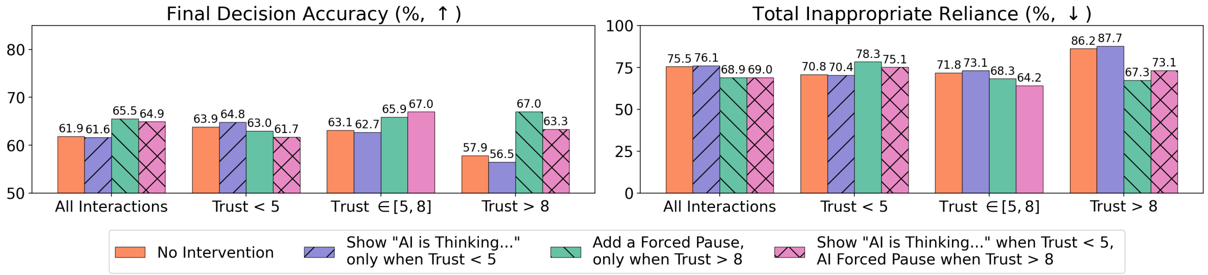


Figure 8: Effect of decelerating interventions on reducing inappropriate reliance and improving decision accuracy.

## 5 Discussion

Our results highlight the promise of trust-adaptive interventions based on experiments in a controlled setting. We highlight some considerations when designing trust-adaptive interventions for real-world decision-making scenarios.

**Applying trust-adaptive interventions.** A key assumption of our setup does not hold in most real-world settings: that both parties have access to real-time feedback about decision accuracy. Environment feedback in response to user decisions will provide a sparse signal in some contexts. Users may still internally update their trust in the AI based on their confidence in their own decision and their (potentially incorrect) perception of their expertise in comparison to the AI assistant's ability. Additionally, some interventions may not be suitable for certain tasks; for example, LLM-generated explanation frequently known to hallucinate details, which may be undesirable in high-stakes applications.

**Can we model user trust?** In our setting, the AI assistant can observe the user's trust level after each interaction. In appendix D, we provide an analysis of several heuristic-based and learning-based trust models on their ability to predict user trust levels based on user-AI interaction history instead. We find that these heuristics only achieve moderate correlation with user-reported trust levels, and are especially poor at predicting moments of high and low user trust. Our results point to the challenging nature of modeling user trust, and the inadequacy of surface-level interaction features. Instead, user-specific features such as users' internal confidence, their prior experience with AI systems and task expertise may be better indicators of user trust.

**Is all inappropriate reliance *equally* bad?** Our formulation of `Total Inappropriate Reliance` treats both under- and over-reliance as equally undesirable. However, a company developing an AI assistant for clinicians may be more wary of clinicians over-relying on their assistant, which may leave them liable. The relative importance of mitigating under- and over-reliance can be quantified through a balancing utility function. Such a utility function can also be used as a reward signal for optimizing an intervention policy to signal when an intervention should be applied.

## 6 Conclusion and Future Directions

We explore the utility of *trust-adaptive interventions* for mitigating inappropriate reliance when users have low or high trust in AI assistants. We demonstrate that low and high levels of trust result in increased inappropriate reliance on AI recommendations for laypeople answering science questions and for doctors making medical diagnoses. We conduct controlled between-subjects studies and find that adaptively providing supporting explanations during low trust and counter-explanations during high trust reduces inappropriate reliance and improves users' decision accuracy. These findings generalize to decelerating interventions; forcing users to pause and deliberate before making their final decision helps reduce over-reliance.

Our findings present an initial exploration into adapting AI behaviors based on user trust levels. We adopted a simple thresholding criterion for deciding when to intervene, but more sophisticated criteria that also account for user and AI confidence may have potential. Further, rather than looking at whether user trust is too high or low, we can consider whether the trust is calibrated with the assistant's trustworthiness. High user trust may not be as undesirable when the AI is significantly more accurate than the user on the task. We hope our findings inspire the community to more closely consider the effect of user trust in user-AI interactions and the potential benefits of modeling and adapting to user trust levels.

## Limitations

Our user studies were performed using a simulated AI assistant, rather than a real system such as a Large Language Model. Similar to Buçinca et al. (2021); Dhuliawala et al. (2023), we use a simulated AI for the controllability of AI accuracy and confidence calibration. However, a simulated AI may be unrealistic compared to real AI systems that people use. For instance, since the answer correctness is decided independently for each problem, our AI assistant may provide contradicting predictions for near-identical problems.

Furthermore, the behavior of Prolific participants interacting with an AI in a user study may differ from users of a live system in a real-world scenario (McGrath, 1995), particularly due to misalignment of motivations (Deci et al., 1999). We test for the effect of two confounding variables: trust reporting and user experience with AI (Appendix E), finding that they do not affect user reliance behavior, but there may be other confounding variables we have not yet considered.

We conducted all our user studies with participants from the U.K. and U.S.A. who were fluent in English. Users from other countries and cultures may have different attitudes towards AI systems, and thus behave differently.

Finally, our findings would be more reliable if we could obtain data from more participants, more tasks and more interventions, however this is not possible due to financial restrictions and the dearth of professional doctors on the Prolific platform.

## Ethical Considerations

While modeling user trust can allow AI systems to help users overcome their cognitive biases, care must be taken that such user modeling is not taken advantage of to mislead or manipulate users. We emphasize that our trust-adaptive interventions are not intended to force or manipulate users into behaving a certain way, but to recognize when the user's trust may be hindering their reasoning. We do not condone use of such user modeling to manipulate users by misrepresenting the AI's beliefs or causing the user any distress.

People designing AI assistants with trust-adaptive interventions should ensure the interventions comply with the local ethical standards of the intended users. Further, we encourage promoting transparency and accountability by disclosing to users that the AI assistant is modeling user trust.

## Acknowledgments

We thank Rishi Vanukuru, Gregory Yauney, Sayan Ghosh and Jesse Zhang for their invaluable discussions and brainstorming, as well as Malihe Alikhani, Heather Culbertson, Suvodip Dey, Vardhan Dongre, Mert İnan, Sunnie Kim, Chaitanya Malaviya, Ana Marasović, Weiyan Shi, Anthony Sicilia, Swabha Swayamdipta, Dilek Hakkani-Tür and Gökhan Tür for their valuable feedback. We also thank Dhuliawala et al. (2023) for releasing code for their user study interface, and in particular Vilém Zouhar for his assistance with setting up the user study and data logging. We are also very grateful to Aisha Baptiste for her swift assistance in adding money to our Prolific account during the course of this project. Tejas Srinivasan was supported by the Amazon ML Fellowship from the USC-Amazon Center for Secure and Trusted Machine Learning.

## References

Fatemeh Alizadeh, Oleksandra Vereschak, Dominik Pins, Gunnar Stevens, Gilles Bailly, and Baptiste Caramiaux. 2022. Building Appropriate Trust in Human-AI Interactions. In *European Conference on Computer-Supported Cooperative Work (ECSCW)*.

Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction*.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *CHI conference on Human Factors in Computing Systems*.

Amie J Barda, Christopher M Horvat, and Harry Hochheiser. 2020. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making*.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *International Conference on Healthcare Informatics*.

Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*.

Chang Che, Zengyi Huang, Chen Li, Haotian Zheng, and Xinyu Tian. 2024. Integrating generative AI into financial market prediction for improved decision making. *Applied and Computational Engineering*.

Zeya Chen and Ruth Schmidt. 2024. Exploring a Behavioral Model of "Positive Friction" in Human-AI Interaction. In *International Conference on Human-Computer Interaction*.

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *CHI Conference on Human Factors in Computing Systems*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1*.

Karl de Fine Licht and Bengt Brülde. 2021. On Defining "Reliance" and "Trust": Purposes, Conditions of Adequacy, and New Definitions. *Philosophia*.

Edward L Deci, Richard Koestner, and Richard M Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A Diachronic Perspective on User Trust in AI under Uncertainty. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies*.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A New Dataset for Automatic Medical Diagnosis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me? Evaluating Machine Learning Interpretations in Cooperative Play. In *International Conference on Intelligent User Interfaces (IUI)*.

Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, Development and Evaluation of a Human-Computer Trust Scale. *Behaviour & Information Technology*.

Fateme Hashemi Chaleshtori, Atreya Ghosal, Alexander Gill, Purbid Bambroo, and Ana Marasovic. 2024. On evaluating explanation utility for human-AI decision making in NLP. In *Findings of the Association for Computational Linguistics: EMNLP*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Mert İnan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. 2025. Better Slow than Sorry: Introducing Positive Friction for Reliable Dialogue Systems. *arXiv preprint arXiv:2501.17348*.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a Dcience of human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Eunhae Lee. 2024. *The Power of Perception in Human-AI Interaction: Investigating Psychological Factors and Cognitive Biases that Shape User Belief and Behavior*. Ph.D. thesis, Massachusetts Institute of Technology.

John D Lee and Katrina A See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human factors*, 46.

Xiaolin Lin, Xuequn Wang, and Nick Hajli. 2019. Building e-commerce satisfaction and boosting sales: The role of social commerce trust and its antecedents. *International Journal of Electronic Commerce*.

Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024a. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. *arXiv preprint arXiv:2403.16812*.

Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024b. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *CHI Conference on Human Factors in Computing Systems*.

Joseph E McGrath. 1995. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in human–computer interaction*. Elsevier.

Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors*.

Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *CHI Conference on Human Factors in Computing Systems*.

Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*.

James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *International Conference on Intelligent User Interfaces (IUI)*.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on ai Advice: Conceptualization and the Effect of Explanations. In *International Conference on Intelligent User Interfaces (IUI)*.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large Language Models Help Humans Verify Truthfulness–Except When They Are Convincingly Wrong. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Judith Sieker, Simeon Junker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, and Sina Zarrieß. 2024. The Illusion of Competence: Evaluating the Effect of Explanations on Users' Mental Models of Visual Question Answering Systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. 2022. Uncalibrated Models can Improve Human-AI Collaboration. *Advances in Neural Information Processing Systems (NeurIPS)*.

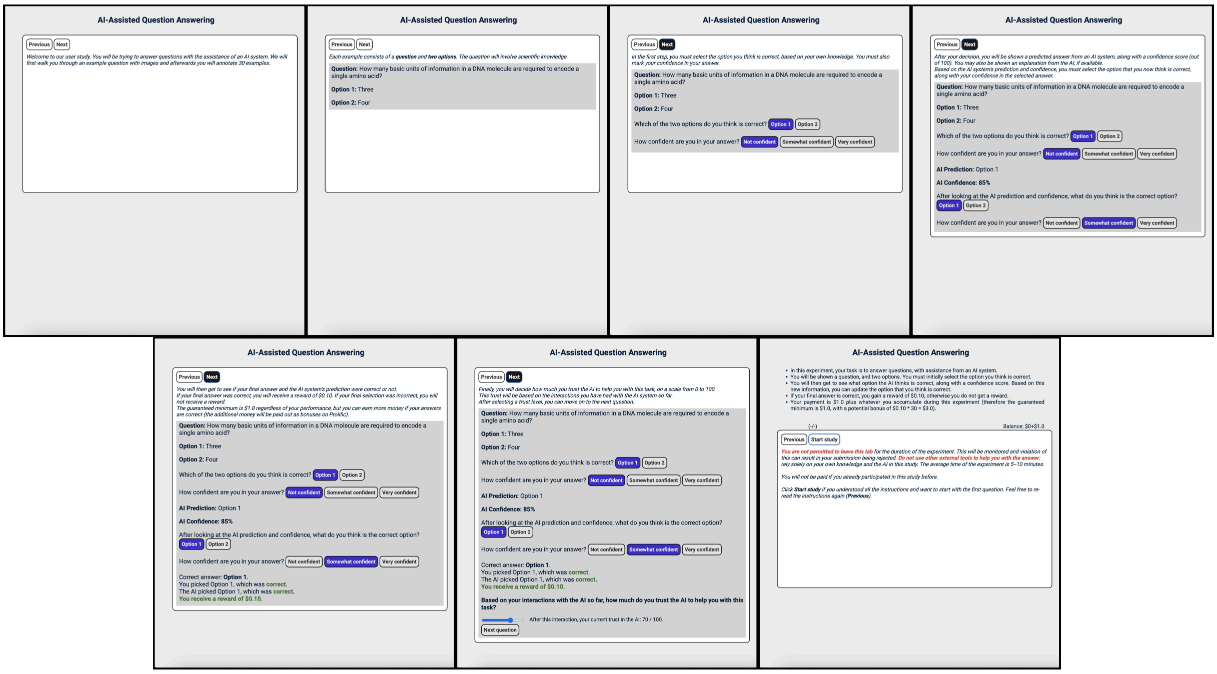
Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *International Conference on Intelligent User Interfaces (IUI)*.

Stephen Wright. 2010. Trust and trustworthiness. *Philosophia*.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.



AI-Assisted Question Answering

*Welcome to our user study. You will be trying to answer questions with the assistance of an AI system. We will first walk you through an example question with images and afterwards you will annotate 30 examples.*

AI-Assisted Question Answering

*Each example consists of a **question** and **two options**. The question will involve scientific knowledge.*

**Question:** How many basic units of information in a DNA molecule are required to encode a single amino acid?

**Option 1:** Three

**Option 2:** Four

AI-Assisted Question Answering

*In the first step, you must select the option you think is correct, based on your own knowledge. You must also mark your confidence in your answer.*

**Question:** How many basic units of information in a DNA molecule are required to encode a single amino acid?

**Option 1:** Three

**Option 2:** Four

Which of the two options do you think is correct? [✓ Option 1] Option 2

How confident are you in your answer? [✓ Not confident] Somewhat confident Very confident

AI-Assisted Question Answering

*After your decision, you will be shown a predicted answer from an AI system, along with a confidence score (out of 100). You may also be shown an explanation from the AI, if available.*
*Based on the AI system's prediction and confidence, you must select the option that you now think is correct, along with your confidence in the selected answer.*

**Question:** How many basic units of information in a DNA molecule are required to encode a single amino acid?

**Option 1:** Three

**Option 2:** Four

Which of the two options do you think is correct? [✓ Option 1] Option 2

How confident are you in your answer? [✓ Not confident] Somewhat confident Very confident

**AI Prediction:** Option 1

**AI Confidence: 85%**

After looking at the AI prediction and confidence, what do you think is the correct option?
[✓ Option 1] Option 2

How confident are you in your answer? Not confident [✓ Somewhat confident] Very confident

AI-Assisted Question Answering

*You will then get to see if your final answer and the AI system's prediction were correct or not.*
*If your final answer was correct, you will receive a reward of $0.10. If your final selection was incorrect, you will not receive a reward.*
*The guaranteed minimum is $1.0 regardless of your performance, but you can earn more money if your answers are correct (the additional money will be paid out as bonuses on Prolific).*

**Question:** How many basic units of information in a DNA molecule are required to encode a single amino acid?

**Option 1:** Three

**Option 2:** Four

Which of the two options do you think is correct? [✓ Option 1] Option 2

How confident are you in your answer? [✓ Not confident] Somewhat confident Very confident

**AI Prediction:** Option 1

**AI Confidence: 85%**

After looking at the AI prediction and confidence, what do you think is the correct option?
[✓ Option 1] Option 2

How confident are you in your answer? Not confident [✓ Somewhat confident] Very confident

Correct answer: **Option 1**.
You picked Option 1, which was correct.
The AI picked Option 1, which was correct.
**You receive a reward of $0.10.**

AI-Assisted Question Answering

*Finally, you will decide how much you trust the AI to help you with this task, on a scale from 0 to 100.*
*This trust will be based on the interactions you have had with the AI system so far.*
*After selecting a trust level, you can move on to the next question.*

**Question:** How many basic units of information in a DNA molecule are required to encode a single amino acid?

**Option 1:** Three

**Option 2:** Four

Which of the two options do you think is correct? [✓ Option 1] Option 2

How confident are you in your answer? [✓ Not confident] Somewhat confident Very confident

**AI Prediction:** Option 1

**AI Confidence: 85%**

After looking at the AI prediction and confidence, what do you think is the correct option?
[✓ Option 1] Option 2

How confident are you in your answer? Not confident [✓ Somewhat confident] Very confident

Correct answer: **Option 1**.
You picked Option 1, which was correct.
The AI picked Option 1, which was correct.
**You receive a reward of $0.10.**

**Based on your interactions with the AI so far, how much do you trust the AI to help you with this task?**

After this interaction, your current trust in the AI: 70 / 100.

AI-Assisted Question Answering

- In this experiment, your task is to answer questions, with assistance from an AI system.
- You will be shown a question, and two options. You must initially select the option you think is correct.
- You will then get to see what option the AI thinks is correct, along with a confidence score. Based on this new information, you can update the option that you think is correct.
- If your final answer is correct, you gain a reward of $0.10, otherwise you do not get a reward.
- Your payment is $1.0 plus whatever you accumulate during this experiment (therefore the guaranteed minimum is $1.0, with a potential bonus of $0.10 * 30 = $3.0).

(-/-) Balance: $0+$1.0

***You are not permitted to leave this tab** for the duration of the experiment. This will be monitored and violation of this can result in your submission being rejected. **Do not use other external tools to help you with the answer**: rely solely on your own knowledge and the AI in this study. The average time of the experiment is 5--10 minutes.*

*You will not be paid if you already participated in this study before.*

*Click **Start study** if you understood all the instructions and want to start with the first question. Feel free to re-read the instructions again (**Previous**).*

Figure 9: Screenshots of task instructions shown to users.

# A Task Details

We conduct user studies on two decision-making tasks: the **ARC** task and the **Diagnosis** task.

**ARC Task.** This task consists of answering science questions. Questions are sourced from the ARC dataset (Clark et al., 2018), which consists of more than 7000 grade-school science multiple-choice questions written for examinations. The authors manually reviewed questions from this dataset and selected questions that were challenging (i.e. the correct answer is not immediately obvious, and at least one option was not obviously incorrect) but still understandable (did not contain any scientific jargon that laypeople may not be familiar with). All questions in the original dataset had four options, but the author only selected the correct answer and the most plausible incorrect option for the decision-making problem. The final filtered set consisted of 39 questions. For each of the 10 problem sequences that users solve, we sample 30 of the 39 questions without repetition.

**Diagnosis Task.** This task consists of diagnosing patients based on patient symptoms. Patient symptoms are sourced from the DDXPlus dataset (Fansi Tchango et al., 2022), which contains 1.3 million synthetic patients with a differential diagnosis. The symptoms are presented as either binary, categorical or continuous variables, but each symptom has a corresponding patient intake question and answer in English, which we translate into a descriptive third-person statement using GPT-4o.. For example, the intake question "Do you feel pain somewhere?" and patient response "Knee (R)" is translated into "The patient feels pain in their right knee.". We filter down to questions with only 10–15 intake responses so that users do not need to spend a long time understanding the problem. To convert the task into a multiple-choice problem, we select the top three negative conditions from the differential diagnosis as the incorrect options. Our final set of problems includes 55 cases corresponding to eleven different conditions, which we use to sample 10 sequences of 30 problems each.

Table 6 contains examples of decision-making problems from both tasks.

# B User Study Details

We present additional details about the user studies.

**User Payment.** As mentioned earlier, users in the ARC task are paid a base payment of $1.0, with an incentive of $0.10 per correct answer. Users achieved 65–75% accuracy on the task, which translates to a bonus of ≈ $2.0 per user, or $3.0 total payment. The tasks took a median time of 15 minutes to complete, which translates to a pay rate of $12 per hour.

For the ARC task, users are paid a higher base payment of $2.0, since the task takes slightly longer

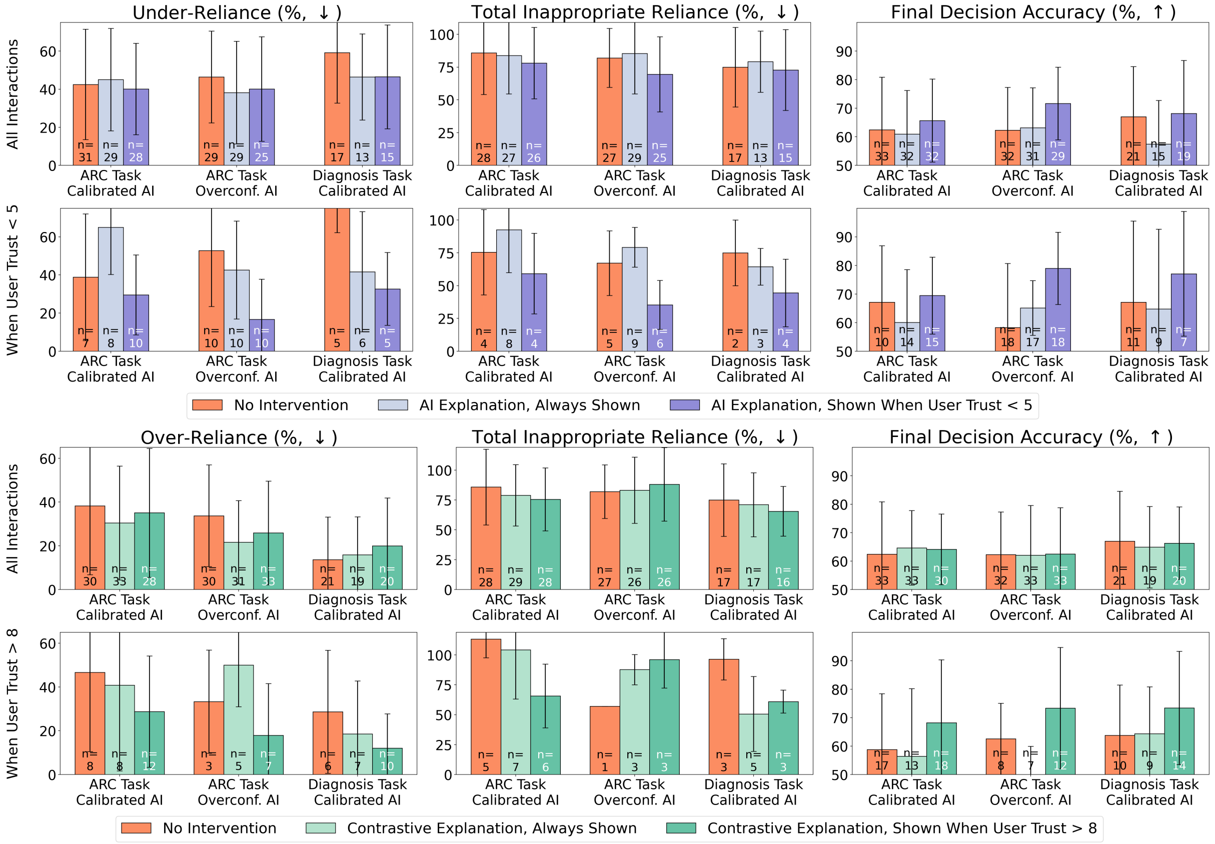


Figure 10: Macro-aggregation results for mitigating under- and over-reliance. $n$ represents the number of users who have at least 3 interactions meeting the required criteria.

to complete (≈ 20 minutes). Users achieve 75% accuracy, translating to an average bonus of \$2.0 per user, or \$4.0 total payment, which translates to a pay rate of \$12 per hour.

In total, including Prolific fees, we spent ≈ \$2500 on the user studies reported in this paper.

**User Demographics.** Users were recruited on the Prolific platform. We recruited participants from the U.K. and U.S.A. who self-identified as fluent in English, and had at least 99% approval rate on previously completed Prolific studies.

**Instructions.** Users were informed that they were participating in a research study. Figure 9 contains screenshots of the task instructions and payment details presented to users.

Our user study was approved by our institution's Institutional Review Board (camera-ready version will include the name of the institution).

## C Macro-Aggregation Results

We report macro-aggregation results, where we first compute our metrics for each user, and then aggregate across all users. Because some users may only have a few interactions that fit the required criteria for computing metrics, the metrics for those users are very sensitive to a single interactions. Therefore, when computing metrics, we filter out users with fewer than 3 interactions that meet the required criteria ($y_i^{AI} \neq y_i^{u,\text{init}}$, and user trust being below/above the corresponding threshold for the "low trust"/"high trust" interaction subsets). For the `Under-Reliance` metric we also need at least 3 interactions where the AI prediction is correct, and for the `Over-Reliance` metric we need at least 3 interactions where the AI prediction is incorrect. For the `Total Inappropriate Reliance` metric we *both* need 3 interactions where the AI prediction is correct and 3 interactions where the AI is incorrect (so that we can calculate both `Under-Reliance` and `Over-Reliance` for that user).

In Figure 10, we observe trends similar to those observed in Section 4. However, when looking at the low/high trust subsets, most conditions have fewer than 10 users for calculating `Under-Reliance` and `Over-Reliance`, and fewer than 5 users for `Total Inappropriate Reliance`.

| Trust estimator | Train corr. | Test corr. | High-trust F1 | Low-trust F1 |
|---|---|---|---|---|
| AIAcc5 | 0.388 | 0.345 | 0.345 | 0.458 |
| CapabilityDiff | 0.185 | 0.248 | 0.382 | 0.500 |
| SmoothOutcomes | 0.466 | 0.434 | 0.495 | 0.447 |
| SmoothConfs | 0.483 | 0.430 | 0.486 | 0.423 |
| TrustEffectModel | 0.478 | 0.509 | 0.392 | 0.498 |

Table 1: Correlation of trust estimation methods with user-reported trust levels, and F1 for detecting moments of low and high trust.

## D Trust Modeling

In this section, we attempt to model user trust directly based on user-AI interaction history, and highlight the difficulties of simple trust heuristics at identifying moments of low and high user trust.

We experiment with several simple heuristics that are computed by looking at interaction outcomes. **AIAcc5** estimates trust based on the AI's accuracy in the last 5 interactions. **CapabilityDiff** calculates the difference between the accuracy of the AI's prediction and the user's initial decision, over all interactions so far. **SmoothOutcomes** updates the trust $\tau_t \in [-1, 1]$ after interacton $t$, based on the AI's prediction correctness $a_t \in \{0, 1\}$

$$\tau_0 = 0; \tau_t = r \cdot (2 * a_t - 1) + (1 - r) \cdot \tau_{t-1}$$

where $r$ is a smoothing parameter. **SmoothConfs** is similar to the above, except the update term is weighted by the AI's confidence $c_t^{AI}$

$$\tau_0 = 0; \tau_t = r \cdot (2 * a_t - 1) \cdot c_t^{AI} + (1 - r) \cdot \tau_{t-1}$$

Finally, we train a **TrustModel**, a linear regression model which estimates the change in user trust after each interaction.

We use a set of 45 additional user sessions as training data for the model and selecting smoothing parameters, and evaluate on a held-out set of 30 user sessions, totaling 1350 and 900 user-AI interactions in train and test set. We evaluate the trust estimation methods on their correlation with user-reported trust levels for the test set interactions, and F1 for predicting low and high trust.

Table 1 shows correlation between trust scores estimated by the various heuristics and the user-reported trust levels, on both train and tests, along with the F1 for detecting moments of low and high user trust. We observe that all of the above methods achieve only moderate correlation with user-reported trust, and achieve very low F1 ($\leq 0.5$) at detecting both low and high trust instances. These results point to the challenging nature of modeling user trust, and the inadequacy of surface-level interaction features.

## E Testing for Confounding Variables

| Users Report Trust? | Final Decision Accuracy | Switch Rate | Under-Reliance | Over-Reliance |
|---|---|---|---|---|
| ✓ | 61.9% | 48.6% | 40.7% | 34.8% |
| × | 60.2% | 49.9% | 42.0% | 37.2% |

Table 2: Effect of asking users to report trust on reliance metrics and decision accuracy, on the ArcC task setting.

We evaluate the effect of asking users to report trust on their decision-making performance and reliance on the AI assistant, finding that user performance stays largely the same (Table 2).

| How often user uses AI | Switch Rate | Under-Reliance | Over-Reliance |
|---|---|---|---|
| Once a week or more | 43.4% | 43.5% | 29.9% |
| About once a month | 41.7% | 41.4% | 28.9% |
| Never | 41.8% | 44.8% | 26.9% |

Table 3: Effect of user experience with AI on reliance metrics, on the ArcC task setting.

| How often user uses AI for checking facts | Switch Rate | Under-Reliance | Over-Reliance |
|---|---|---|---|
| Regularly | 41.9% | 46.7% | 30.1% |
| Have used, not regularly | 45.4% | 39.3% | 29.9% |

Table 4: Effect of user experience with AI for fact checking on reliance metrics, on the ArcC task setting.

Table 3 and Table 4 show how user experience with AI, and specifically using it for fact checking, affects user reliance on AI assistance, again finding that reliance behavior is similar.

| Supporting explanation generation prompt | Counter-explanation generation prompt |
|---|---|
| Provide a convincing 3-4 sentence explanation (under 100 words) for why option $\{i\}$ is the correct diagnosis, even if you don't believe option $\{i\}$ is the correct diagnosis. It is absolutely essential that you justify that option $\{i\}$ is the correct diagnosis, rather than any of the other options.<br>Do not explicitly refer to any options by their option number, only their content. For example, if you are explaining why option 1: Anemia is correct, do not say "Option 1 is correct because..." or "For the above reasons, the correct diagnosis is option 1"), instead say "Anemia is the correct diagnosis because...". | Provide a convincing 3-4 sentence explanation (under 100 words) for why option $\{i\}$ may not be the correct diagnosis, even if you believe that option $\{i\}$ is the correct diagnosis. Your explanation should reflect that you believe option $\{i\}$ is correct (e.g. by saying "While I think <option $\{i\}$> is the correct diagnosis..."), but also provide 1-2 reason(s) for why it may not be correct. Your explanation should not merely restate the option, but provide a (short) reasoning. Do not explicitly refer to any options by their option number, only their content. For example, if you are explaining why option 1: Anemia is correct, do not say "Option 1 is correct because..." or "For the above reasons, the correct diagnosis is option 1"), instead say "Anemia is the correct diagnosis because...". |

Table 5: Prompts used for generating supporting and counter-explanations for the Diagnosis task. Prompts for ARC task were slightly modified (e.g. replacing "correct diagnosis" with "correct answer").

| Task | Question | Options | Supporting Explanation | Counter-Explanation |
|---|---|---|---|---|
| ARC | Water has many unique chemical properties. Which property of water makes water a good solvent of crystalline salts? | Low Ph<br>**Strong polarity** | Water's strong polarity allows it to interact with the charged ions in crystalline salts, breaking them apart and dissolving them. This is because water molecules are polar, with oxygen being more electronegative than hydrogen, creating a partial negative charge near the oxygen and partial positive charges near the hydrogen atoms. This polarity allows water molecules to surround and solvate the ions in the salt, leading to the dissolution of the salt in water. | Although strong polarity is the primary reason water is a good solvent, low pH could potentially enhance the solubility of certain crystalline salts by increasing the availability of hydrogen ions to interact with the salt molecules. |
| Diagnosis | Patient is a 7 year old male.<br>- The patient lives with 4 or more people.<br>- The patient attends or works in a daycare.<br>- The patient has pain somewhere, related to their reason for consulting.<br>- The patient's pain is sensitive.<br>- The patient feels pain in the right tonsil, under the jaw, and in the trachea.<br>- The intensity of the patient's pain is 3 (on a scale of 0 to 10).<br>- The patient's pain does not radiate to another location.<br>- The patient's pain appeared at a speed of 2 (on a scale of 0 to 10).<br>- The patient has a fever, either felt or measured with a thermometer.<br>- The patient has had a cold in the last 2 weeks.<br>- The patient has noticed that the tone of their voice has become deeper, softer, or hoarse. | - Epiglottitis<br>- Viral pharyngitis<br>- **Acute laryngitis**<br>- Chagas | Acute laryngitis is the correct diagnosis because the patient exhibits hoarseness and a change in voice tone, which are hallmark symptoms of laryngitis. The presence of a recent cold and mild pain in the throat area align with viral-induced laryngitis, typically following upper respiratory infections. The patient's fever and pain in the right tonsil and trachea further support inflammation consistent with laryngitis. | While I think acute laryngitis is the correct diagnosis, it may not be entirely accurate because the patient's pain is specifically located in the right tonsil, under the jaw, and in the trachea, which might also suggest a more localized infection like tonsillitis or pharyngitis. Additionally, acute laryngitis typically involves a more pronounced hoarseness of voice, which may not align perfectly with the patient's description of a deeper or softer tone, potentially pointing towards other conditions. |

Table 6: Examples of supporting and counter-explanations generated by GPT-4o. The AI prediction (which is also the correct option) is highlighted.