Understanding User Reliance on AI in Assisted Decision-Making

SHIYE CAO, Johns Hopkins University, USA
CHIEN-MING HUANG, Johns Hopkins University, USA

Proper calibration of human reliance on AI is fundamental to achieving complementary performance in AI-assisted human decision-making. Most previous works focused on assessing user reliance, and more broadly trust, retrospectively, through user perceptions and task-based measures. In this work, we explore the relationship between eye gaze and reliance under varying task difficulties and AI performance levels in a spatial reasoning task. Our results show a strong positive correlation between percent gaze duration on the AI suggestion and user AI task agreement, as well as user perceived reliance. Moreover, user agency is preserved particularly when the task is easy and when AI performance is low or inconsistent. Our results also reveal nuanced differences between reliance and trust. We discuss the potential of using eye gaze to gauge human reliance on AI in real-time, enabling adaptive AI assistance for optimal human-AI team performance.

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

Additional Key Words and Phrases: human-AI interaction, decision support tools, gaze, decision making, trust

ACM Reference Format:
Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 471 (November 2022), 23 pages. https://doi.org/10.1145/3555572

1 INTRODUCTION

AI-assisted human decision-making aims to augment the human-AI team performance to exceed both parties' individual performances. However, prior studies have found that in experimental tasks such as text classification [4], deception detection [33, 34], and treatment selection [31], even though a human-AI team often outperforms the individual human's performance, its team performance is inferior to that of the AI alone. Part of this sub-optimal human-AI team performance may be attributed to the user's failure to properly calibrate how much they should rely on, or trust in, the AI. When the user has little trust for the AI assistance, they do not properly rely on it. When the user has too much trust for the AI, they over-rely on it [9, 32, 50]. In the latter case, the human-AI team performance may be bounded by the AI capability, which can be problematic especially in critical task domains since AI systems are not error-free. Therefore, it is important to assess user reliance on AI during assisted decision-making and apply necessary strategies to help users calibrate their reliance to achieve enhanced human-AI team performance.

Indeed, prior research was able to manipulate user agency in decision-making successfully through manipulating the level of machine assistance [35]. However, most prior works assess reliance and trust in human-AI interaction through self-reported measures (perceived reliance and trust) and task measures (user AI agreement fraction). These measures are only used to analyze user

Authors' addresses: Shiye Cao, scao14@jhu.edu, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA, 21218; Chien-Ming Huang, chienming.huang@jhu.edu, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA, 21218.

2573-0142/2022/11-ART471
https://doi.org/10.1145/3555572

reliance and trust *retrospectively*, and therefore are inadequate for timely, adaptive calibration of reliance and trust in human-AI collaboration. Recent works have explored the use of physiological measures including heart rate variability, galvanic skin response, and electroencephalography (EEG) for assessing human-AI trust. While physiological measures would allow for real-time measurements of trust and reliance, they require special equipment, thereby making them difficult to generalize for everyday use. The rise of real-time eye gaze tracking with regular webcams [55, 56, 69] points to a new possible way of assessing human reliance on AI in assisted decision-making for a wider range of everyday situations. Prior research has shown how eye gaze may provide insights into human decision-making and collaboration ([28]); for instance, gaze duration is indicative of people's preferences in selective decision-making [72]. In this work, we explore how eye gaze may be used to gauge human reliance on AI in assisted decision-making in hopes to enable future adaptive AI assistance for optimal human-AI team performance.

Our exploration was contextualized in an AI-assisted spatial reasoning task (Fig. 1). We conducted an in-person experiment and recorded participants' eye gaze during the experimental task. We manipulated task difficulty and AI accuracy level and sought to understand 1) the role of eye gaze in AI-assisted human decision-making; 2) the relationship between gaze behavior and human reliance on the AI; and 3) the effects of task difficulty and AI performance on task accuracy, reliance, and trust. Our results revealed 1) a strong positive correlation between the percent gaze duration on the AI suggestion and human reliance on the AI; 2) nuanced differences between the concepts of reliance and trust in human-AI interaction; and 3) preserved user agency when the task is easy and AI performance is low or inconsistent. Our findings point toward a possibility of enabling real-time adaptive human-AI collaboration through gaze awareness. Next, we describe relevant prior research to help situate this work.

2 RELATED WORK

2.1 Human-AI Collaboration

Artificial Intelligence (AI) agents have much potential to improve efficiency and productivity of humans in many domains such as agriculture [36, 54] and medical care [21, 30, 54, 61]. In fact, agents are already out-performing their human counterparts in select scenarios in recidivism prediction [41, 76], cancer diagnosis [21, 68, 80], and speech recognition [19]. Under ethical and legal concerns, particularly in critical tasks with high-stakes like medical diagnosis, AI agents are called to be used as decision support tools for human decision-makings [10, 80, 81]. However, studies have found the human-AI collaborative performance to be inferior to the AI performance alone in tasks ranging from text classification [4], to deception detection [33, 34], to treatment selection [31]. A main reason for this is because users have trouble balancing their trust for the AI and their agency, and frequently under/over-rely on the AI agent [9, 32, 50]. Prior works have explored providing explanations and more model information to help users better understand AI's reasoning so that they can make a more informed decision on the trustworthiness of the AI recommendation [4, 10]. However, explanations did not appear to offer more benefits than simply displaying the model accuracy [4, 13, 52], and in some cases even reduced human agency in the decision-making [35].

2.2 Human-AI Trust

Previous work in human-AI interaction often adopted the definition of trust from human-human trust, illustrating trust as "an attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [38]. Research to date has adopted a variety of quantitative measures to study human-AI trust and reliance. Many studies used task-based metrics of trust such as agreement fraction (frequency at which a person's final response agreed with the

AI) [35, 42, 44, 50, 70, 85], switch fraction (given that a person's initial response was different from the AI, the frequency at which the person revised their initial response for their final response to match that of the AI) [44, 47, 85], and the number of trials at which a person authorized the AI agent to make predictions on behalf of them [15]. Moreover, studies often used post-task or post-study questionnaires to gauge user's self-perceived trust in the AI. Questions about constructs related to trust including reliance [79], confidence [3, 15, 16, 44, 47, 50], reliability [3, 11, 44, 53], and predictability [3, 11, 27, 45] are also commonly included in the questionnaires to help aid the characterization of trust. However, task-based measures and self-reported perception measures only afford retrospective analysis of user trust.

Some studies have explored if human trust in intelligent systems can be measured through physiological measures, such as heart rate variability (HRV), galvanic skin response (GSR), electroencephalography (EEG). The use of these measures could allow for real-time estimation of trust. However, no studies have observed a correlation between HRV and GSR with human trust in machines. In a study on human trust in virtual reality agents, no significant effects of cognitive load of the task and accuracy of the agent (highly correlated with perceived trust in the experiment) and their interaction were found on any of the physiological measures (HRV, GSR, and EEG) [23]. A study on traffic augmentation in automated driving system found no correlation between HRV and trust based on user self-reported Technology Acceptance Model (TAM) and the Trust Scale (TS) values [83]. However, several time-domain EEG features and GSR were found predictive of whether or not the user would trust or distrust an AI suggestion [1]. Furthermore, EEG signal components were also found to be correlated with the trustworthiness of the AI in an investment game scenario with AI agents of varying trustworthiness levels [82]. Particularly, the frontal and the occipital area of the brain were identified to be correlated with trust. To the best of our knowledge, research thus far has not explored the use of eye tracking to measure trust and reliance in human-AI interaction, but prior works have noted it as an area worth studying [23, 79].

2.3 Gaze Tracking and Decision Making

Gaze tracking is the analysis of measured eye movement data with respect to the visual scene [14]. Since gaze can reveal underlying attentional patterns of people selectively seeking the information they need from the environment, we can gain insights into people's thoughts and intentions based on where they are looking [8, 24, 60]. Prior works have explored eye movement data in a large variety of tasks, from basic scene viewing tasks [26, 43] and visual search tasks [5, 17], to complex food preparation tasks [37] and driving tasks [39, 40]. Fixations on a scene encode the features in the scene, and the duration of a fixation is determined by the amount of time required to carry out the intended feature encoding [43]. Thus, gaze fixation duration and positions are reflective of the person's perceptual and cognitive processing of the scenery [43]. For instance, in reading, as text becomes conceptually more difficult, fixation duration increases [63]. Moreover, semantically informative objects in a picture tend to have longer total fixations duration [26]. In visual preference-based selection decision-making, people also tend to like things that they spend more time looking at (preferential looking) [2, 67, 72]. Furthermore, people spend longer time looking at the option they ultimately choose (gaze bias effect) [20, 48, 58, 65, 72, 73, 77]. Together, the two phenomena form the "Gaze Cascade" model, in which exposure to an item increases preference, preference increases gaze duration, which in turn increases exposure, forming a positive feedback loop that leads to selection of the item [48, 72, 73].

In human-computer interaction, studies have explored the use of gaze tracking as a multi-modal input device for the system to read the intentions of the users [66, 74, 75, 86]. Some studies have also used gaze tracking for user experience research and usability testing of websites and smartphone apps [22, 46, 51, 62]. A study analyzed the gaze behavior of drivers interacting with navigation system of different display sizes and positions while driving to determine which combination causes the least level of visual distraction for the driver [87]. Similarly, a study in human-AI interaction also employed gaze tracking to gain more insights into the user decision-making process while engaging with an automated driving system [49]. The study used eye gaze tracking to detect poor use of automated driving system, less attention paid to the road ahead. However, no study has explored the characterization of human trust in and reliance on AI through eye gaze tracking.

3 METHODS

3.1 Hypotheses

We designed a user study to explore the relationship between gaze and user trust in and reliance on AI. We hypothesized that manipulating task difficulty and AI accuracy level will result in changes in the user's level of reliance and trust in the AI agent and that this change will be reflected in their gaze duration on the AI. More specifically, our hypotheses are as follows:

- Hypothesis 1: By increasing task difficulty, regardless of the accuracy of the AI, users will rely more on the AI. This hypothesis is informed by previous work that showed greater user reliance on decision aids when a task is more difficult [7, 57].
- Hypothesis 2: By decreasing the accuracy of the AI, regardless of the task difficulty, users will rely on the AI less. This hypothesis is based on previous findings suggesting that the observed accuracy of a model can affect people's reliance on the model [85].
- Hypothesis 3: User gaze duration on the AI suggestion increases with higher user reliance on AI. Higher user reliance on AI can be reflected in higher user AI agreement in the final response. Furthermore, the gaze bias effect suggests that preference for the AI suggestion and adoption of the AI suggestion means longer gaze duration on the AI suggestion [58].
- Hypothesis 4: User gaze duration on the AI suggestion increases with higher user trust in the AI. We expect that higher trust for the AI suggestion increases preference, which in turn would increase gaze duration on the AI suggestion [73].

3.2 Experimental Task

To investigate the above hypotheses, we adopted a visuospatial task (Fig. 1) that is commonly used to study cognitive ability in cognitive psychology research. The spatial ability involved in this task is of importance in many disciplines including architecture, mathematics, and medicine [59, 64, 78]. For example, in medicine, spatial ability is critical for medical professionals, such as radiologists and dentists, to understand medical images (CT, MRI, X-ray, and ultrasound) [25]. They must comprehend complex three-dimensional (3D) structures based on two-dimensional image slices [6]. Our experimental task represents a simpler version of such complex spatial reasoning where the participants are asked to apprehend the top-view or bottom-view of a 3D block structure based on one 2D image slice. As AI continues to be developed to assist humans in complex visuospatial tasks as in, for example, diagnostic radiology, we hope that the results of this investigation will help inform the development of gaze-aware AI assisted decision support systems for future real-world applications.

In our experiment, the participants were asked to complete a 10-trial of the task with the assistance of an AI agent on the computer. In each trial, the participants were shown a new three dimensional structure made up of 15 cubes of one of four colors (black, orange, purple, or blue) and asked to determine what the structure would look like from the top view or the bottom view on a 4 by 4 grid (Fig. 1). The color of the square in the top right corner of the response grid was provided as reference for how the structure should be oriented in the answer. The structures were crafted

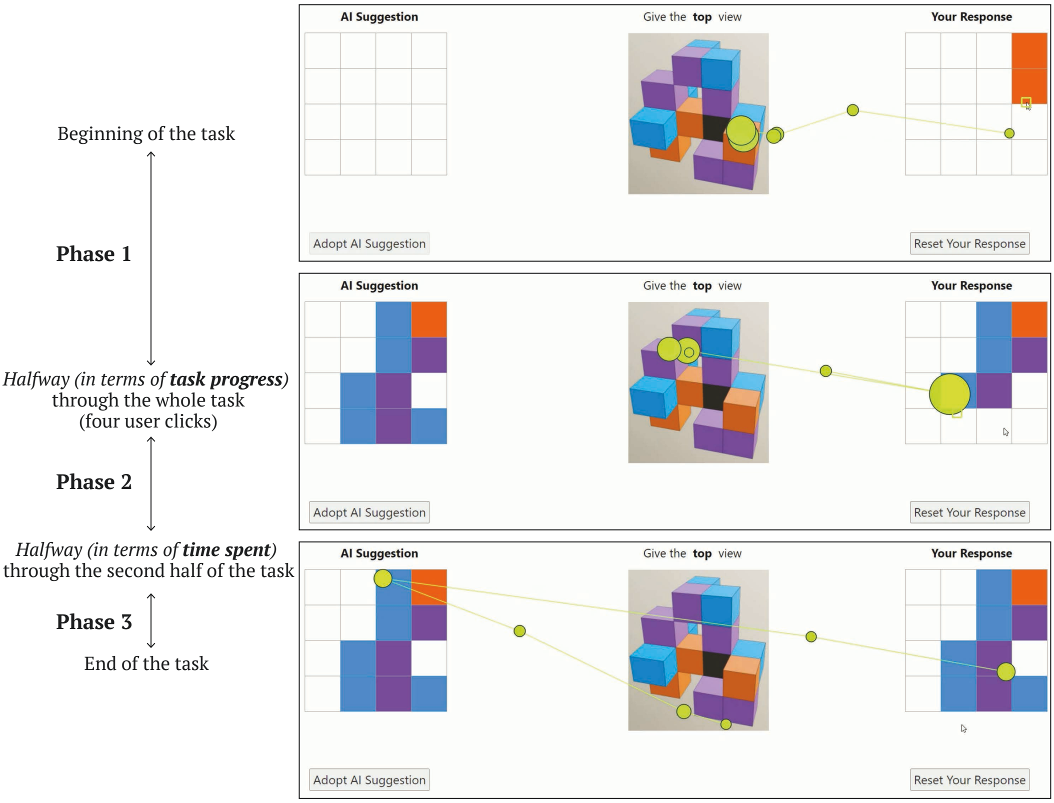

Fig. 1. Overview of the 3 phases in the task. Phase 1 is the time before the AI appears, which is roughly halfway through the task in terms of the progress in the user response; Phase 2 is the time after the AI appears until halfway in term of the time spent after the AI appears; Phase 3 is the second half of the time spent after the AI appears. The phases are demonstrated through screen shots from a participant's screen recording overlaid with their real-time eye gaze in green. Each green circle shows the location of a fixation and the size of the circle is representative of the length of the fixation at that location. A screen recording of a participant completing a full trial over-laid with their eye gaze is available at https://www.youtube.com/watch?v=eFIPCvm3Lqg.

using Blender[1] and designed so that there is only one possible correct response given the color of the top right corner square. The structures were also designed to have nine blocks in the correct answer; however, this information was not revealed to participants.

The user response is entered by clicking on the square that the participant would like to change the color of. The color of that square will change following the sequence of white, blue, orange, purple, and black. With each click, the square will change to the next color in the sequence and when the user clicks on a black square, the square will change back to white. This method was chosen to preserve the gaze fixation on the square when the user is updating their response. It does not involve the user looking down at the keyboard to look for specific keys, nor looking else where on the screen.

[1] https://www.blender.org/

The experimental task was implemented as a custom web application using the React[2] and Flask frameworks[3], and deployed with Heroku[4].

3.3 Study Design

In this study, we employed a 2 by 3 mixed factorial design with the task difficulty (top-view and bottom-view) as a within-subjects factor and the AI accuracy (high, medium, and low) as a between-subjects factor. In the beginning of an experiment, a participant was randomly assigned to a particular AI accuracy level by the system; neither the participant nor the experimenter were made aware of the accuracy level of the AI.

3.3.1 Task Difficulty. To manipulate task difficulty, we chose to let participants solve 3D spatial reasoning tasks from two perspectives, namely top and bottom. We expected bottom-view tasks to be more difficult as they required one more level mental rotation than top-view tasks. The participants rated the bottom-view task to be more difficult than the top-view task in the practice round (Section 4.1). Apart from the manipulation of the task viewpoint, difficulty of each task trial was controlled by a consistent number ($n = 15$) of blocks making up of each structure and a consistent number ($n = 9$) of filled-in blocks in the response. Each participant alternated between doing a top-view task and a bottom-view task in hopes to decrease the learning effect throughout the experiment.

3.3.2 AI Accuracy Level. The accuracy of AI suggestions is defined as the number of squares in the AI suggestion that matched the ground truth grid divided by 16 (the total number of cells). We designed three accuracy levels in this experiment for the AI agent.

- High performance. AI with high performance had 100.00% accuracy on all ten trials.
- Medium performance. AI with medium performance had an average cell-level accuracy of 97.50% over the 10 trials. The agent had one minor error in four trials (two trials for each viewpoint) and 100% accurate response on the rest of the six trials. A suggestion with a minor error was defined as either having one extra square filled in, one square missing on the response grid, or one block of an incorrect color. As a result, an AI suggestion with a minor error has an accuracy of 93.75%. To ensure consistency, the trials with a minor error were set to be round 2, 3, 7, 8 in the main experiment.
- Low performance. AI with low performance had an average cell-level accuracy of 49.38% over the 10 trials. The agent had one major error in suggestion on all ten trials. A suggestion with a major error was defined as having a minor error and on top of that at least three more square-level mistakes. The additional mistakes may be a result of a reflection error (answer rotated by 90 degrees) or having shifts in parts of the response. As a result, AI suggestions with a major error had an accuracy between 31.25% and 75.00%.

3.3.3 Timing of an AI Suggestion. An AI suggestion was not shown to the user until the user entered a response for four different squares on the user response grid, regardless of the correctness (color and location) of the blocks they entered. Since the correct answer for the task always contained nine squares, four squares corresponds to half way through completing the task in terms of progress (Fig. 1); note that the first block in the top right corner was always provided. Participants were not told how many blocks were in the final response and when the AI would appear. This design choice was to encourage the participants to form an opinion about the task answer before they saw the AI suggestion so they did not solely rely on the AI.

[2] https://reactjs.org
[3] https://flask.palletsprojects.com/en/2.0.x
[4] https://www.heroku.com

For simplicity, we refer to the task time before an AI suggestion appears as phase 1, half way into the task time (in terms of time spent) after an AI suggestion appears as phase 2, and the rest of the task time until the user clicks "Done" as phase 3. Fig. 1 illustrates the process of a user completing a top-view task.

3.4 Study Procedure

Upon agreeing to participate in the study, the participants were asked to sit in front of a computer screen on a chair without wheels. They were told to move as close to the desk as possible and lean back on the back of the chair. Then, an experimenter adjusted the position and angle of the Gazepoint GP3 remote eye tracker (60 Hz) accordingly and calibrated the eye tracker using the built-in 9-point calibration program. The calibration procedure was repeated until all points were successfully calibrated and captured by the eye tracker for both of the participants' eyes. After the eye tracker was calibrated, the experimenter began screen recording and eye movement recording.

Before pulling up the web interface for the experiment, the experimenter emphasized the importance of the participant staying still in the chair throughout the entire experiment. The experimenter also iterated that an AI suggestion might not appear right away at the beginning of the task and rather the AI might take some time to come up with the answer and that the system was not broken because of the absence of an AI suggestion. The participants were told that the experimenter will stay in the room while they read the instructions for the experiment and the task in case of any question or concerns, but will leave the study room after that and the participant will not be able to ask any more questions.

In the web application, participants were first asked to fill out a demographic survey regarding their gender, age, education background, level of familiarity with AI, and level of trust for AI. After the demographic survey, participants were presented with the task instructions and a video explaining and walking through the task and some tips and strategies for the task for both top-view and bottom-view. Then, they were given one sample top-view task and a sample bottom-view task with the correct answers. Then, if the participant did not have any further questions, the experimenter left the room for the participant to proceed with the practice round of the study.

In the practice round of the study, the participants were asked to complete one top-view task and one bottom-view task for practice without the AI with unlimited time. The order at which the two practice tasks appeared was randomized and after each practice task the participants were asked to rate the difficulty of the task they just completed on a 7-point Likert scale.

After completing the two practice tasks and the post task survey, the participants were told that they will now be working with an AI agent. After they clicked "Start", a task image appeared at the center of the screen with a blank 4 by 4 grid to the left of it labeled "AI suggestion" and another blank 4 by 4 grid to the right of it labeled "User response" (Fig. 1). An AI suggestion would only appear after the user clicked on four different squares on the user response grid. The user had unlimited time to complete each task and was able to click "Done" whenever they were satisfied with their response. After each task trial, the participant was asked to complete a brief post-task survey concerning their agreement with, reliance on, and trust in the AI suggestion, as well as their confidence in their answers.

After the survey, no feedback on their response to the previous trial was given to the participant to reduce possible learning effects. Every task trial contained a completely different block structure and the task alternated between top-view and bottom-view. After the ten task trials, the participant was presented with a final survey with an open-ended question on the strategy they used for completing the task and any feedback they might have.

3.5 Measures

We used a set of metrics to evaluate task accuracy, reliance on AI (agreement, gaze duration, and perceived reliance), and other perception-based metrics. Eye gaze movement was recorded and processed using the Gazepoint Control and Gazepoint Analysis software.

3.5.1 Task Accuracy. The task accuracy (Range: 0–1) of each trial is defined as the number of squares in the user response that matched the ground truth grid divided by 16 (the total number of cells). Thus, if the user response matched the ground truth perfectly, then the accuracy of the user response would be 1.00.

3.5.2 User Reliance on AI. To understand user reliance on the AI, we use three metrics aiming to capture user reliance through the lens of user-AI agreement on task outcomes, subjective perception, and gaze behavior.

- User-AI Agreement (Range: 0–1). We defined the rate of agreement between the final user response and the AI suggestion as the number of squares in the user response that matched the AI suggestion divided by 16. Thus, if the user response matched the AI suggestion perfectly, then the user AI agreement ratio would be 1.00. This metric is a *task*-based measure of reliance.
- Perceived Level of Reliance on AI (Range: 1–7). We used one item "I relied on the AI suggestion in the previous task" to assess the participant's self-reported level of agreement with the statement after each task trial in the main experiment. The item is on a 7-point Likert scale with 1 being strongly disagree and 7 being strongly agree. This metric is a *perception*-based measure of reliance.
- Percent Gaze Duration on AI Suggestion (Range: 0–1). The total amount of time the participant spent fixating on the AI suggestion divided by the total amount of time (in seconds) the participants spent fixating on the task, including fixations on AI suggestions, the task image, and user responses. This metric is a *behavior*-based measure of reliance.

3.5.3 Other Perception-based Metrics. In addition to perceived reliance, we include other questionnaire items to assess participants' agreement with and trust in the AI, as well as their confidence in their responses. All items below are on a 7-point Likert scale with 1 being strongly disagree and 7 being strongly agree.

- Perceived Level of Agreement with AI Suggestion (Range: 1–7). We used one item "I agree with the AI suggestion" to assess the participant's self-reported level of agreement with the AI suggestion after each task trial in the main experiment.
- Perceived Level of Trust in AI (Range: 1–7). We used one item "I trusted the AI suggestion in the previous task" to assess the participant's self-reported trust in the AI suggestion after each task trial in the main experiment.
- Confidence Level in Response (Range: 1–7). We used one item "I was confident in my answers" to assess the participant's self-reported confidence in their responses after each task trial in the main experiment.

We additionally included an open-ended question, "What strategies did you use that helped you to complete the task?" to obtain a qualitative understanding of how the participants might use the AI in their problem-solving processes.

3.6 Participants

A total of 42 participants (22 males, 20 females) were recruited through convenience sampling from the local community, through physical posters, electronic newsletter posts, and posts to student group mailing lists. The participants' age range was between 19 and 60 years ($M = 23.69, SD = 6.13$). Participants self-reported their level of familiarity with AI ($M = 3.38, SD = 1.02$) on a scale of 5, with 1 being not familiar at all and 5 being extremely familiar; the range of their reported familiarity level was between 1 and 5, with a median of 3 out of 5. The participants also self-reported their perceived level of trust for AI ($M = 3.33, SD = 0.69$) on a scale of 5, with 1 being not familiar at all and 5 being extremely familiar; the range of their reported trust level was between 2 and 4, with a median of 3 out of 5.

On average, the participants took roughly between 10 minutes and 50 minutes ($M = 25.47, SD = 7.89$) to complete the study. The participants received a $10.00 gift card as compensation for their time. The participants were incentivized to perform better with an extra $5 reward for the top 5% of performers. The study was approved by our institutional review board (IRB).

Data from two participants were excluded because they left the experiment room in the middle of the study. Data from another five participants were excluded because they moved too much during the experiment and their eyes were completely out of range and unable to be tracked by the gaze tracker for more than five task trials. Among the resulting 35 participants, 14 were assigned to the AI with high performance, 10 assigned to the AI with medium performance, and 11 assigned to the AI with low performance. For the data from the 35 participants, three more trials where the participant's eyes were out of range from the gaze tracker was removed from two participants. As a result, our data analysis included a total of 347 trials.

4 RESULTS

For the analyses reported below, if not specified otherwise, we performed two-way repeated measure analysis of variance (ANOVA) tests where task difficulty was set as a within-subjects factor, AI accuracy level as a between-subjects factor, and participants as a random effect. All post-hoc pairwise comparisons were conducted using Tukey's HSD test. We considered a $p < .05$ as a significant effect.

4.1 Manipulation Check

We first checked whether we successfully created two different task difficulty levels through the manipulation of the task orientation using Pearson's chi-square test for the ordinal dependent variable, self-report perceived task difficulty, and one-way analysis of variance (ANOVA) test for the continuous dependent variable, practice accuracy.

4.1.1 Perceived Task Difficulty. A Pearson's chi-square test revealed that the bottom-view task ($M = 4.94, SD = 1.26$) was perceived by the participants to be more difficult than the top-view task ($M = 3.17, SD = 1.29$), $\chi^2(6, N = 70) = 25.60, p < .001$

4.1.2 Practice Round Accuracy. A one-way ANOVA test was conducted to examine the effect of manipulating the task orientation on the participants' task performance in the practice round without the help from the AI. We found that the participants performed significantly worse in the bottom-view task ($M = 0.49, SD = 0.24$) than they did in the top-view task ($M = 0.99, SD = 0.05$), $F(1, 68) = 147.67, p < .001$.

4.2 Learning Effect

We examined a potential learning effect throughout the ten main trials using a three-way repeated measure analysis of variance (ANOVA) test where task difficulty and the trial number were set as within-subjects factors, AI accuracy level as a between-subjects factor, and participants as a random effect.

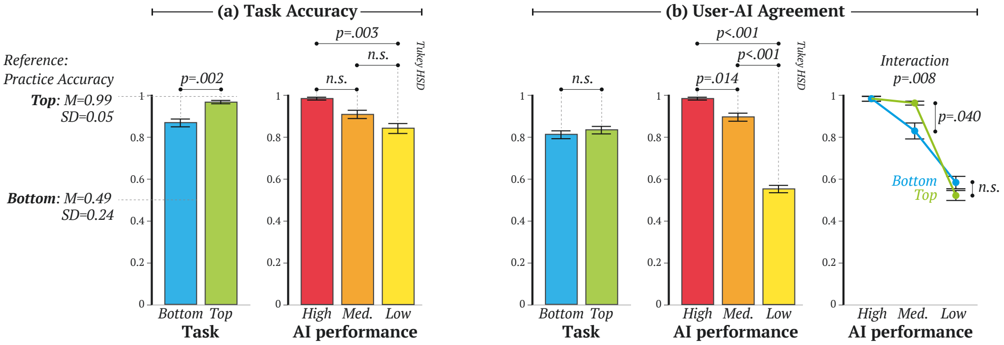

Fig. 2. (a) Bar plots demonstrating the final user response accuracy under varying task orientations (difficulty) and AI performance. (b) Bar plots showing the agreement rate between the final user response and the AI suggestion under varying task orientations and AI performance. The error bars shown in the plots represent the standard error and only significant results are emphasized.

A three-way repeated measure ANOVA test was conducted to explore the effect of AI accuracy level, task difficulty, and trial number on the accuracy of the user response right before the AI suggestion is provided (half-way through the task). Therefore, we did not observe a significant learning effect over the ten task trials.

4.3 Task Accuracy

A two-way repeated measure ANOVA test was conducted to examine the effect of AI accuracy level and task difficulty on the accuracy of the user response (Fig. 2, a). We found a significant effect of task difficulty on task accuracy, $F(1, 31.91) = 11.06, p = .002$, with a lower user task accuracy in the bottom-view task ($M = 0.87, SD = 0.24$) than the top-view task ($M = 0.97, SD = 0.10$). Additionally, we found a significant effect of AI accuracy level on the participant's task accuracy, $F(2, 32.02) = 6.64, p = .004$. Pairwise comparisons using Tukey's HSD test showed a significantly higher accuracy in participants with the high performance AI ($M = 0.98, SD = 0.08$) than participants with the low performance AI ($M = 0.84, SD = 0.25$), $p = .003$. No significant difference in performance was observed in participants with the high performance AI and participants with the medium performance AI ($M = 0.90, SD = 0.20$), $p = .165$, nor in participants with the medium performance AI and participants with the low performance AI, $p = .270$. There was no interaction effect of AI accuracy level and task difficulty on the user task accuracy, $F(2, 31.91) = 3.25, p = .052$. Fig. 2 (a) visualizes our results.

4.4 User Reliance on AI

4.4.1 User-AI Agreement. A two-way repeated measure ANOVA test was carried out to study the effect of AI accuracy level and task difficulty on the similarity rate between the user response and the AI suggestion (Fig. 2, b). We observed a significant effect of AI accuracy level on the agreement ratio between the final participant response and the AI suggestion, $F(2, 32.15) = 124.58, p < .001$. Pairwise comparisons using the Tukey's HSD test revealed a significantly higher similarity level in participants with the high performance AI ($M = 0.98, SD = 0.08$) than participants with the medium performance AI ($M = 0.90, SD = 0.19$), $p = .014$. Moreover, participants with the medium performance AI had a significantly higher agreement level than participants with the low performance AI ($M = 0.55, SD = 0.18$), $p < .001$, and participants with the high performance AI had a significantly higher agreement level than participants with the low performance AI,

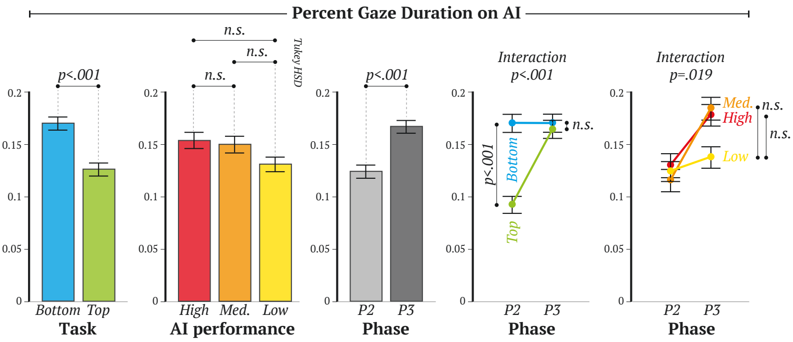

Fig. 3. Bar plots demonstrating the behavior-based metrics (percent gaze duration on AI) under varying task orientations, AI performance, and phase in the task. The error bars shown in the plots represent the standard error and only significant results are emphasized.

$p < .001$. No significant effect of task difficulty on the similarity rate between the final participant response and the AI suggestion was found, $F(1, 32.02) = 1.19, p = .283$. However, there existed a significant interaction effect of AI accuracy level and task difficulty on the user AI agreement rate, $F(2, 32) = 5.58, p = .008$. Pairwise comparisons using Tukey's HSD test (see Table 3 in Appendix) found that the difference in similarity is significantly lower for the bottom-view task ($M = 0.83, SD = 0.25$) than the top-view task ($M = 0.96, SD = 0.05$) for participants interacting only with the medium performance AI, $p = .040$. Fig. 2 (b) visualizes our results.

4.4.2 Perceived Level of Reliance on AI. A two-way repeated measure ANOVA test found a significant difference in the participants' perceived reliance on the AI under varying AI accuracy levels, $F(2, 32.11) = 12.73, p < .001$ (Fig. 4, a). Pairwise comparisons using Tukey's HSD test showed that participants with the high performance AI ($M = 4.68, SD = 1.99$) reported that they relied on the AI more than participants with the low performance AI ($M = 2.53, SD = 1.57$), $p < .001$. Participants with the high performance AI reported significantly higher perceived reliance levels than participants with medium performance AI ($M = 3.58, SD = 1.92$), $p = .043$. The difference between the reported perceived reliance level of participants with the medium performance AI and that of participants with the low performance AI was not significant, $p = .079$. Participants reported that they relied more on the AI in the bottom view task than the top view task, ($F(1, 31.8) = 3.89, p = .057$). No interaction effect of AI accuracy level and task difficulty ($F(2, 31.78) = 0.31, p = .739$) were identified on the participant reported perceived level of reliance on the AI agent. Fig. 4 (a) visualizes our results.

4.4.3 Percent Gaze Duration on AI Suggestion. In addition to understanding the effect of AI accuracy level and task difficulty on the participant's percent gaze duration on the AI suggestion, we were interested in how gaze duration on the AI might be influenced by task phase (time). To this end, we used a three-way repeated measure analysis of variance (ANOVA) test where task difficulty and task phase (phase 2 vs. phase 3) (Fig. 1) were set as within-subjects factors, AI accuracy level as a between-subjects factor, and participants as a random effect. The test revealed a significant difference in the participants' percent gaze duration on the AI suggestion under the bottom-view task ($M = 0.17, SD = 0.11$) and the top-view task ($M = 0.13, SD = 0.12$), $F(1, 32.42) = 16.18, p < .001$ (Fig. 3).

The test did not find a significant difference in the participants' percent gaze duration on the AI suggestion under varying AI accuracy levels, $F(2, 32.14) = 0.74, p = .487$. However, there was a significant difference in the participants' percent gaze duration on the AI suggestion in phase 2 of the task ($M = 0.12, SD = 0.12$) and phase 3 of the task ($M = 0.17, SD = 0.11$), $F(1, 634.5) = 22.95, p < .001$.

Moreover, there was a significant interaction effect of task difficulty and task phase on the participants' percent gaze duration on the AI suggestion, $F(1, 625.1) = 17.11, p < .001$. Pairwise comparisons using Tukey's HSD test (see Table 1 in Appendix) showed no difference between percent gaze duration on the AI suggestion for the bottom-view task in phase 2 ($M = 0.17, SD = 0.10$) and the bottom-view task in phase 3 ($M = 0.17, SD = 0.11$), $p = .974$, nor difference between bottom-view task in phase 3 and top-view task in phase 3, $p = .999$. However, we found a significant difference between percent gaze duration on the AI suggestion for the bottom-view task in phase 2 and the top-view task in phase 2 ($M = 0.09, SD = 0.12$), $p < .001$, and between the top-view task in phase 2 and the top-view task in phase 3, $p < .001$.

The interaction effect of AI accuracy level and the task phase on the participant percent duration on the AI suggestion was also significant, $F(2, 634.7) = 3.98, p = .019$. Pairwise comparison using Tukey's HSD Test (see Table 2 in Appendix) identifies a significantly higher percent duration on the AI suggestion in participants with the medium performance AI in phase 3 ($M = 0.18, SD = 0.11$) than participants with the medium performance AI in phase 2 ($M = 0.12, SD = 0.11$) ($p < .001$). There was no significant difference in duration on the AI suggestion in participants with the high performance AI (phase 2: $M = 0.13, SD = 0.14$, phase 3: $M = 0.18, SD = 0.12$) and the low performance AI between phases 2 and 3 of the task (phase 2: $M = 0.12, SD = 0.10$, phase 3: $M = 0.14, SD = 0.11$). No interaction effect on the participants' percent gaze duration on the AI suggestion was found from AI accuracy level and task difficulty ($F(2, 32.47) = 1.25, p = .301$), nor the AI accuracy level, task phase, and task difficulty ($F(2, 625.4) = 1.06, p = .347$). Fig. 3 visualizes our results.

Percent Gaze Duration on AI vs. User-AI Agreement. Since gaze duration on the top-view task and the bottom-view task was significantly different, we stratified our further analysis of gaze duration by task difficulty (Fig. 5). Pearson correlation coefficient was used to examine the linear relationship between the participants' percent gaze duration on the AI suggestion and user-AI agreement. For the bottom-view task, there existed a significant positive correlation between the participants' percent gaze duration on the AI suggestion and the similarity of their final response and the AI suggestion, $r(172) = .19, p = .011$. For the top-view task, there was a significant positive correlation between the participants' percent gaze duration on the AI suggestion and the similarity of their final response and the AI suggestion, $r(171) = .19, p = .012$. Fig. 5 (a) visualizes our results.

Percent Gaze Duration on AI vs. Perceived Reliance. Spearman's rank correlation was used to assess the relationship between the participant's percent duration on the AI and their perceived reliance on the AI. Among the bottom-view tasks, there was a positive correlation between the two variables, $r(172) = .37, p < .001$. Among the top-view tasks, there was also a positive correlation between the two variables, $r(171) = .30, p < .001$. Fig. 5 (b) visualizes our results.

4.5 Other User Perception Measures

We analyzed the effect of AI accuracy level and task difficulty on the participants' perceptions of their agreement with the AI, trust in the AI, and confidence in their response.

4.5.1 Perceived Agreement with AI Suggestion. A two-way repeated measure ANOVA test revealed a significant difference in the participants' perceived level of agreement with the AI suggestion under varying AI accuracy levels, $F(2, 31.98) = 111.46, p < .001$ (Fig. 4, b). No significant effect of task difficulty on perceived agreement was identified, $F(1, 32.2) = 1.48, p = .233$. However, there was a significant interaction effect of AI accuracy level and task difficulty on the participants' perceived

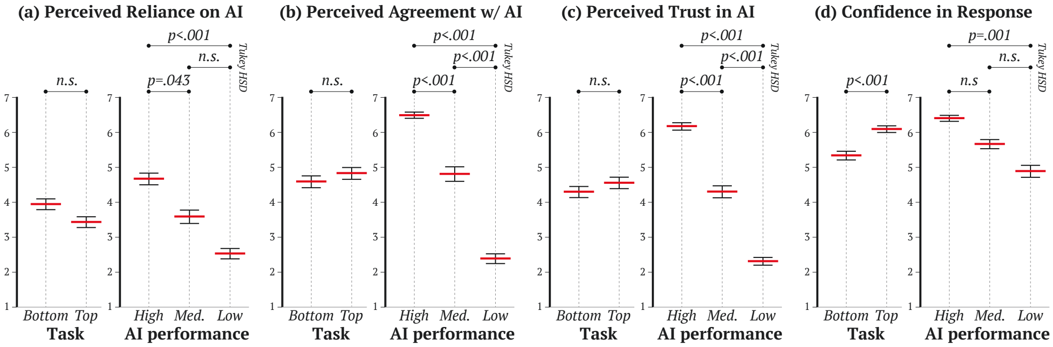

Fig. 4. Plots demonstrating the effects of varying task orientations and AI performance on user perception-based metrics: (a) perceived reliance, (b) perceived agreement, (c) perceived trust, and (d) user confidence in response. The error bars shown in the plots represent the standard error and only significant results are emphasized.

agreement with the AI suggestion, $F(2, 32.18) = 4.36, p = .046$. Pairwise comparisons using Tukey's HSD test (see Table 4 in Appendix) showed that participants with the high performance AI ($M = 6.48, SD = 1.01$) reported that they agreed with the AI suggestion more than participants with the medium performance AI ($M = 4.81, SD = 2.09$), $p < .001$; participants with the medium performance AI reported that they agreed with the AI suggestion more than participants with the low performance AI ($M = 2.37, SD = 1.44$), $p < .001$; and participants with the high performance AI reported that they agreed with the AI suggestion more than participants with the low performance AI, $p < .001$. Fig. 4 (b) visualizes our results.

4.5.2 Perceived Trust in AI. A two-way repeated measure ANOVA test identified a significant difference in the participants' self-reported perceived trust in the AI under varying AI accuracy levels, $F(2, 32.07) = 109.22, p < .001$ (Fig. 4, c). The ANOVA test revealed no significant effect of task difficulty on the participants' perceived trust, $F(1, 31.65) = 1.71, p = .200$. However, there was a significant interaction effect of AI accuracy level and task difficulty on the participants' level of trust in the AI agent, $F(2, 31.64) = 4.09, p = .026$. Pairwise comparisons using Tukey's HSD test (see Table 5 in Appendix) showed that participants with the high performance AI ($M = 6.19, SD = 1.26$) reported that they trusted the AI more than participants with the medium performance AI ($M = 4.29, SD = 1.71$), $p < .001$; participants with the medium performance AI reported that they trusted the AI more than participants with the low performance AI ($M = 2.30, SD = 1.19$), $p < .001$; and participants with the high performance AI reported that they trusted the AI more than participants with the low performance AI, $p < .001$. Fig. 4 (c) visualizes our results.

4.5.3 Confidence in Response. A two-way repeated measure ANOVA test showed a significant difference in the participants' self-reported confidence level in their answer under varying AI accuracy levels, $F(2, 32.07) = 8.39, p = .001$ (Fig. 4, d). Pairwise comparisons using Tukey's HSD test found that participants with the high performance AI ($M = 6.41, SD = 0.99$) reported that they are more confident in their answer than participants with the low performance AI ($M = 4.87, SD = 1.80$), $p = .001$. However, there was no significant difference in the reported confidence level in participants with the high performance AI and participants with the medium performance AI ($M = 5.67, SD = 1.33$), $p = .152$. Furthermore, the difference in the reported confidence level in the participants with the medium performance AI and the participants with the low performance AI were not significant, $p = .138$. The ANOVA test also found a significant

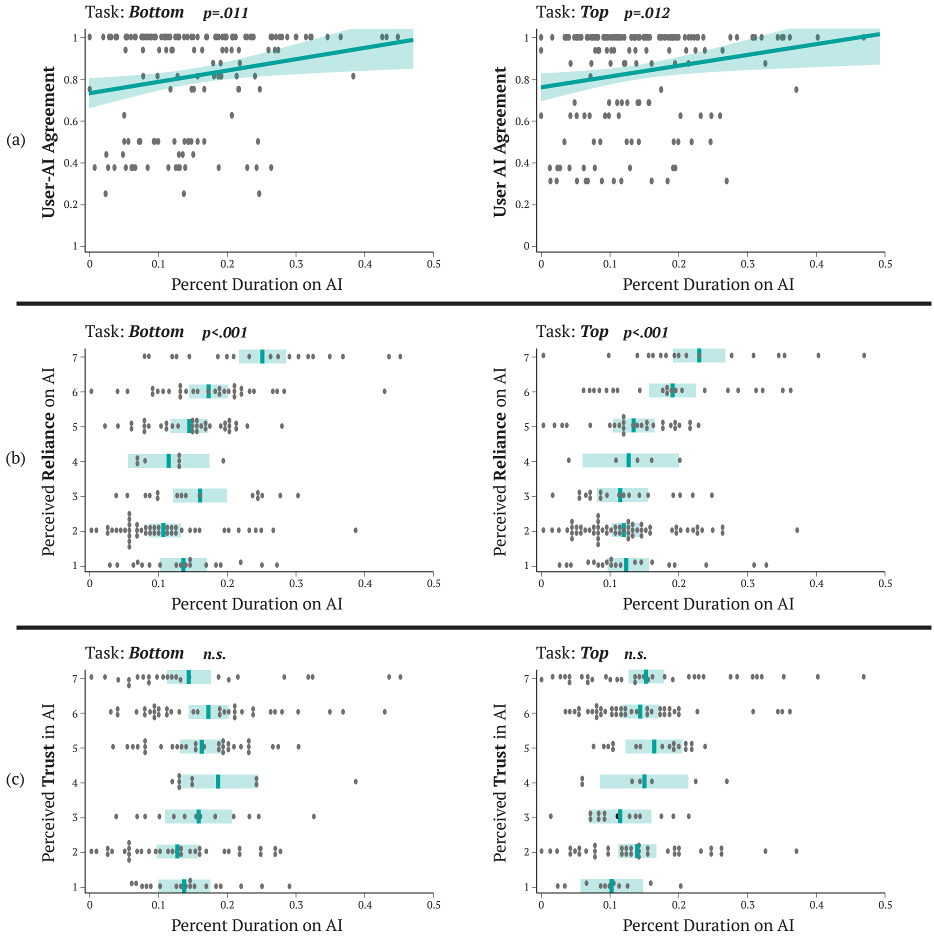

Fig. 5. Scatter plots demonstrating the correlation between percent gaze duration on AI suggestion and (a) final user AI agreement, (b) perceived reliance, and (c) perceived trust for bottom-view tasks and top-view tasks. The error bars shown in the plots represent confidence intervals.

effect of task difficulty on the participants' confidence in their answer, $F(1, 32.27) = 30.01, p < .001$. Participants had a lower confidence in their response to the bottom-view tasks ($M = 5.34, SD = 1.67$) than the top-view tasks ($M = 6.10, SD = 1.26$). However, no significant interaction effect of AI accuracy level and task difficulty on the participants' perceived confidence was observed, $F(2, 32.25) = 1.60, p = .217$. Fig. 4 (d) visualizes our results.

Perceived user-AI Agreement vs. Confidence in Response. We studied the effect of the participants' perceived agreement with the AI suggestion on their perceived confidence level in their answer. A Pearson's chi-square test revealed that the relationship between perceived agreement and confidence was significant, $\chi^2(36, N = 347) = 173.93, p < .001$.

Percent Gaze Duration on AI vs. Perceived Trust. Stratified by task difficulty, we used the Spearman's rank correlation to examine the relationship between the participants' percent gaze duration on the AI suggestion and their self-reported perceived trust in the AI. For the bottom-view task, there was no evidence of significant correlation between percent duration on the AI suggestion and perceived trust in the AI, $r(172) = .04, p = .596$. For the top-view task, no evidence was found of a significant correlation between percent duration on the AI suggestion and perceived trust in the AI, $r(171) = .07, p = .334$. Fig. 5 (c) visualizes our results.

5 DISCUSSION

5.1 The Impact of Task Difficulty

We found that task difficulty had significant impacts on task accuracy (Fig. 2, a), the participants' perceived confidence in their response (Fig. 4, d), and how much time the participants looked at the AI suggestion during the task (Fig. 3). However, we did not observe a significant effect of task difficulty on user-AI agreement in their final response (Fig. 2, b) and perceived reliance (Fig. 4, a). These results suggest nuanced differences between perception-based reliance, task-based reliance, and behavior-based reliance, illustrating the complexity of the *reliance* construct. In all, we do not have enough evidence to support Hypothesis 1 (the higher the task difficulty, the higher the user reliance on AI).

This finding contradicts the results from prior work [57], which showed that increase in task complexity and difficulty heightens reliance. The prior work also showed the higher the level of expertise the participant has in the task, the smaller the effect of task difficulty on user reliance. A possible explanation of our finding is that the task employed in our study does not require special domain knowledge; the participants in fact performed really well on the top-view tasks. User expertise in the task was also reflected in their ability to accurately differentiate between varying AI performance, as evidenced by their agreement ratio with the AI (Fig. 2, b) and perceived trust in AI (Fig. 4, c) decreased when AI performance was low, showing that the participants had enough expertise in the task to make an accurate judgement of the true accuracy of the AI.

5.2 The Impact of AI Accuracy

We observed a significant effect of AI performance on perceived reliance, perceived trust, perceived agreement with AI, and confidence in response (Fig. 4). As AI performance increased, the user perceived reliance, perceived trust, perceived agreement with AI also increased. This finding is consistent with prior work and Hypothesis 2 (increasing AI performance increases participant trust for AI). Furthermore, this result shows that even though the participants were not informed the AI performance, they were able to figure out the quality of the AI based on the quality of its suggestions and adjust their trust and reliance accordingly. This is also demonstrated through the user-AI agreement metric as participants agreed more with the AI suggestion when the AI performance was better (Fig. 2, b). This result reflects that the participants were able to maintain their agency and did not blindly follow nor ignore the AI suggestion. Additionally, user agency is particularly apparent as participants performed well on the top-view tasks regardless of the accuracy of the AI suggestion. Participants in top-view tasks had an average final response accuracy of $0.98(SD = 0.08)$, $0.98(0.05)$, and $0.94(SD = 0.15)$ when interacting with the AI with high, medium, and low performance, respectively. In bottom-view tasks, participants had an average final response accuracy of $0.98(SD = 0.09)$, $0.84(SD = 0.26)$, and $0.75(SD = 0.28)$ when interacting with the AI with high, medium, and low performance, respectively. This shows that participants benefited from the AI suggestions even when it was inaccurate or inconsistent. Participants' description of their problem-solving strategy collected in the post-study survey aligned with the quantitative finding. While one participant assigned to the low performance AI reported that they "*no longer trust[ed] [the] AI after see[ing] its error on the first task*", most participants with low performance or inconsistent agents reported to have found the AI "*suggestion to be helpful cues*", and "*used the AI* *suggestion to help [them] visualize better*" or "*utilized the AI's chosen orientation but chose the colors of the blocks [themselves] because [they] felt that some of the AI's selected colors were inaccurate.*"

5.2.1 The Importance of Consistency in AI Performance. We found that participants agreed less with the medium performance AI in the bottom-view tasks than the top-view tasks (Fig. 2, b). This difference did not occur in participants interacting with the low performance AI nor the high performance AI. One potential explanation is that when the AI is inconsistent regarding its performance (the medium performance AI makes minor errors in four out of ten trials), people are more cautious towards trust for the AI. As a result, when the task is more difficult, they are more hesitant towards trusting the AI, even though more trust in the AI would lead to a higher performance in this case.

5.3 Eye Gaze as a Window to Understand User Reliance on AI

In this study, we employed three different metrics to measure user reliance on AI. One, the degree of agreement between the final user response and the AI suggestion as a task-based metric for reliance—the higher the agreement ratio, the higher the reliance. Two, the participant's self-reported perceived reliance level on AI as a user perception-based metric of reliance. Last, the percent gaze duration on the AI suggestion during the task as a potential behavior-based metric of reliance. In our results, we found that duration on AI is positively correlated with perceived reliance (Fig. 5, b) and user agreement with the AI response. This finding is consistent with Hypothesis 3 (the longer the gaze duration on AI, the higher the reliance on AI). This result follows the gaze cascade model, after the AI suggestion appears, as the user spends more time looking at the AI suggestion, their preference for the AI suggestion increases, which increases the amount of time they spend with the AI suggestion, which in turn increases user exposure to the AI suggestion, forming a positive feedback that leads them to agree more with the AI suggestion. In addition to helping understand user reliance on AI, this behavioral metric allows for real-time estimation of user reliance, which can be used to enable adaptive human-AI collaboration.

5.3.1 Nuanced Differences between Reliance and Trust. We found that duration on AI is not correlated with perceived trust (Fig. 5, c), even though it has a strong positive correlation with perceived reliance (Fig. 5, b). This result is inconsistent with Hypothesis 4 (the longer duration on AI, the higher the trust in AI). This result demonstrates that there exists some nuanced differences between reliance and trust in human-AI interaction. One possible explanation is that reliance is how much the user think the AI was involved in the formation of their final response, whereas trust has more emotional value and is about how much the user is willingly to use the AI in the formation of their final response. More research is need to further characterize the differences between user trust and reliance in human-AI interactions.

5.3.2 How AI is Used in Decision Making. The participants' qualitative accounts of their problem-solving strategies may explain the difference in total percent duration on AI in phase 2 and phase 3; out of the 24 participants who mentioned AI in the strategy, 18 of them wrote something along the lines of "*complete[d] the task myself, and then compare[d] [it] to the AI answer*" and "*complet[ed] the task on my own first and then watch[ed] the AI result to check if they are the same*". Additionally, this use of AI for answer checking is coherent with the finding that the participants were more confident in their response when they had high perceived agreement with the AI suggestion (Section 4.5.3).

5.4 Design Implications for Enabling Adaptive Human-AI Collaboration

Our empirical findings have design implications for enabling real-time adaptive human-AI collaboration. Most prior works assess reliance, or more broadly trust, retrospectively through measures such as user-AI agreement and user perceptions. In this work, we show how gaze duration on AI is strongly correlated with perceived reliance and user-AI task agreement. This relationship between eye gaze and user reliance on AI opens the possibility of real-time adaptive human-AI collaboration through gaze monitoring. Prior research in the domain of human-robot collaboration has demonstrated how a robot can anticipate its human collaborator's task needs based on the observation of the collaborator's eye gaze and provide anticipatory robot assistance [29]. Similarly, gaze awareness has the potential to enhance human-AI team performance. As an example, an AI agent may based on observed gaze behavior provide timely feedback to help its user better calibrate their reliance on the AI for more optimal task performance. More research is needed to investigate the design of different strategies that AI may use to productively shape user reliance and to understand the effects of real-time gaze awareness on calibrated user reliance on AI.

Beyond user reliance, eye tracking and pupillometry have been used to gauge the attentional, cognitive, and emotional state of a user [12, 18]; for instance, pupil diameter, eyelid closure, and gaze entropy [71] are highly correlated with a person's cognitive load [84]. A holistic view of the user state can further allow the AI to adaptively assist the user in complex decision making. For example, real-time gaze attention information (fixation location and pattern) can be used to guide AI behavior to direct the user's attention to places where the user might have overlooked. The user's cognitive load as estimated by pupil diameter and gaze entropy can be used to regulate the amount of information provided by the AI. All in all, this work contributes an initial exploration of how eye behaviors may be incorporated into the computational characterization and estimation of the holistic user state and paves the way for enabling productive human-AI collaboration.

5.5 Limitations and Future Work

There are limitations to this study that call for further investigations. First, the spatial reasoning task used in this study is a non-critical task. Though we attempted to motivate the participants with an extra monetary reward given to top 5% of performers, the nature of the task is still low-stakes. The participants' interaction with the AI and their level of reliance on the AI may be different if they were in a higher stakes scenario. Future work may use our research paradigm to study human-AI collaboration in critical, high-stakes domains such as diagnostic radiology and autonomous driving. Second, the results presented in this paper are based on a one-session experimental study. Though our results suggest a subtle difference between human-AI trust and reliance, particularly in their relationship with gaze duration on the AI, it is unclear if this difference will continue to exist over multiple interaction sessions. Finally, we only focused on using gaze duration in understanding the complex process of human-AI collaboration. Future work should investigate other aspects of gaze behavior. For instance, a temporal analysis of gaze shift patterns may provide a more fine-grained view of how and when people turn to the AI suggestion. Future research may also use gaze tracking to explore other aspects of human-AI collaborative decision-making, evaluating user's ability in completing a task and their contribution to the task.

6 CONCLUSION

In this paper, we present empirical findings from an in-person user study that relates eye gaze and human reliance on AI during assisted decision-making in a spatial reasoning task. Our work shows the potential for eye gaze to be used to behaviorally understand how people consider AI suggestions in a human-AI teaming context. In particular, our findings indicate a strong positive correlation between total percent gaze duration on the AI suggestion and task-based user-AI agreement and self-reported perceived reliance. These findings present eye gaze as a plausible source of information for estimating human reliance on AI computationally in real-time, which can be used toward enabling adaptive human-AI collaboration that avoids sub-optimal collaboration where people over- or under-rely on AI assistance. Additionally, our results suggest nuanced differences between the conceptions of user trust and reliance. We also find that participants tend to maintain agency, particularly when the task is easy and when AI performance is low or inconsistent. Altogether, this work offers an empirical understanding of how people consider AI suggestions during spatial reasoning through gaze behavior and points toward the possibility of gaze-aware, adaptive human-AI collaboration.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation award #1840088.

REFERENCES

[1] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–20.

[2] A Selin Atalay, H Onur Bodur, and Dina Rasolofoarison. 2012. Shining in the center: Central gaze cascade effect on product choice. *Journal of Consumer Research* 39, 4 (2012), 848–866.

[3] Nora Balfe, Sarah Sharples, and John R Wilson. 2018. Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human factors* 60, 4 (2018), 477–495.

[4] 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.

[5] James H Bertera and Keith Rayner. 2000. Eye movements and the span of the effective stimulus in visual search. *Perception & psychophysics* 62, 3 (2000), 576–585.

[6] D Birchall. 2015. Spatial ability in radiologists: a necessary prerequisite? *The British journal of radiology* 88, 1049 (2015), 20140511.

[7] Eric Bogert, Aaron Schecter, and Richard T Watson. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports* 11, 1 (2021), 1–9.

[8] Ali Borji and Laurent Itti. 2014. Defending Yarbus: Eye movements reveal observers' task. *Journal of vision* 14, 3 (2014), 29–29.

[9] 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.

[10] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[11] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety science* 47, 9 (2009), 1260–1270.

[12] James F Cavanagh, Thomas V Wiecki, Angad Kochar, and Michael J Frank. 2014. Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General* 143, 4 (2014), 1476.

[13] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

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

[15] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[16] 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* 58, 6 (2003), 697–718.

[17] John M Findlay and Iain D Gilchrist. 1998. Eye guidance and visual search. In *Eye guidance in reading and scene perception*. Elsevier, 295–312.

[18] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. 2007. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin* 133, 4 (2007), 694.

[19] Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*. 1–8.

[20] Mackenzie G Glaholt and Eyal M Reingold. 2009. Stimulus exposure and gaze bias: A further test of the gaze cascade model. *Attention, Perception, & Psychophysics* 71, 3 (2009), 445–450.

[21] S Larry Goldenberg, Guy Nir, and Septimiu E Salcudean. 2019. A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology* 16, 7 (2019), 391–403.

[22] Fu Guo, Yi Ding, Weilin Liu, Chang Liu, and Xuefeng Zhang. 2016. Can eye-tracking data be measured to assess product design?: Visual attention mechanism should be considered. *International Journal of Industrial Ergonomics* 53 (2016), 229–235.

[23] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billinghurst. 2019. In ai we trust: Investigating the relationship between biosignals, trust and cognitive load in vr. In *25th ACM Symposium on Virtual Reality Software and Technology*. 1–10.

[24] Mary M Hayhoe. 2017. Vision and action. *Annual review of vision science* 3 (2017), 389–413.

[25] Mary Hegarty, Madeleine Keehner, Peter Khooshabeh, and Daniel R Montello. 2009. How spatial abilities enhance, and are enhanced by, dental education. *Learning and Individual Differences* 19, 1 (2009), 61–70.

[26] John M Henderson and Andrew Hollingworth. 1998. Eye movements during scene viewing: An overview. *Eye guidance in reading and scene perception* (1998), 269–293.

[27] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[28] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology* 6 (2015), 1049.

[29] Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 83–90.

[30] Shigao Huang, Jie Yang, Simon Fong, and Qi Zhao. 2020. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer letters* 471 (2020), 61–71.

[31] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

[32] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[33] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[34] 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.

[35] 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.

[36] Vijaya Lakshmi and Jacqueline Corbett. 2020. How artificial intelligence improves agricultural productivity and sustainability: A global thematic analysis. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

[37] Michael F Land and Mary Hayhoe. 2001. In what ways do eye movements contribute to everyday activities? *Vision research* 41, 25-26 (2001), 3559–3565.

[38] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[39] Sophie Lemonnier, Roland Brémond, and Thierry Baccino. 2014. Discriminating cognitive processes with eye movements in a decision-making driving task. *Journal of Eye Movement Research* 7, 4 (2014).

[40] Firas Lethaus and Jürgen Rataj. 2007. Do eye movements reflect driving manoeuvres? *IET Intelligent Transport Systems* 1, 3 (2007), 199–204.

[41] Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science advances* 6, 7 (2020), eaaz0652.

[42] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *arXiv preprint arXiv:2101.05303* (2021).

[43] Geoffrey R Loftus. 1981. Tachistoscopic simulations of eye fixations on pictures. *Journal of Experimental Psychology: Human Learning and Memory* 7, 5 (1981), 369.

[44] 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.

[45] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[46] Maria Laura Mele and Stefano Federici. 2012. A psychotechnological review on eye-tracking systems: towards user experience. *Disability and Rehabilitation: Assistive Technology* 7, 4 (2012), 261–281.

[47] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[48] Takashi Mitsuda and Mackenzie G Glaholt. 2014. Gaze bias during visual preference judgements: Effects of stimulus category and decision instructions. *Visual Cognition* 22, 1 (2014), 11–29.

[49] Chiyomi Miyajima, Suguru Yamazaki, Takashi Bando, Kentarou Hitomi, Hitoshi Terai, Hiroyuki Okuda, Takatsugu Hirayama, Masumi Egawa, Tatsuya Suzuki, and Kazuya Takeda. 2015. Analyzing driver gaze behavior and consistency of decision making during automated driving. In *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1293–1298.

[50] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. 2020. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. *arXiv preprint arXiv:2007.12358* (2020).

[51] Róbert Móro, Jakub Daráz, and Mária Bieliková. 2014. Visualization of Gaze Tracking Data for UX Testing on the Web.. In *HT (Doctoral Consortium/Late-breaking Results/Workshops)*.

[52] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

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

[54] Harikumar Pallathadka, Malik Mustafa, Domenic T Sanchez, Guna Sekhar Sajja, Sanjeev Gour, and Mohd Naved. 2021. Impact of machine learning on management, healthcare and agriculture. *Materials Today: Proceedings* (2021).

[55] Alexandra Papoutsaki. 2015. Scalable Webcam Eye Tracking by Learning from User Interactions. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 219–222.

[56] Alexandra Papoutsaki, James Laskey, and Jeff Huang. 2017. Searchgazer: Webcam eye tracking for remote studies of web search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 17–26.

[57] Alison Parkes. 2017. The effect of individual and task characteristics on decision aid reliance. *Behaviour & Information Technology* 36, 2 (2017), 165–177.

[58] Rik Pieters and Luk Warlop. 1999. Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of research in Marketing* 16, 1 (1999), 1–16.

[59] Stefanie Pietsch and Petra Jansen. 2012. Different mental rotation performance in students of music, sport and education. *Learning and Individual Differences* 22, 1 (2012), 159–163.

[60] Michael I Posner and Steven E Petersen. 1990. The attention system of the human brain. *Annual review of neuroscience* 13, 1 (1990), 25–42.

[61] Julia M Puaschunder, Josef Mantl, and Bernd Plank. 2020. Medicine of the future: The power of Artificial Intelligence (AI) and big data in healthcare. *RAIS Journal for Social Sciences* 4, 1 (2020), 1–8.

[62] Qing-Xing Qu, Le Zhang, Wen-Yu Chao, and Vincent Duffy. 2017. User experience design based on eye-tracking technology: a case study on smartphone APPs. In *Advances in applied digital human modeling and simulation*. Springer, 303–315.

[63] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

[64] Erika Rogers. 1996. A study of visual reasoning in medical diagnosis. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. 213–218.

[65] Toshiki Saito, Ryunosuke Sudo, and Yuji Takano. 2020. The gaze bias effect in toddlers: Preliminary evidence for the developmental study of visual decision-making. *Developmental science* 23, 6 (2020), e12969.

[66] Rafael Santos, Nuno Santos, Pedro M Jorge, and Arnaldo Abrantes. 2014. Eye gaze as a human-computer interface. *Procedia Technology* 17 (2014), 376–383.

[67] Elizabeth R Schotter, Raymond W Berry, Craig RM McKenzie, and Keith Rayner. 2010. Gaze bias: Selective encoding and liking effects. *Visual Cognition* 18, 8 (2010), 1113–1132.

[68] D Selvathi and A Aarthy Poornila. 2017. Breast cancer detection in mammogram images using deep learning technique. *Middle-East Journal of Scientific Research* 25, 2 (2017), 417–426.

[69] Weston Sewell and Oleg Komogortsev. 2010. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. 3739–3744.

[70] Mona SharifHeravi, John R Taylor, Christopher J Stanton, Sandra Lambeth, and Christopher Shanahan. 2020. It's a Disaster! Factors Affecting Trust Development and Repair Following Agent Task Failure. In *Proceedings of the 2020 Australasian Conference on Robotics and Automation (ACRA 2020), 8-10 December 2020, Brisbane, Queensland*.

[71] Brook Shiferaw, Luke Downey, and David Crewther. 2019. A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience & Biobehavioral Reviews* 96 (2019), 353–366.

[72] Shinsuke Shimojo, Claudiu Simion, Eiko Shimojo, and Christian Scheier. 2003. Gaze bias both reflects and influences preference. *Nature neuroscience* 6, 12 (2003), 1317–1322.

[73] Claudiu Simion and Shinsuke Shimojo. 2006. Early interactions between orienting, visual sampling and decision making in facial preference. *Vision research* 46, 20 (2006), 3331–3335.

[74] Ronal Singh, Tim Miller, Joshua Newn, Eduardo Velloso, Frank Vetere, and Liz Sonenberg. 2020. Combining gaze and AI planning for online human intention recognition. *Artificial Intelligence* 284 (2020), 103275.

[75] Rainer Stiefelhagen and Jie Yang. 1997. Gaze tracking for multimodal human-computer interaction. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, 2617–2620.

[76] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).

[77] Armin W Thomas, Felix Molter, Ian Krajbich, Hauke R Heekeren, and Peter NC Mohr. 2019. Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour* 3, 6 (2019), 625–635.

[78] David H Uttal and Cheryl A Cohen. 2012. Spatial thinking and STEM education: When, why, and how? In *Psychology of learning and motivation*. Vol. 57. Elsevier, 147–181.

[79] 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.

[80] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).

[81] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[82] Min Wang, Aya Hussein, Raul Fernandez Rojas, Kamran Shafi, and Hussein A Abbass. 2018. EEG-based neural correlates of trust in human-autonomy interaction. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 350–357.

[83] Philipp Wintersberger, Tamara von Sawitzky, Anna-Katharina Frison, and Andreas Riener. 2017. Traffic augmentation as a means to increase trust in automated driving systems. In *Proceedings of the 12th biannual conference on italian sigchi chapter*. 1–7.

[84] Chuhao Wu, Jackie Cha, Jay Sulek, Tian Zhou, Chandru P Sundaram, Juan Wachs, and Denny Yu. 2020. Eye-tracking metrics predict perceived workload in robotic surgical skills training. *Human factors* 62, 8 (2020), 1365–1386.

[85] 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.

[86] Thorsten O Zander, Matti Gaertner, Christian Kothe, and Roman Vilimek. 2010. Combining eye gaze input with a brain–computer interface for touchless human–computer interaction. *Intl. Journal of Human–Computer Interaction* 27, 1 (2010), 38–51.

[87] Rencheng Zheng, Kimihiko Nakano, Hiromitsu Ishiko, Kenji Hagita, Makoto Kihira, and Toshiya Yokozeki. 2015. Eye-gaze tracking analysis of driver behavior while interacting with navigation systems in an urban area. *IEEE Transactions on Human-Machine Systems* 46, 4 (2015), 546–556.

A APPENDIX

Table 1. Results from pairwise comparisons using Tukey's HSD test for interaction effect of task difficulty and phase of task on the percent gaze duration on AI suggestion

Phase	Task	-Phase	-Task	p-value	Significant
2	bottom	3	bottom	0.974	No
2	bottom	2	top	<.001	Yes
2	bottom	3	top	0.990	No
3	bottom	2	top	<.001	Yes
3	bottom	3	top	0.999	No
2	top	3	top	<.001	Yes

Received January 2022; revised April 2022; accepted May 2022

Table 2. Results from pairwise comparisons using Tukey's HSD test for interaction effect of task difficulty and phase of task on the percent gaze duration on AI suggestion

Phase	AI performance	-Phase	-AI performance	p-value	significant
2	high	3	high	0.055	No
2	high	2	med.	0.921	No
2	high	3	med.	0.614	No
2	high	2	low	0.988	No
2	high	3	low	1.000	No
3	high	2	med.	0.178	No
3	high	3	med.	1.000	No
3	high	2	low	0.326	No
3	high	3	low	0.592	No
2	med.	3	med.	<.001	Yes
2	med.	2	low	0.999	No
2	med	3	low	0.977	No
3	med	2	low	0.292	No
3	med	3	low	0.529	No
2	low	3	low	0.972	No

Table 3. Results from pairwise comparisons using Tukey's HSD test for interaction effect of AI accuracy level and task difficulty on the user AI agreement rate

AI performance	Task	-AI performance	-Task	p-value	Significant
high	bottom	high	top	1.000	No
high	bottom	med.	bottom	0.007	Yes
high	bottom	med.	top	0.996	No
high	bottom	low	bottom	<.001	Yes
high	bottom	low	top	<.001	Yes
high	top	med.	bottom	1.000	No
high	top	med.	top	0.006	Yes
high	top	low	bottom	0.994	No
high	top	low	top	<.001	Yes
high	bottom	low	top	<.001	Yes
med.	bottom	med.	top	0.040	Yes
med.	bottom	low	bottom	<.001	Yes
med.	bottom	low	top	<.001	Yes
med.	top	low	bottom	<.001	Yes
med.	top	low	top	<.001	Yes
low	bottom	low	top	0.684	No

Table 4. Results from pairwise comparisons using Tukey's HSD test for interaction effect of AI accuracy level and task difficulty on perceived agreement with AI suggestion

AI performance	Task	-AI performance	-Task	p-value	Significant
high	bottom	high	top	.449	No
high	bottom	med.	bottom	<.001	Yes
high	bottom	med.	top	.031	Yes
high	bottom	low	bottom	<.001	Yes
high	bottom	low	top	<.001	Yes
high	top	med.	bottom	<.001	Yes
high	top	med.	top	<.001	Yes
high	top	low	bottom	<.001	Yes
high	top	low	top	<.001	Yes
med.	bottom	med.	top	.504	No
med.	bottom	low	bottom	<.001	Yes
med.	bottom	low	top	<.001	Yes
med.	top	low	bottom	<.001	Yes
med.	top	low	top	<.001	Yes
low	bottom	low	top	.715	No

Table 5. Results from pairwise comparisons using Tukey's HSD test for interaction effect of AI accuracy level and task difficulty on perceived trust in AI

AI performance	Task	-AI performance	-Task	p-value	Significant
high	bottom	high	top	.352	No
high	bottom	med.	bottom	<.001	Yes
high	bottom	med.	top	.015	Yes
high	bottom	low	bottom	<.001	Yes
high	bottom	low	top	<.001	Yes
high	top	med.	bottom	<.001	Yes
high	top	med.	top	<.001	Yes
high	top	low	bottom	<.001	Yes
high	top	low	top	<.001	Yes
med.	bottom	med.	top	.417	No
med.	bottom	low	bottom	.009	Yes
med.	bottom	low	top	<.001	Yes
med.	top	low	bottom	<.001	Yes
med.	top	low	top	<.001	Yes
low	bottom	low	top	.623	No