Eyes are the Windows to AI Reliance: Toward Real-Time Human-AI Reliance Assessment

SHIYE CAO, Johns Hopkins University, USA
SHICHANG KE[*], Johns Hopkins University, USA
YANYAN MO[*], Johns Hopkins University, USA
ANQI LIU, Johns Hopkins University, USA
CHIEN-MING HUANG, Johns Hopkins University, USA

Appropriate user reliance on artificial intelligence (AI) is fundamental to achieving synergistic human-AI collaboration; however, human users frequently struggle with reliance calibration. In this paper, we conduct a statistical analysis to explore how AI performance affects users' reliance across trials and build machine learning models for earlier prediction of user reliance. Our results show that AI performance has varying effects on different proxies of user reliance, suggesting that nuanced differences exist between *task-based* (human-AI agreement), *gaze-based* (gaze duration on an AI suggestion), and *perception-based* (perceived reliance) proxies of reliance. Moreover, we model gaze behavior, task behavior, and user demographics for reliance prediction at various points in time before the final human-AI team decision; our models demonstrate reliable estimation of the task- and gaze-based user reliance on AI halfway through the task while also achieving generalizability for new users. Our work indicates the possibility of real-time human-AI reliance assessment to facilitate adaptive reliance calibration.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: human-AI interaction, appropriate reliance, eye gaze, decision-making

ACM Reference Format:
Shiye Cao, Shichang Ke, Yanyan Mo, Anqi Liu, and Chien-Ming Huang. 2023. Eyes are the Windows to AI Reliance: Toward Real-Time Human-AI Reliance Assessment. In *CHI TRAIT '23: Workshop on Trust and Reliance in AI-Assisted Tasks at CHI 2023, April 23, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 MOTIVATION AND BACKGROUND

Proper reliance calibration is hard for users in human-AI teaming. Artificial intelligence (AI) systems are increasingly being developed to assist humans in image analysis tasks such as pothole detection [17, 24] and diagnostic radiology [11–13], colonoscopy [31], and dermatology [8, 15]. The goal of human-AI teamwork is to enhance the human performance of such tasks, as humans and AI are presumed to have areas of complementarity in decision-making [8, 28]. Therefore, successful human-AI collaboration requires users to know when and how much to rely on themselves as opposed to the AI assistant [22, 23, 37], given both human decisions and AI predictions are not and will never be perfect. However, this

[*] Both authors contributed equally to this research.

imposes significant challenges for users to calibrate their reliance as they tend to lack awareness of both their and AI's capabilities [23, 26].

Prior work has identified several approaches to facilitating users in determining how much to rely on an AI suggestion. Commonly practiced techniques include presenting the model uncertainty on the prediction (i.e., the model's uncalibrated/calibrated softmax output [30, 34, 37]) or local explanations for the prediction [3, 18, 20, 21]. Presentation of other model information, such as global explanations [27], model performance information [30, 36], and model error boundary [2], has been experimented with in hopes of helping users build better mental models of AI capabilities. Simple cognitive forcing functions (i.e., delaying the display of AI recommendations [23, 25] or only presenting AI suggestions upon request [5]) have also been explored to help directly reduce inappropriate user reliance on AI recommendations. Applications of these existing reliance calibration supports are often rigid and do not account for users' individual differences and changes in user status over time. This is because retrospective quantifications of user reliance limit existing approaches. Common metrics of user reliance on AI (*i.e.,* human-AI agreement [20, 22, 33], adoption of AI recommendation [19, 33, 36], and self-reported perceived reliance [19, 32, 33]) require user input and thus are typically only collected after the final human-AI team decision is made. Thus, these reliance measures can only inform the implementation of interventions for future collaborations and not the current collaboration [24].

Contrastingly, eye gaze tracking provides objective and quantitative evidence of users' visual and attentional processes, allowing us to gain insights into user decision-making processes and intentions without interrupting the user workflow to ask for their input [7, 9]. Prior work has discovered a strong positive correlation between user gaze duration on an AI suggestion (a *gaze-based* measure of reliance) and human-AI agreement (a *task-based* measure of reliance), as well as perceived reliance (a *perception-based* measure of reliance) [6]. This relationship between eye gaze and user reliance opens the opportunity for real-time assessment of AI reliance via gaze monitoring. Furthermore, prior work in human-computer interaction has explored the use of gaze tracking as an input source to estimate user status and intention and additionally to adapt user interfaces based on user behavior [10]; our ultimate goal is to use gaze tracking as an input source to estimate user reliance on an AI's suggestions in real-time and adaptively adjust the choice of reliance calibration support strategy based on this estimation.

Using data collected in prior work [6] through a user study contextualized in a spatial reasoning task, we conducted analyses to understand 1) how an AI's performance shapes user reliance across trials; 2) the differences between task-, gaze-, and perception-based characterizations of user reliance on AI; and 3) the possibility of assessing user reliance on an AI before a final human-AI team decision is made. Our analyses revealed that 1) the effect of AI performance on user reliance differed across users and different proxy measures of reliance; 2) nuanced differences exist between task-based (human-AI agreement), gaze-based (gaze duration on AI), and perception-based (user perception) measures of reliance; and 3) task- and gaze-based reliance measures can be predicted reliably as early as halfway through the total task time. Our findings indicate the potential for real-time assessment of human-AI reliance, which in turn can help inform the selection of interventions to facilitate user reliance calibration.

2 DATA

Data were obtained from a previous user study with AI performance as a between-subjects factor. The data contained participants' task behavior and recordings of their eye gaze during the experiment, collected via the Gazepoint GP3 remote eye tracker (60 Hz). 347 trials from 35 participants were used in the analysis below (10 trials from each participant, excluding the corrupted gaze data present in three trials from two participants). Among the 35 participants, 14 were assigned to a high-performance AI, 10 to a medium-performance AI, and 11 to a low-performance AI. Participants

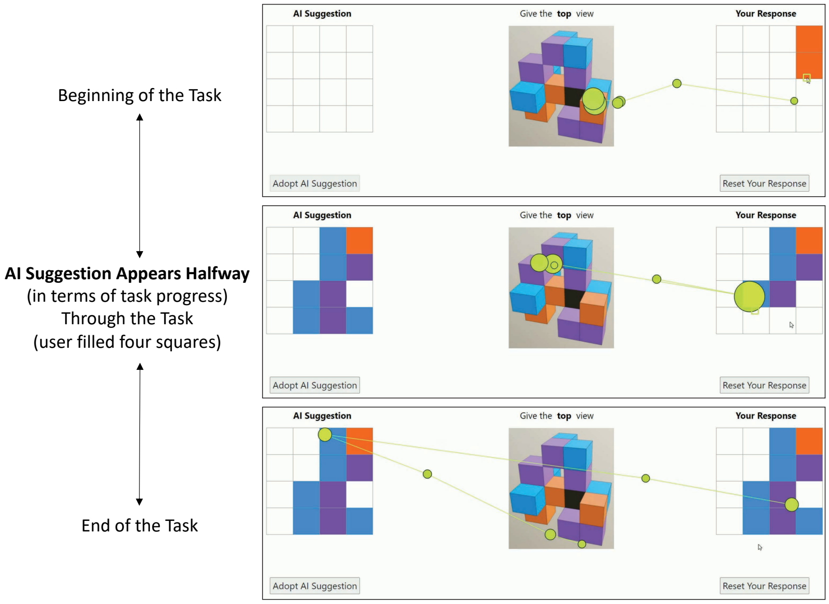

Fig. 1. Overview of the experimental task. The AI suggestion appears when the user is about halfway through the task in terms of task progress. Three screenshots from three phases of the task are shown overlaid with their associated real-time eye gaze denoted in green. Each green circle shows the location of gaze fixation; the circle's size represents the length of the fixation at that location. This figure is reproduced from prior work [6].

(16 female, 19 male) were recruited through convenience sampling from the local community and randomly assigned to a particular AI accuracy level by the system; neither the participant nor the experimenter was made aware of the accuracy level of the AI. More details on the participants and an in-detail description of the data collection procedure can be found in the original study from which these data were obtained [6].

2.1 Task Description

Participants collaborated with an AI on a visuospatial task in which they are asked to apprehend the top or bottom view of a three-dimensional block structure based on a two-dimensional image slice taken from a side view (see Figure 1 for a sample task trial). Task difficulty varied across trials; each participant completed five trials providing the top view of the structure (easier; one degree of mental rotation needed) and five trials providing the bottom view of the structure (harder; two degrees of mental rotation needed). The order in which the trials appeared was randomized.

2.2 AI Recommendation Generation

AI performance was manipulated manually via the accuracy of its recommendations. The high-performance AI had 100.00% accuracy in all its recommendations; the medium-performance AI had 93.75% accuracy in 40% of its recommendations and 100.00% accuracy in the remaining 60% of its recommendations; and the low-performance AI had between 31.25% and 75% accuracy in all its recommendations (averaging 49.38% accuracy overall).

An AI suggestion was shown to the user after they entered a response for four different squares on the user response grid, regardless of the correctness (color and location) of the squares they selected. As the correct answer for the task always contained nine squares, four squares corresponded to the halfway point in completing the task. (Note that the first block in the top-right corner was always provided; see Figure 1 for an example.) Participants were not told how many blocks were in the final response, nor when the AI's suggestion would appear.

3 PROXY METRICS FOR USER RELIANCE

Drawn from prior research, we employed three metrics to evaluate user reliance on AI assistance:

- *Human-AI Agreement* (Range: 0–1). Commonly used in prior work (e.g., [19, 33]), this metric is a **task-based proxy of user reliance on AI**. It was computed as the number of squares in the user response that matched the AI suggestion divided by the total number of squares (16); i.e., if the user response matched the AI suggestion exactly in a particular trial, then the human-AI agreement value for that trial was 1.00.
- *Perceived Reliance* (Range: 1–7). Also commonly used in prior work (e.g., [19, 33]), this metric is a **perception-based proxy of reliance on AI**. Participants rated their level of agreement with the statement "I relied on the AI suggestion in the previous task" on a 7-point Likert scale, with 1 being "strongly disagree" and 7 being "strongly agree."
- *Gaze Duration on AI* (Range: 0–1). This metric is a **gaze-based proxy of user reliance on AI**, computed as the amount of time a participant spent fixating on the AI's suggestion divided by the total time the participant spent fixating on the task (including the fixation on the AI suggestion, the task image, and the response grid). Prior work has shown that users' human-AI agreement is highly correlated with their gaze duration on an AI's input [6].

4 THE DYNAMIC NATURE OF USER RELIANCE

We observed that even though no AI-recommendation-related information was provided to support user reliance calibration, user reliance on the AI fluctuated from trial to trial (see Figure 2). We analyzed the effect of AI performance on the variance in user reliance across trials measured by standard deviation. For the analyses reported below, if not otherwise specified, we performed one-way analyses of variance (ANOVA) with participant ID as a random effect. All post-hoc pairwise comparisons were conducted using Tukey's HSD test. We considered $p < .05$ as a significant effect.

4.1 Results

A one-way ANOVA revealed a significant main effect of AI performance on the standard deviation of participants' task-based reliance, $F(2, 32) = 9.71, p < .001$. Pairwise comparisons using Tukey's HSD test revealed that participants paired with high-performance AI had a significantly lower spread in their task-based reliance ($M = 0.05, SD = 0.08$) than participants paired with medium-performance AI ($M = 0.14, SD = 0.11$), $p = .019$. Furthermore, participants with high-performance AI had a significantly lower variance in their task-based reliance than participants with low-performance AI ($M = 0.18, SD = 0.03$), $p < .001$. No significant differences were observed between the variance in task-based reliance among those with medium-performance AI and those with low-performance AI.

A one-way ANOVA revealed no significant effect of AI performance on the standard deviation of participants' gaze duration on the AI suggestion, $F(2, 32) = 0.04, p = .964$.

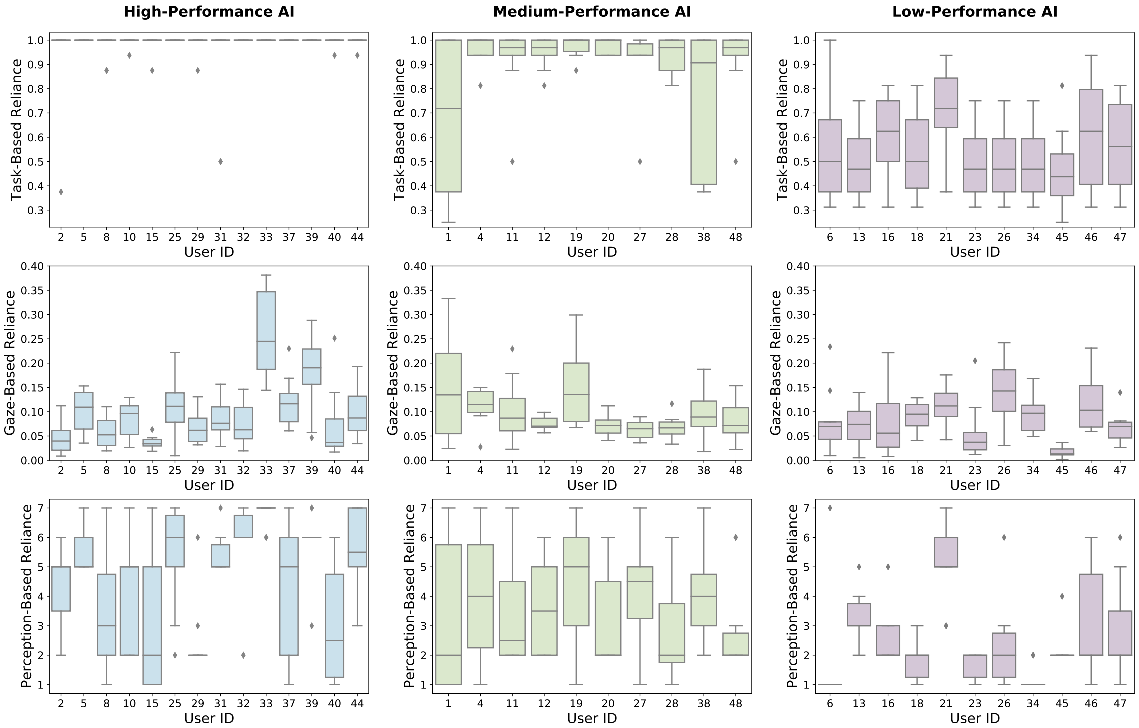

Fig. 2. Box plots demonstrating the effects of AI performance on the dynamics of user reliance for each user across all their trials. User reliance is characterized by three metrics: task-based human-AI agreement, gaze-based gaze duration on the AI's suggestion, and perception-based perceived reliance.

A one-way ANOVA revealed a significant effect of AI performance on the standard deviation of participants' perception-based reliance on AI, $F(2, 32) = 6.24, p = .005$. Pairwise comparisons using Tukey's HSD test showed that participants with medium-performance AI had a significantly higher variance in their perception-based reliance ($M = 1.92, SD = 0.32$) than participants with low-performance AI ($M = 1.13, SD = 0.55$), $p = .004$. No significant differences were observed between the variances in perception-based reliance of each user among those with high-performance AI and those with medium-performance and low-performance AIs.

4.2 Discussion

Through these analyses, we found that, in general, the impact of AI performance on the distribution of user reliance varied across the three measures. AI performance significantly affected the spread of human-AI agreement and perceived reliance, but not participants' gaze duration on the AI's suggestion. On the other hand, working with the high-performance AI led to a smaller variance in users' human-AI agreement but a greater variance in their perceived reliance on the agent's suggestions; this shows that **fundamental differences exist between these three measures of reliance** and further demonstrates the complexity of the construct of reliance overall.

Additionally, **we observed individual differences in how user reliance was shaped by AI performance**. For instance, Users 1 and 38 varied in their level of agreement with the medium-performance AI much more than other users paired with the same agent; User 45 varied their gaze duration on the low-performance AI's suggestions much less than other users paired with the same agent; and Users 31 and 32 varied in their perceived reliance on the high-performance AI much less than compared to other users paired with the same agent.

We observed that users often shifted their reliance trial by trial (See Figure 2). However, opportunities remain for the user reliance on AI to be better calibrated to help achieve more optimal human-AI team task performance (see Appendix D for more details). Prior works have examined factors that may cause such shifts in user reliance, such as task-related factors including task difficulty [6] and user-experience-related factors, such as the user perceived AI performance [1, 36]. Additionally, interventions (*e.g.,* cognitive forcing functions [5, 29] and adding model explanations [3]) have been found to help adjust user reliance on an AI successfully. However, individual differences exist in users' reactions to these interventions [23]. Numerous factors (*i.e.,* users' varying familiarity with AI, level of expertise in the task, age, gender, trust in the AI, perception of task criticality, and so on) can contribute to the users' individual differences, making it difficult to predict the effectiveness of an intervention on user reliance using a limited set of factors. Instead, we propose real-time assessment of human-AI reliance so that the effectiveness of an intervention for appropriate reliance can be evaluated *before* the user reaches a final decision. This would allow system designers to adjust interventions based on a user's status in the middle of the task. In the following section, we test whether we can reliably predict ultimate user reliance on an AI during a task and, if so, how early that prediction can be made.

5 MODELING USER RELIANCE IN "REAL TIME"

We trained models using input features extracted from 0%–100% of the task time to predict final user reliance on the AI within the task context. In this section, we focus on the prediction of task- and gaze-based reliance; please refer to Appendix B for more details on our attempt to predict perception-based reliance (the features considered had little influence on perception-based reliance).

5.1 Model Training

We considered a combination of features extracted from users' gaze behavior (gaze duration and shift in gaze fixation), task behavior (initial human-AI agreement and task difficulty), and demographics (pre-study trust in AI and pre-study familiarity with AI) as our models' input. Please refer to Appendix A.1 for a full list of the features considered.

We trained stepwise multiple linear regression models to model task- and gaze-based measures of reliance with backward elimination as the stepwise method [16]; likelihood ratio as the selection criterion [35]; and $p < .25$ as the stop criterion [35]. For more details on the feature selection process, please refer to Appendices A.2 and A.3.

Interestingly, while there were some overlapping features, significant features for gaze-based reliance were partially different from those of task-based reliance. For task-based reliance, the following features were significant: gaze duration on the user response grid, gaze shifts from the user response grid to the AI suggestion, gaze shifts to the AI suggestion, task duration, task difficulty, initial human-AI agreement, and familiarity with AI. For gaze-based reliance, the identified features included: gaze duration on the task image, gaze shifts from the AI suggestion to the task image, gaze shifts from the AI suggestion to the user response grid, gaze shifts from the task image to the user response grid, gaze shifts from the task image to the AI suggestion, gaze shifts from the user response grid to the AI suggestion, gaze shifts from the user response grid to the task image, and familiarity with AI. For more details on the feature selection process, please refer to Appendix A.

With these sets of features identified from our feature selection process using data from all the trials, we trained a model per time stage (0%–100%) for each of the two reliance measures. Input features were generated using behavior

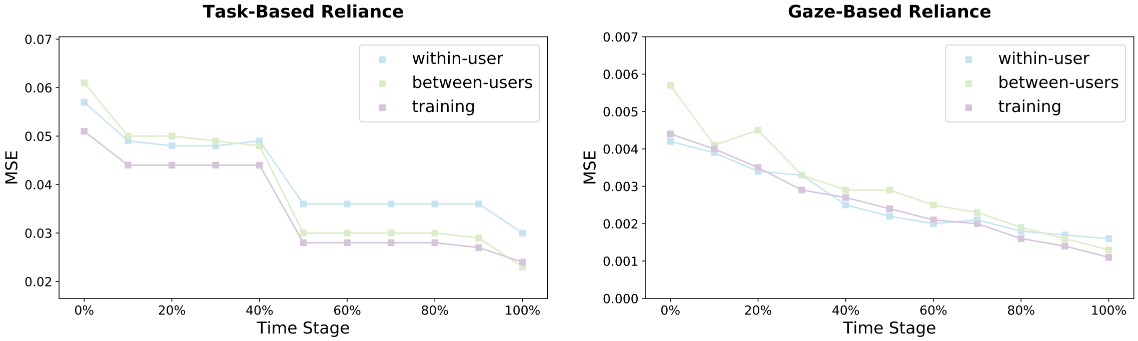

Fig. 3. Left: A line plot showing the mean squared error of the model predicting a task-based measure of reliance (human-AI agreement) throughout the task at different time stages (0%–100%). Right: A line plot showing the mean squared error of the model predicting a gaze-based measure of reliance (duration on AI) throughout the task at different time stages (0%–100%). A smaller mean square error signifies better model performance.

data only from up until that time stage; for instance, models at the 0% time stage only included user demographics, whereas models at the 50% time stage included user demographics, task behavior up to 50% task time, and eye gaze movement recorded from within the first half of the task.

Table 1. Two linear regression models predicting task- and gaze-based reliance halfway through the task. Please refer to Appendix A.1 for descriptions of the features used. β = regression coefficient, SE = standard error, t = t-value, p = probability of committing a Type I error.

Modeling Task-Based Reliance at 50% Task Time

Features	β	SE	t	p
Task duration (50%)	-0.25	0.04	-5.61	<.001
Familiarity with AI	-0.08	0.04	-2.07	.039
Gaze duration on user	0.00	0.05	0.03	.975
Gaze shifts (user to AI)	-0.08	0.06	-1.27	.204
Gaze shifts onto AI	0.20	0.07	2.74	.006
Initial human-AI agreement	0.56	0.04	13.07	<.001
Task difficulty	-0.06	0.04	-1.54	.124

Modeling Gaze-Based Reliance at 50% Task Time

Features	β	SE	t	p
Familiarity with AI	-0.12	0.04	-2.91	.004
Gaze shifts onto task	0.04	0.05	0.67	.501
Gaze shifts (AI to task)	0.03	0.11	0.25	.805
Gaze shifts (AI to user)	0.01	0.08	0.12	.902
Gaze shifts (task to user)	-0.10	0.10	-1.02	.306
Gaze shifts (task to AI)	0.31	0.10	3.18	.002
Gaze shifts (user to AI)	0.38	0.09	4.18	<.001
Gaze shifts (user to task)	-0.06	0.11	-0.53	.594

5.2 Model Evaluation

We evaluated the generalizability of our models using two methods: 1) *between-users* testing, where we tested the models on 90 trials from nine new users; and 2) *within-users* testing, where we tested the models on 70 unseen trials from existing users (two trials from each user). We evaluated the task- and gaze-based reliance models using mean square error (MSE). See Appendix C for model performance at varying time stages in the task.

We observed a trade-off between model performance and the earliest point at which to make the model prediction (see Figure 3); particularly, the MSE of the task-based measure of reliance prediction dropped to 0.036 in the within-users evaluation and to 0.031 in the between-users evaluation at 50% task time and remained stable until the end of the trial.

The models used to predict gaze-based reliance had very low MSE across all time stages; for comparison, MSE was 0.0022 in the within-users evaluation and 0.0029 in the between-users evaluation at 50% task time. Our model evaluation results illustrate that our two models (see model details in Table 1) are able to reliably predict task- and gaze-based reliance when a user is halfway through the task time and additionally are generalizable to new trials and users.

An in-depth look at the factors used by the 50% time models (Table 1) shows that task-based reliance is significantly higher when 1) task duration is shorter, 2) user familiarity with AI technology is lower, 3) gaze shifts to the AI suggestion are more frequent, and 4) initial human-AI agreement is higher. Gaze-based reliance is significantly higher when 1) user familiarity with AI technology is lower, 2) gaze shifts from the task image to the AI suggestion are more frequent, and 3) gaze shifts from the user response grid to the AI suggestion are more frequent. Surprisingly, model MSE was relatively low even without considering any behavioral features (at 0% time) for both predictions of the task- and gaze-based reliance. In other words, user pre-study familiarity with AI technology and task difficulty are highly correlated with task-based reliance.

6 DISCUSSION

6.1 Differences Between Three Proxies of Human-AI Reliance

Task-based, gaze-based, and perception-based proxies of reliance are known to be strongly correlated [6]. However, our analyses revealed that task-based, gaze-based, and perception-based reliance were impacted by different subsets of factors, indicating that they may be capturing different aspects of reliance (See Table 1 and Appendix B). Moreover, AI performance's effect on the variance of the three measures across trials also differed (See 4.1). AI performance only significantly affected task-based and perception-based reliance, but not gaze-based reliance; we speculate that this may be because users always looked at the AI suggestion regardless of their trust in or their intention to use or ignore the suggestion. Prior work found that while participants with low-performance AI often had lower gaze-based and perception-based reliance on AI, they still relied on the suggestions only for task structure orientation information, which was inferred from the suggestions, rather than directly adopting them [6]. In these cases, users' reliance on and trust in AI is more refined, and their reliance would likely only be captured by the gaze-based reliance metric (high gaze duration on AI since the user took time considering the AI suggestion) but not by task-based (low agreement since the user did not directly adopt AI suggestion) nor perception-based reliance (user might give low perceived reliance rating because they did not directly adopt AI suggestion). This shows the limitation of using one of the reliance proxies to capture user reliance.

Other limitations can also be identified from each of the task-based, gaze-based, and perception-based proxies of user reliance on AI. For instance, high human-AI agreement may be incidental and may not necessarily reflect high user reliance on the AI; the user may have completed the task while ignoring the AI assistance and happened to have reached the same conclusion as the agent. Moreover, perception-based reliance may be misleading at times, as humans are illogical and user perception is a reflection of their belief, which may not always match their behavior [4]; this could potentially explain why perception-based reliance was much more difficult to model as compared to task-based and gaze-based reliance. Lastly, gaze-based reliance may also be limited, as users may theoretically look at an AI suggestion due to *distrust* and decide to choose the answer opposite to the AI's recommendation; this would be the *opposite* of reliance, even though the user spent more time looking at the AI's suggestion.

In summary, each of the three proxy measures of reliance captures a different aspect of user reliance and has its specific limitations as a stand-alone measure of reliance. Therefore, future work should consider using a combination of these and possibly other metrics that may facilitate gaining a more holistic view of user reliance on AI.

6.2 Toward Real-Time Gaze-Aware Human-AI Reliance Assessment

Existing reliance calibration support is limited by its dependence on retrospective evaluations of user reliance. As user reliance on AI is dynamic and challenges remain for more appropriate reliance in users (See Appendix D), an adaptive reliance calibration support system would provide more flexibility in adjusting its interventions to real-time feedback. We speculate that a more flexible system would have a higher success rate in inducing appropriate reliance among its users since it would allow for selecting interventions that are more tailored to each user in real-time. However, such an adaptive system requires real-time estimation of user reliance so that the most appropriate intervention for the user at that time can be determined. In this study, we trained models to predict user reliance via gaze tracking before the final task decision was made; our models were able to reliably predict task-based and gaze-based reliance halfway through task time. Furthermore, we demonstrated that our model predictions are robust for new trials of existing users as well as new users. However, several more steps are necessary before the model can be used for real-time assessment of user reliance and to inform an adaptive reliance calibration support system.

Firstly, total task duration is not known during real-time assessments of reliance. Therefore, the proposed system must have the means to estimate task progress to determine when to use the trained models to assess user reliance; that is, the system must estimate when the user is roughly halfway through the task time to use the 50% task time model for reliance prediction. Estimating task progress is relatively easy in this task specifically, as we can judge task progress by how many task squares the users have interacted with in their response; however, task progress can be more difficult to judge in other, more realistic contexts, such as medical diagnosis.

Secondly, our models predict reliance rather than *inappropriate* reliance (e.g., over-reliance and under-reliance). Prediction of inappropriate reliance is difficult, as it is determined based on the ground truth of the task instance. High user reliance on an AI agent can be desirable if the AI recommendation is correct; however, this is undesirable when the AI recommendation contains errors. Prior works have used AI uncertainty (the correctness likelihood of the AI suggestion) to guide the selection of strategies to facilitate user reliance calibration [29]; for instance, if AI uncertainty is high, user decision time is extended to reduce user anchoring bias on the AI recommendation. Recent work takes into account both the predicted correctness likelihood of the user response and the predicted correctness likelihood of the AI suggestion to inform intervention selection to encourage more appropriate user reliance [23]. Taking advantage of people's anchoring bias, the adaptive workflow system from prior work [23] adaptively adjusts whether an AI suggestion is shown before or after the user makes an initial decision on the task instance; when the user has a greater correctness likelihood than the AI, the AI suggestion is only revealed to the user after they enter an initial decision, and vice versa when the user has a lower correctness likelihood than the AI. While estimations of AI and user correctness likelihood may be able to provide general direction for reliance calibration, more research is required to estimate the range of appropriate reliance without ground truth knowledge so that user reliance calibration can be encouraged on a finer scale. One potential approach is to identify mindless reliance/non-reliance in users by analyzing users' gaze behavior.

Lastly, more experimentation is needed to determine whether our trained reliance assessment models are generalizable to new interface setups (*i.e.,* AI explanations overlaid on the task [21]) and interaction schemes (*i.e.,* involving users in the AI's prediction generation process [14]). Moreover, the spatial reasoning task from this study was low-stake in nature. In high stake tasks, user reliance may be impacted differently as users may generally be more cautious (*i.e.,* longer gaze duration on AI recommendation) even if their reliance on and trust in the AI was the same. Further exploration is also needed to determine whether the trained models are applicable to new task types such as feature-based prediction tasks, visual inspection tasks, and so on. However, we speculate that while specific coefficient values might vary, the features we identified could still be useful for human-AI reliance modeling on new tasks.

In future work, we plan to implement an adaptive reliance calibration support system and evaluate its effectiveness compared to a static system. Suppose we had an appropriate reliance range in mind based on the predicted user and AI correctness likelihoods; an intervention could then be executed at the beginning of the trial, *i.e.,* using the adaptive workflow system [23] or the time-based de-anchoring strategy [29]. Due to individual differences in user backgrounds and preferences, the same intervention might incur varying levels of response in users [23]. Thus, a real-time human-AI reliance assessment would give us the opportunity to evaluate the effectiveness of the initial intervention and make personalized adjustments to the intervention as necessary (*e.g.,* further extend decision time if user reliance remains undesirably high [29]). We believe that a real-time assessment of human-AI reliance can provide valuable feedback to the AI system, which allows for the implementation of a more personalized adaptive human-AI reliance calibration support. We hope that the flexibility such an adaptive system provides can better facilitate user reliance calibration and help achieve more optimal human-AI team performance.

ACKNOWLEDGMENTS

We thank Jaimie Patterson for their valuable feedback and assistance in this work. This work was supported in part by National Science Foundation Award #1840088.

REFERENCES

[1] Anthony L Baker, Elizabeth K Phillips, Daniel Ullman, and Joseph R Keebler. 2018. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–30.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[3] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[5] 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* 5, CSCW1 (2021), 1–21.

[6] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.

[7] HR Chennamma and Xiaohui Yuan. 2013. A survey on eye-gaze tracking techniques. *arXiv preprint arXiv:1312.6410* (2013).

[8] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science advances* 8, 31 (2022), eabq6147.

[9] Andrew T Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods Instruments and Computers* 34, 4 (2002), 455–470.

[10] Greg Edwards. 1998. A tool for creating eye-aware applications that adapt to changes in user behaviors. In *Proceedings of the third international ACM conference on Assistive technologies*. 67–74.

[11] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1362–1374.

[12] Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias FC Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, et al. 2023. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific Reports* 13, 1 (2023),

1383.

[13] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 31.

[14] Catalina Gomez, Mathias Unberath, and Chien-Ming Huang. 2023. Mitigating knowledge imbalance in AI-advised decision-making through collaborative user involvement. *International Journal of Human-Computer Studies* 172 (2023), 102977.

[15] Seung Seog Han, Young Jae Kim, Ik Jun Moon, Joon Min Jung, Mi Young Lee, Woo Jin Lee, Chong Hyun Won, Mi Woo Lee, Seong Hwan Kim, Cristian Navarrete-Dechent, et al. 2022. Evaluation of Artificial Intelligence–Assisted Diagnosis of Skin Neoplasms: A Single-Center, Paralleled, Unmasked, Randomized Controlled Trial. *Journal of Investigative Dermatology* 142, 9 (2022), 2353–2362.

[16] Chien-Ming Huang and Bilge Mutlu. 2014. Multivariate evaluation of interactive robot systems. *Autonomous Robots* 37 (2014), 335–349.

[17] Young-Mok Kim, Young-Gil Kim, Seung-Yong Son, Soo-Yeon Lim, Bong-Yeol Choi, and Doo-Hyun Choi. 2022. Review of Recent Automated Pothole-Detection Methods. *Applied Sciences* 12, 11 (2022), 5320.

[18] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[19] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[20] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[21] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[22] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[23] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. *arXiv preprint arXiv:2301.05809* (2023).

[24] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.

[25] 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* 3, CSCW (2019), 1–15.

[26] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI: Literature review. (2022).

[27] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[28] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2022. A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making. *arXiv preprint arXiv:2204.10806* (2022).

[29] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2020. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *arXiv preprint arXiv:2010.07938* (2020).

[30] Amy Rechkemmer and Ming Yin. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*. 1–14.

[31] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[32] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures. *arXiv preprint arXiv:2203.12318* (2022).

[33] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[34] Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. 2022. Uncalibrated Models Can Improve Human-AI Collaboration. *arXiv preprint arXiv:2202.05983* (2022).

[35] Qinggang Wang, John J Koval, Catherine A Mills, and Kang-In David Lee. 2007. Determination of the selection statistics and best significance level in backward stepwise logistic regression. *Communications in Statistics-Simulation and Computation* 37, 1 (2007), 62–72.

[36] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[37] 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 *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

A FEATURE SELECTION FOR RELIANCE MODELS

A.1 List of All Features Considered

- **Trust in AI**: users' self-reported pre-study trust in AI technology as reported on a 7-point Likert scale, with 1 being "strongly disagree" that they trust AI technology and 7 being "strongly agree" that they trust AI technology.
- **Familiarity with AI**: users' self-reported pre-study familiarity with AI technology as reported on a 7-point Likert scale, with 1 being "strongly disagree" that they are familiar with AI technology and 7 being "strongly agree" that they are familiar with AI technology.
- **Start of first gaze fixation on AI suggestion**: the time in seconds of the start time of the user's first gaze fixation on the AI's suggestion.
- **Duration of first gaze fixation on AI suggestion**: the time in seconds of the duration of the user's first gaze fixation on the AI's suggestion.
- **Gaze duration on AI**: the sum of the durations of all gaze fixations on the AI's suggestions during the task divided by the total time fixated on completing the task (including the AI suggestions, the user response grid, and the task image).
- **Gaze duration on task**: the sum of the durations of all gaze fixations on the task image during the task divided by the total time fixated on completing the task (including the AI suggestions, the user response grid, and the task image).
- **Gaze duration on user**: the sum of the durations of all gaze fixations on the user response grid during the task divided by the total time fixated on completing the task (including the AI suggestions, the user response grid, and the task image).
- **Gaze shifts from AI to task**: the total number of times the user shifted their eye gaze from the AI suggestion to the task image divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Gaze shifts from AI to user**: the total number of times the user shifted their eye gaze from the AI suggestion to the user response grid divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Gaze shifts from task to user**: the total number of times the user shifted their eye gaze from the task image to the user response grid divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Gaze shifts from task to AI**: the total number of times the user shifted their eye gaze from the task image to the AI suggestion divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Gaze shifts from user to AI**: the total number of times the user shifted their eye gaze from the user response grid to the AI suggestion divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Gaze shifts from user to task**: the total number of times the user shifted their eye gaze from the user response grid to the task image divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).

- **Gaze shifts onto AI**: the total number of times the user shifted their eye gaze from the task image or user response grid to the AI suggestion divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Gaze shifts onto task**: the total number of times the user shifted their eye gaze from the AI suggestion or user response grid to the task image divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Gaze shifts onto user**: the total number of times the user shifted their eye gaze from the AI suggestion or task image to the user response grid divided by the total number of gaze shifts between aspects of the task (from AI suggestion/task image/user response grid to AI suggestion/task image/user response grid).
- **Initial human-AI agreement**: the number of squares in the user response that matched the AI suggestion when the user was halfway through the task divided by the total number of squares (16). This feature was only included in models after the 50% time stage.
- **Task difficulty**: whether the task at hand was asking for the top view (easier) or bottom view (more difficult) of the structure. Providing the top view of the structure involved only one mental rotation, while providing the bottom view of the structure involved two mental rotations.
- **Task Duration**: the total amount of time in seconds spent on completing the task. The value of this feature was multiplied by the time stage before being used as input; *i.e.,* at the 50% time stage, the input task duration used was calculated as the total task duration in seconds multiplied by 0.5.

A.2 Feature Selection for Task-Based Measure of Reliance Model

Table 2. Details from the feature selection for the task-based reliance model. We trained stepwise multiple linear regression models using gaze and task behavior data from the entire trial, along with user demographics. We used backward elimination as the stepwise method, likelihood ratio as the selection criterion, and $p < .25$ as the stop criterion.

Iteration	Feature	χ^2	p-value	Removed or Kept
1	Gaze shifts from AI to user	<.01	1.000	Removed
2	Gaze shifts from task to AI	<.01	1.000	Removed
3	Duration of first gaze fixation on AI suggestion	0.01	.994	Removed
4	Gaze shifts from user to task	0.05	.973	Removed
5	Gaze shifts onto AI	0.31	.857	Removed
6	Trust in AI	1.00	.604	Removed
7	Gaze shifts from task to user	1.42	.492	Removed
8	Gaze shifts from AI to task	1.05	.592	Removed
9	Start of first gaze fixation on AI suggestion	1.17	.557	Removed
10	Gaze shifts onto user	1.40	.497	Removed
11	Task duration	10.91	.004	Kept
11	Familiarity with AI	8.01	.018	Kept
11	Gaze shifts onto user	43.91	<.001	Kept
11	Gaze shifts from user to AI	5.21	.074	Kept
11	Gaze shifts onto AI	39.71	<.001	Kept
11	Initial human-AI agreement	139.79	<.001	Kept
11	Task difficulty	3.39	.183	Kept

A.3 Feature Selection for Gaze-Based Measure of Reliance Model

Table 3. Details from the feature selection for the gaze-based reliance model. We trained stepwise multiple linear regression models using gaze and task behavior data from the entire trial, along with user demographics. We used backward elimination as the stepwise method, likelihood ratio as the selection criterion, and $p < .25$ as the stop criterion.

Iteration	Feature	χ^2	p-value	Removed or Kept
1	Gaze shifts onto user	<.01	1.000	Removed
2	Gaze shifts onto task	<.01	1.000	Removed
3	Total duration	<.01	1.000	Removed
4	Task difficulty	0.09	.956	Removed
5	Start of first gaze fixation on AI suggestion	0.96	.620	Removed
6	Gaze shift from task to user	0.02	.992	Removed
7	Trust in AI	1.32	.515	Removed
8	Gaze duration on task	13.83	<.001	Kept
8	Gaze shifts from AI to task	16.32	<.001	Kept
8	Gaze shifts from task to AI	2.80	.246	Kept
8	Gaze shifts from user to AI	1.99	.370	Kept
8	Gaze shifts from user to task	1.43	.490	Kept
8	Gaze shifts from task to user	2.92	.231	Kept
8	Gaze shifts from AI to user	12.5	.002	Kept
8	Familiarity with AI	2.27	.321	Kept

B MODELING PERCEPTION-BASED RELIANCE

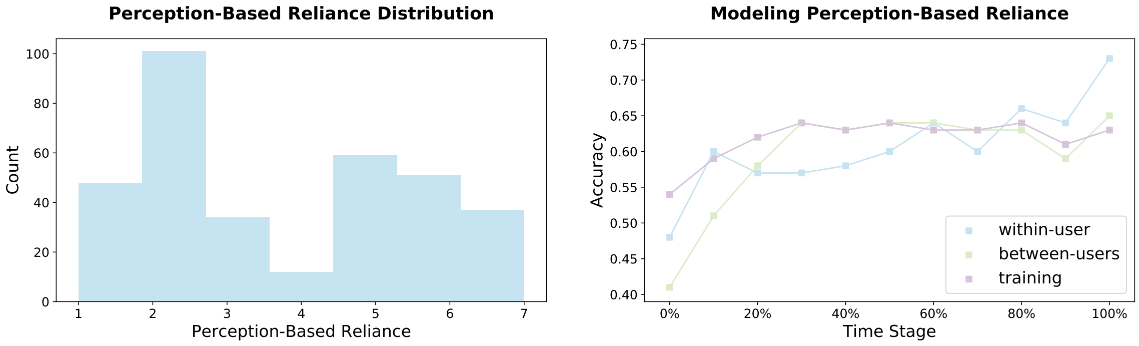

Fig. 4. Left: A bar plot showing the distribution of user-perception-based reliance (perceived reliance self-reported on a 7-point Likert scale, with 1 being “strongly disagree” that the user relied on the AI and 7 being “strongly agree” that they relied on the AI). Right: A line plot showing the accuracy of the model predicting the perception-based measure of reliance (perceived reliance) at different time stages. Greater accuracy signifies better model performance.

Due to the uneven distribution of perception-based reliance values in the data (see Figure 4, left), we grouped the 7-point Likert scale data into three categories: 1–3, 4, and 5–7. We trained stepwise multiple ridge regression models to model task- and gaze-based measures of reliance with backward elimination as the stepwise method [16], likelihood ratio as the selection criterion [35], and $p < .25$ as the stop criterion [35]. However, none of the behavioral, task-based, or demographic features significantly affected the perception-based measure of reliance at the $p = .25$ level. Therefore, we lowered the stop criterion to $p < 1.00$. We also experimented with modeling perception-based reliance with these input features using non-linear and other more complex model architectures (e.g., ordinal regression, random forest, deep neural network), but the ridge regression model had the best performance.

B.1 Feature Selection for Perception-Based Measure of Reliance Model

Table 4. Details from the feature selection for the perception-based reliance model. We trained stepwise multiple linear regression models using gaze and task behavior data from the entire trial, along with user demographics. We used backward elimination as the stepwise method, likelihood ratio as the selection criterion, and $p < 1.00$ as the stop criterion.

Iteration	Feature	χ^2	p-value	Removed or Kept
1	Gaze shifts onto task	<.01	1.000	Removed
2	Start of first gaze fixation on AI suggestion	<.01	1.000	Removed
3	Duration of first gaze fixation on AI suggestion	<.01	1.000	Removed
4	Familiarity with AI	<.01	1.000	Removed
5	Initial human-AI agreement	<.01	1.000	Removed
6	Gaze shift onto task	<.01	1.000	Removed
7	Gaze shift from user to task	<.01	1.000	Removed
8	Gaze shift from user to AI	<.01	1.000	Removed
9	Gaze from shift task to user	<.01	1.000	Removed
10	Gaze duration on user	0.61	.739	Kept
10	Gaze shifts from AI to task	0.81	.668	Kept
10	Gaze shifts from AI to user	1.21	.546	Kept
10	Gaze shifts from task to AI	0.81	.558	Kept
10	Gaze shifts onto AI	0.40	.817	Kept
10	Gaze shifts onto user	1.01	.604	Kept
10	Task duration	0.40	.817	Kept
10	Gaze duration on AI	0.81	.668	Kept

B.2 Modeling Perception-Based Reliance Halfway Through the Task

Table 5. Model details for a model predicting three classes of perception-based reliance (< 4, = 4, and > 4) halfway through the task time. β = coeffcient. SE = standard error, t = t-value, p = probability of committing a Type I error.

		Perception-Based Reliance								
		perceived reliance < 4			perceived reliance = 4			perceived reliance > 4		
Features	SE	β	t	p	β	t	p	β	t	p
Gaze duration on task	0.05	0.05	1.12	.265	0.03	0.67	.504	-0.09	-1.79	.075
Gaze shifts onto AI	0.12	0.18	1.53	.126	-0.10	-0.86	.389	-0.08	-0.67	.503
Trust in AI	0.05	-0.04	-0.97	.335	0.00	0.00	.998	0.04	0.96	.336
Gaze shifts onto user	0.07	-0.06	-0.90	.371	0.03	0.39	.700	0.03	0.51	.610
Gaze shift from AI to task	0.11	-0.16	-1.44	.152	0.10	0.94	.349	0.05	0.50	.619
Gaze shift from task to AI	0.09	0.03	0.36	.718	0.02	0.18	.855	-0.05	-0.54	.587
Gaze shift from AI to user	0.09	-0.10	-1.20	.231	0.07	0.79	.429	0.04	0.41	.682
Gaze shifts onto AI	0.18	-0.29	-1.64	.101	-0.03	-0.18	.859	0.32	1.82	.069
Gaze shifts onto user	0.10	0.13	1.38	.169	-0.01	-0.15	.878	-0.12	-1.22	.222
Task difficulty	0.05	0.10	2.02	.045	0.00	-0.05	.963	-0.09	-1.97	.050

C MODEL EVALUATION RESULTS

Table 6. Dynamic model performance for a model evaluated by training, within-users, and between-users methods at different time stages for task-, gaze-, and perception-based reliance.

Measures	Evaluation Method	Time Stages										
		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Human-AI Agreement (MSE)	training	0.051	0.044	0.044	0.044	0.044	0.028	0.028	0.028	0.028	0.028	0.024
	within	0.057	0.049	0.048	0.048	0.049	0.036	0.036	0.036	0.036	0.036	0.030
	between	0.061	0.050	0.050	0.049	0.049	0.031	0.031	0.031	0.030	0.030	0.024
Gaze Duration on AI (MSE)	training	0.004	0.004	0.004	0.003	0.003	0.002	0.002	0.002	0.002	0.001	0.001
	within	0.004	0.004	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.001
	between	0.006	0.004	0.005	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.001
Perceived Reliance (Accuracy)	training	0.54	0.59	0.62	0.64	0.63	0.64	0.63	0.63	0.64	0.61	0.63
	within	0.48	0.54	0.56	0.60	0.59	0.60	0.59	0.60	0.61	0.59	0.63
	between	0.41	0.51	0.58	0.64	0.63	0.64	0.64	0.63	0.63	0.59	0.65

Though the model performance on perception-based reliance prediction increased across time stages, the model accuracy was still low at 0.630 in the within-users evaluation and 0.716 in the between-users evaluation at its peak, when 100% of the data from the task trial was used. This was expected, as features had low influence (based on likelihood ratio test) on perception-based measures of reliance. This may be due in part to the fact that few data were available for the perception-based reliance = 4 class, which limited our ability to model for that class.

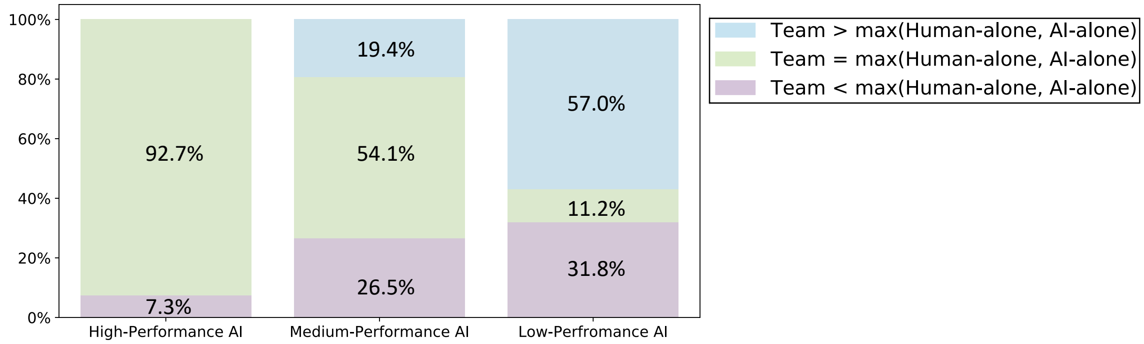

Fig. 5. A stacked bar plot showing the distribution of final task outcome classified into three categories: 1) final performance superior to both the individual human and the individual AI (blue), 2) final performance equal to the performance of the better agent out of the individual human and the individual AI (green), and 3) final performance inferior to the individual human, the individual AI, or both.

D USER RELIANCE CAN BE BETTER CALIBRATED TO IMPROVE PERFORMANCE

To investigate whether the variance in participant reliance between trials led to better-calibrated user reliance on the AI, we compared participants' final response accuracy to the maximum of the human and the AI's individual performance. We estimated human-alone performance on the task by computing the accuracy of the user response completed in the first half of the task, before the AI suggestion was shown to them (e.g., a participant's human-alone performance was 1.00 if all the cells they filled out during the first half of the task were correct). The AI's individual performance was calculated as the accuracy of the AI recommendation in that trial.

Without assuming complementary capabilities between the human and the AI, if the user relied on the AI suggestion appropriately, the accuracy of the final response should be at least as good as the more accurate response out of the human-alone response and the AI suggestion (a less-strict definition of appropriate reliance). Assuming complementary capabilities between the human and the AI, the criteria for appropriate reliance would be stricter, such that the accuracy of the final response must be better than the human-alone response and the AI suggestion unless the accuracy of the human-alone response or the AI suggestion is 1.00. In this study, users are not necessarily experts in the task. Thus, we do not assume human-AI complementary capabilities and adopt the less-strict evaluation of appropriate reliance.

We observed that team decision accuracy was greater than or equal to human-alone and AI-alone accuracy in 79.53% of the trials (272 out of 342). More specifically, among participants paired with the high-performance AI, team decision accuracy was greater than or equal to that of the human-alone and the AI alone in 92.70% of the trials (127 out of 137); among those paired with the medium-performance AI, complementary performance was observed in 73.47% of the trials (72 out of 98); and among those paired with the low-performance AI, team decision accuracy was greater than or equal to that of the human-alone and the AI-alone in 68.22% of the trials (73 out of 107). Figure 5 illustrates the distribution of performance comparisons between the team and the individual participants and AI agents.

While our estimation of the human-alone performance is limited, the presence of cases in which team performance was worse than that of the approximated human or AI alone indicates that more opportunities exist for interventions to assist user reliance calibration toward more optimal team performance.