How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration

SHIYE CAO*, Johns Hopkins University, USA
CATALINA GOMEZ*, Johns Hopkins University, USA
CHIEN-MING HUANG, Johns Hopkins University, USA

Human cognitive and decision-making abilities depreciate under pressure, motivating the emergence of artificial intelligence (AI) systems as decision support tools to assist people in performing tasks under stress. In this work, we study human decision-making behavior and task performance under time pressure—induced from limited *initial observation time* (time to perform the task before providing an initial response without AI input) and *final decision time* (time to weigh an AI's suggestion before reaching a collective human-AI team answer)—for spatial reasoning and count estimation tasks. Our results show that, while the impact of initial observation time on AI-assisted decision-making was dependent on task nature, participants were more likely to follow AI suggestions when they were provided with longer final decision time; moreover, although participants generally tended to adhere to their initial responses, they had more agency when they were more logically engaged in a task. Our results offer a nuanced understanding of human-AI collaboration under time pressure in different phases of the decision-making process.

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

Additional Key Words and Phrases: human-AI interaction, decision support tools, time pressure, decision-making

ACM Reference Format:
Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 277 (October 2023), 26 pages. https://doi.org/10.1145/3610068

1 INTRODUCTION

Decisions in real-world scenarios such as aviation [59], medicine [22], and finance [33] often have to be made under intense time pressure—e.g., brokers trading stocks or radiologists interpreting emergency room X-rays; in fact, radiologists' overall workload measured in terms of relative value units during on-call hours has quadrupled [7], causing them to feel added stress from time pressure—or in their words, "having too great an overall volume of work" while "under pressure to meet deadlines" [22]. Previous research has shown that time pressure lowers people's cognitive complexity and flexibility, negatively affecting their decision-making and decreasing the quality of their task performance [29, 33, 39, 51, 63]; in the context of radiology, reckless reading lawsuits proclaiming that radiologists have missed findings due to insufficient time spent viewing imaging

*Both authors contributed equally to this research.

Authors' addresses: Shiye Cao, scao14@jhu.edu, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA, 21218; Catalina Gomez, cgomezc1@jhu.edu, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA, 21218; Chien-Ming Huang, cmhuang@cs.jhu.edu, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA, 21218.

2573-0142/2023/10-ART277
https://doi.org/10.1145/3610068

results have become increasingly common [1]. Recent advances in artificial intelligence (AI) have enabled its application as a decision support tool in diverse real-world scenarios, including stressful tasks with high stakes or limited time, such as financial trading or medical diagnosis [14, 35, 55, 74]. Additionally, AI systems are unaffected by any type of stress and are optimized to solve specific tasks; thus they have the potential to assist humans effectively in stressful decision-making situations. A common implementation of AI-assisted decision-making is to have AI systems provide task predictions and recommendations, with humans still making the final decisions [5]. The ideal outcome of such human-AI collaboration is an improvement in overall decision quality such that the team performs better than both the human and AI system alone. However, AI systems are not flawless, making the development of appropriate trust in and reliance on such systems critical in facilitating the achievement of improved team performance [10, 80].

Effective human-AI teaming is challenging to design and achieve [2, 77]; to enable successful human-AI collaboration, previous research has investigated how a range of factors—including model capabilities, user backgrounds, and task contexts—may shape people's performance with and trust in an AI system. For example, prior works have explored how information elements, such as explanations of model outputs [6, 11], performance and confidence values [78, 81], and details about training data and model architecture [12, 66], or user involvement in joint decision making tasks [20] may influence human-AI collaboration. People's domain expertise and knowledge of AI technology [45, 71], as well as their math and logic skills [66], have also been studied to understand the complex interplay between human cognition and data-driven AI models. Likewise, contexts such as task complexity and time constraints play a key role in shaping the collaborative dynamics between humans and AI systems [62]; for instance, it has been demonstrated that when people are under time pressure they are more likely to over-trust automation in a single-phase visual inspection decision-making task [54, 59].

Building on the growing body of research on user trust and reliance in human-AI interactions, we sought to further understand 1) how time pressure may influence people's trust and reliance behaviors in AI-assisted decision-making tasks and 2) how these behavioral differences may affect task performance. In this work, we designed and conducted an online user study with participants recruited through convenience sampling in the local university community to investigate the effects of time pressure at more granular levels, considering both *initial observation time* (the time allotted to observe and perform the task before considering the AI's suggestion) and *final decision time* (the time allotted to consider the AI's suggestion and make a final decision). In the remaining sections, we refer to these time variables as *observation time* and *decision time,* respectively.

We contextualized our investigation in two visual interpretation tasks: a *spatial reasoning* task that involves spatial perception and memory to identify modified locations on a piece of paper after folding it (Fig. 1) and a *count estimation* task that requires focused attention to count and estimate the number of items in an image (Fig. 2). Although people know how to perform such tasks, their ability to complete them accurately can be hindered by stress and time pressure [21, 31, 49]. We were interested in whether this effect might lead to greater user reliance on AI assistance while performing tasks under time pressure. By not requiring users have the specific knowledge necessary to evaluate suggestions from an AI assistant, these two experimental tasks allowed us to explore how the nature of tasks requiring different abilities influenced the effects of time pressure in AI-assisted decision-making.

Our investigation revealed that 1) the impact of observation time on AI-assisted human decision-making was dependent on the nature of the task in question; 2) the more decision time users had, the more likely they were to follow the AI's suggestions in their final responses; 3) logical engagement in the task discouraged users from following AI suggestions even when there were potential benefits. Our results contribute a deeper understanding of how time pressure may regulate people's trust

in and reliance on an AI assistant during different phases of the collaborative decision-making process. Our findings have implications for the design of human-AI collaboration when strict time constraints are unavoidable, demonstrating the potential for strategic redistribution of task time between initial observation time and final decision time to facilitate superior calibration of user reliance on an AI assistant. Next, we review relevant background and related work that helped situate this investigation.

2 BACKGROUND AND RELATED WORK

2.1 Time Pressure in Human Decision-Making

Time pressure, which is distinct from time constraint, is a stressor that originates from a fear of failure to complete a task on time [46]; more specifically, time pressure is caused by time constraint, but it is possible to have a time constraint without time pressure and its associated stress. Psychological studies have shown that stress directly affects specific regions of the brain, including the hippocampus, prefrontal cortex, striatum, and insula [34, 41, 52, 69]; as a result, stress impairs cognitive function—reducing the amount of attention one can devote to information processing—inhibits working memory, and increases one's vulnerability to cognitive overload [19, 30]. In turn, the quality of human decisions made under stress is adversely influenced, as has been observed in routine activities such as public speaking or presenting course exams [40, 44]. Stress also changes people's decision-making patterns; studies have shown that stress leads to decisions that are rushed, unsystematic, and lacking full consideration of available options [29, 39, 63]. Under time pressure, people focus more on negative information and effects than positive ones when considering options with associated risks or when concerning their preferences [32, 67, 73]. Additionally, gender differences can affect decision outcomes in simulated gambling tasks; women under stress tend toward less risky options, while men under stress tend to choose riskier options [42]. Overall, time pressure and stress impair cognition and decision-making ability, which in turn causes decreased task performance, particularly in tasks that require "attentional control" or "effortful cognitive processing" [24, 65]. Furthermore, time pressure has been observed to increase people's confidence when making easier judgments, but reduces their confidence in more difficult cases in the context of high-fidelity clinical risk assessment [76]. In our study, we selected two tasks, a count estimation task and a spatial reasoning task, that subjects would likely perform poorly under stress to determine if an AI assistant might help improve task performance under time pressure [51].

2.2 User Trust and Reliance in Human-AI Assisted Decision-Making

AI systems are, and will continue to be, imperfect. Therefore it is critically important to know when and when not to trust in or rely on AI in a joint decision-making collaboration, as under-reliance and over-reliance hinders human-AI team performance and can have severe consequences in critical decisions [37, 57]. Previous works have evaluated how different capabilities of an AI model (e.g., performance [78] and explanations about its predictions [36, 81] and interactive mechanisms to provide feedback [64] or guide the AI's predictions [20]), user-related factors (e.g., domain knowledge [18] and familiarity with AI techniques [28]), and task-related factors (e.g., task difficulty [10, 48]) may affect user trust in and reliance on AI assistance in human-AI collaborative decision-making processes. For instance, providing users with more information about an AI model increases user trust in the AI, but can also reduce human agency during decision-making [36]; higher task familiarity leads users to rely less on AI recommendations, even though they may still self-report to have greater trust in the AI [60]; and higher objective task difficulty increases user tendency to rely on decision aids [48].

Studies have also explored the effects of slowing users down [8, 47]; cognitive forcing functions—such as asking people to provide an initial response before being shown an AI suggestion, delaying the presentation of AI suggestions, or letting users decide whether or not they want to see AI suggestions in the first place—reduces user over-reliance on AI [8, 47] at the cost of decreased user trust in and preference for the decision-support system [8]. Moreover, interaction schemes meant to help increase user efficiency can also induce different reliance behaviors; in a clinical text annotation context, a decision aid with fully pre-populated annotation suggestions led to greater user reliance than a decision aid that provided label recommendations for mapping concepts [38]. In this study, we focus on examining the effects of time pressure, a task-related factor, on user trust in and reliance on AI suggestions.

2.3 Time Constraints and Time Pressure in Human-AI Assisted Decision-Making

AI recommendations have the potential to effectively assist human decision-making in time-constrained settings such as clinical practice, where actionable decisions must be determined in a timely manner. However, the successful integration of such decision support tools into human workflows requires careful consideration of user expectations and contextual factors under varying time constraints. For instance, a recent study [27] reports that clinicians who already have limited time with patients may not be able to make in-the-moment determinations of trust in suggestions provided by an ML decision support tool when selecting the optimal treatment for a patient. Moreover, the presence of time pressure increases how frequently people use an intelligent voice assistant in a creative task, which overall negatively affects the creative outcome of that task [62]. Studies have also found that users are more likely to adopt automation suggestions in visual inspection tasks when the amount of time they have to observe the task image (observation time) is limited [54, 68]; this increased reliance on automation support leads to increased performance when the aid is reliable and decreased performance when the automation's performance is less reliable. A different effect on visual search performance was observed when time pressure did not alter the use of automation support—rather, only the negative effects of time pressure on sensitivity were mitigated when users worked with a decision support system (without improving performance) [56]. Placing constraints on decision time has also been explored as an active mitigation strategy for reducing anchoring bias—people tend to affix their responses to those of an AI after being introduced to its predictions [53]. One study found that increasing the amount of time allocated to consider the task and the AI prediction (decision time) decreased user reliance on the AI and reduced anchoring bias; this finding motivated the design of a confidence-based time allocation strategy, which, with an explanation, effectively de-anchored participants and improved the AI model's performance when it had low confidence and was incorrect.

As we continue to develop and deploy AI-assisted decision-making systems for a wider array of task contexts, it is imperative to understand time pressure's effects on user reliance on and trust in AI systems. The present work builds on previous findings and seeks to further understand the effects of constraining observation time, constraining decision time, and any resulting interaction effects on people's tendency to follow AI suggestions, their perceptions of those suggestions, and overall task performance.

3 METHODS

3.1 Hypotheses

We designed a user study to evaluate the effect of time pressure when completing spatial reasoning and count estimation tasks in an AI-assisted decision-making scenario. We hypothesized that adding time constraints for users at different stages in the task completion process (manipulating

observation time and decision time) would affect their engagement with and attitude toward the AI assistant and, as a result, their task performance. More specifically, we formulated the following hypotheses:

- **H1**: With insufficient observation time, regardless of decision time, users will agree with the AI more than if they had sufficient observation time. Prior studies have found that users have a higher probability of complying with automation recommendations (shown before engagement with the task) when they have less time to observe the task image in a visual search task [54, 68]; for our purposes, the AI suggestion is shown later in the decision-making process, but we believe this previously observed effect will extend into our study.
- **H2**: With insufficient decision time, regardless of observation time, users will agree with the AI more than if they had sufficient decision time. This hypothesis is informed by prior work [53] on how a time allocation strategy may mitigate anchoring bias, suggesting that users tend to adjust their responses away from an AI suggestion with more decision time. Thus, with insufficient decision time, we expect participants to have a higher probability of relying upon, adopting, and trusting the AI suggestions than if they had sufficient decision time.

We expected these two hypotheses to apply to both the spatial reasoning and count estimation tasks.

3.2 Experimental Tasks

Our study focused on investigating the dynamics of time pressure in human-AI collaboration using tasks that humans can perform, but may not execute well under time pressure. We chose two tasks—spatial reasoning and count estimation—that did not require special domain knowledge to complete and in which human performance under time pressure would be significantly impaired. Previous studies in human decision-making under pressure suggest that human performance in tasks that require "effortful cognitive processing" or "attentional control" is significantly impaired by time pressure [24, 65]; spatial reasoning tasks require three-dimensional spatial perception and "effortful cognitive processing" while count estimation tasks require "attentional control." Thus, we hypothesized that human performance would be significantly impaired under time pressure for these two tasks, allowing us to study how people may rely on an AI agent in completing the tasks.

- **Spatial Reasoning.** In this task, participants are presented with a sequence of images that show the folding of a square piece of paper. In the last image of the sequence, one hole is punched through all the paper layers. Participants must deduce where the holes are located in a 4-by-4 grid when the paper is completely unfolded (Fig. 1 shows an example). Task images were drawn from the Paper Folding Test data set from the Working Memory in Spanish–English and Chinese–English Bilinguals study [43]. Spatial skills—more specifically the mental rotation and recall of object positions in this task—are instrumental in many domains, such as civil, mechanical, and aerospace engineering [17, 23]. The inference of a three-dimensional context from a two-dimensional image as in our task is particularly crucial for radiologists and dentists reading medical images (e.g., CT and MRI scans, X-rays, ultrasounds) [16, 25].
- **Count Estimation.** In this task, participants are presented with an image containing a crowd of penguins. Participants are asked to estimate the number of penguins present in the image, including partially occluded penguins, as shown in Fig. 2. Task images were drawn from the penguin data set from the Counting in the Wild study [3]; we hand-selected task images from this data set to ensure that each was unique and avoided images with ambiguity in the number of penguins contained within. Attention to detail in multiple

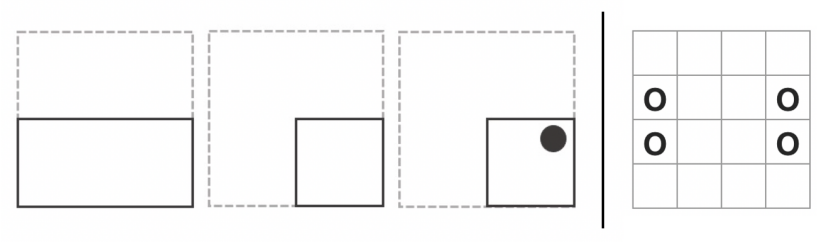

Fig. 1. Example image with two folds in the spatial reasoning task. The two leftmost squares show how the paper is folded. The square to the right of that shows the position of the hole. The right-most square is the solution, showing the position of the holes when the paper is unfolded.

Fig. 2. An instructive image from the count estimation task showing users that all penguins (marked with red dots), including occluded ones, should be counted. The task images in the practice round, calibration round, and main experiment did not have red dots on the penguins.

areas simultaneously is vital to visually estimating a quantity. Crowd counting with AI techniques has received much attention as it poses significant challenges to humans, such as scale variation and time consumption; an AI-assisted tool can therefore provide benefits in multiple applications, including video surveillance, urban planning, and wild animal population census and monitoring [13, 26, 50, 58].

3.3 Experimental Design

The study had a within-subjects 2 (observation time: insufficient and sufficient) × 2 (decision time: insufficient and sufficient) factorial design. We defined the time users had to observe the task image before providing an initial response as *initial observation time*. We defined observation time to only include the time in which users were exposed to the image, rather than the time they had to complete the task and provide an answer; this is because constraining the time to provide an answer would introduce the possibility of users being cut off while entering their responses or missing the opportunity to enter a response. Instead, we decided to control for observation time by manipulating the length of time that users had to look at the image. Even if they took more time

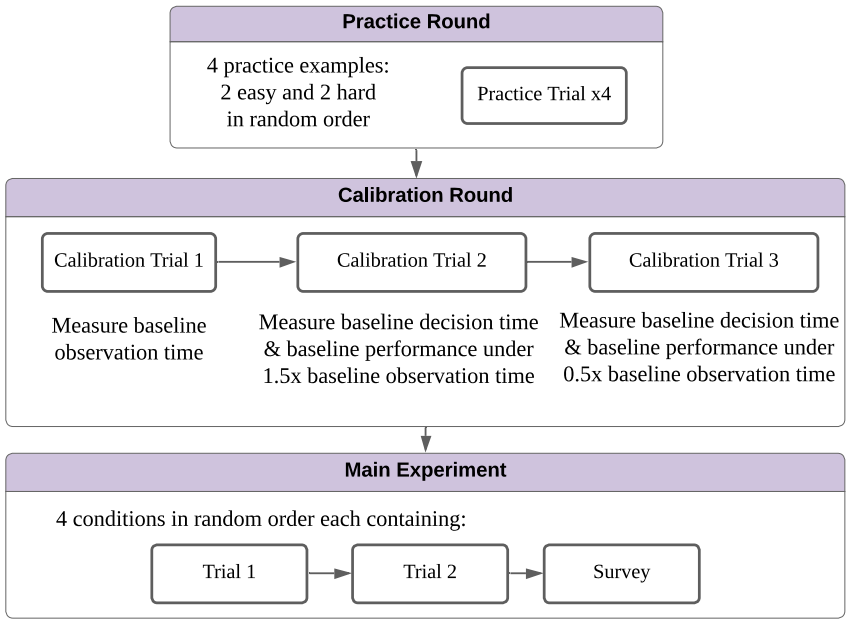

Fig. 3. An overview of our study. The experiment involved three stages: a practice round followed by a calibration round (in which the baseline decision times and baseline performances are determined) and then the main experiment.

to reason afterward, the time pressure effect was still in place and was in fact reinforced with the disappearance of the image.

We defined *final decision time* as the time users had to analyze and consider the AI's suggestion against their own initial response and to come up with a final team response. Participants were not given the option to continue to the next step until the allotted observation or decision time was over. Participants were also not allowed to go back to a previous step (i.e., change their answers) once their allotted observation or decision time was up or if they had already moved on to the next step.

3.3.1 Time Manipulation. In our study, participants were first given four practice examples that were both easier and harder than the actual test to become familiar with the task and its interface. We defined insufficient and sufficient time for the task's completion and subsequent decision-making based on each user's behavior in three calibration trials before the main experiment, allowing us to account for individual differences in problem-solving rather than applying fixed values for all the participants.

Calibration Trial 1. The goal of this trial was to measure the observation time participants needed to provide their initial answer without any time constraints, which we referred to as *baseline observation time.* In this trial, participants were not presented with any AI suggestions, nor were they asked to update their initial response; we were only interested in the time they needed to complete the task by themselves.

Calibration Trial 2. The goal of this trial was to measure the decision time participants needed to consider a suggestion and make any necessary changes to their initial answers when provided with sufficient observation time (baseline observation time $\times 1.5$), which we referred to as *baseline sufficient decision time.* In this trial, participants were given sufficient observation time and allowed

Table 1. Definition of time manipulation values for the main experiment. The three baseline times from the calibration rounds were used to manipulate the sufficient and insufficient times for task observation and decision-making in the main experiment.

Time	Insufficient Observation	Sufficient Observation
Insufficient Decision	0.5 × baseline observation time	1.5 × baseline observation time
	0.5 × baseline decision time under insufficient observation time	0.5 × baseline decision time under sufficient observation time
Sufficient Decision	0.5 × baseline observation time	1.5 × baseline observation time
	1.5 × baseline decision time under insufficient observation time	1.5 × baseline decision time under sufficient observation time

to consider a suggestion posed as originating from another participant and to modify their answers without time constraints. In this trial with sufficient observation time, the displayed suggestion was correct to avoid biased perceptions of the quality of the suggestions that could affect the overall perception of the suggestions in main experiment. We expected participants' decision time to be low because they would have more than enough time to complete the task and feel confident in their answers. Baseline sufficient decision time was used in the main experiment in conditions with sufficient observation time to calculate sufficient decision time (baseline sufficient decision time ×1.5) and insufficient decision time (baseline sufficient decision time ×0.5).

Calibration Trial 3. The goal of this trial was to measure the decision time participants needed to consider a suggestion and make any changes to their initial answers when provided with insufficient observation time (baseline observation time ×0.5), which we referred to as *baseline insufficient decision time*. In this trial, participants were given insufficient observation time and allowed to consider a suggestion posed as originating from another participant and to modify their answers without time constraints. In this trial, the displayed suggestion was slightly off, since the limited observation time might not be enough for participants to identify minor flaws without affecting their initial perception of the quality of the suggestion. We expected that decision time in this trial would be lengthier because participants might not have had enough time to complete the task on their own and would instead take advantage of seeing the image again. Baseline insufficient decision time was used in the main experiment for the conditions with insufficient observation time to calculate sufficient decision time (baseline insufficient decision time ×1.5) and insufficient decision time (baseline insufficient decision time ×0.5). See Table 1 for an illustration of our time manipulation design.

The overall process of the study is summarized in Fig. 3. The images in the practice round for the count estimation task had 4, 16, 60, and 62 penguins, while the calibration and main experiment images had between 29–49 penguins; the practice examples for the spatial reasoning task consisted of two trials with one fold and two trials with three folds, while the calibration and main experiment tasks all had two folds. We sought to control the difficulty level for both tasks such that they were neither too easy nor too difficult based on the number of folds in the spatial reasoning task and the number of penguins in the count estimation task. If the tasks were too easy, participants might be able to complete the task by themselves without considering the AI's suggestions, whereas if the tasks were too difficult, participants might rely on the suggestions blindly. Moreover, for the practice trials, second and third calibration trials, and main experiment trials, participants had the option to see the task image again for three seconds while considering the suggestion from the other participant/AI assistant; this option was added to encourage participants to reconsider their initial answers and the AI's suggestions.

3.3.2 AI Suggestion Generation. To promote the realism of the AI-assisted decision-making process, we experimentally adjusted the AI suggestions to be imperfect with a predetermined task performance slightly superior to that of humans alone as determined through a pilot study. All calculations of percent error in the spatial reasoning task were the number of cells that did not match the ground truth normalized with respect to the total number of cells (16) and reported as a percentage. Calculations of percent error in the count estimation task were the absolute difference in the counts normalized with respect to the ground truth of that specific task instance and reported as a percentage. For the spatial reasoning task, the simulated AI had an error range of 6.25–12.5%, with an overall mean of 7.03% and a standard deviation of 2.07% to keep the suggestions reasonable (equivalent to 1–2 cells out of 16 containing an extra hole or missing a hole); errors were fixed for every test example for each participant. For the count estimation task, the simulated AI had randomly assigned errors within the range 10–20% of the ground truth, with an overall mean of 14.94% and a standard deviation of 3.17%.

3.4 Measures

We used a set of objective and subjective measures to evaluate user behavior and perception, respectively, when interacting with the AI system under time pressure.

3.4.1 Behavioral Metrics. We adopted two behavioral indicators that have been used in prior research to capture participants' willingness to follow AI suggestions [78]:

- **Final Agreement.** This metric captures the percent difference between participants' final responses and the AI's suggestions. In the spatial reasoning task, the metric is calculated as the number of cells that are different between a user's final response and the AI suggestion, normalized with respect to the total number of cells (16 in our experimental task). In the count estimation task, final agreement is computed as the absolute numeric difference between a user's final response and the AI's suggestion, normalized with respect to the AI suggestion value.
- **Switch to AI.** This binary metric captures whether participants' final responses exactly matched the AI suggestions in each trial for cases in which their initial responses were different from the AI's suggestions.

3.4.2 Subjective Metrics. We defined two main subjective metrics collected after participants interacted with each AI agent:

- **Perceived Trust.** This metric aims to capture participants' self-reported trust of the AI agent's suggestions. Participants rated their agreement with the following statement on a 5-point Likert scale, from 1 (strongly disagree) to 5 (strongly agree): "I trusted the AI agent's suggestions."
- **Perceived AI Usefulness.** This metric captures participants' perception of the usefulness of the AI's suggestions in completing the task; improved perception of usefulness may be aligned with higher reliance on the agent. Participants rated their agreement with the following statement on a 5-point Likert scale, from 1 (strongly disagree) to 5 (strongly agree): "The AI agent's suggestions were useful."

3.4.3 Task Performance Metric.

- **Error Improvement.** This metric represents the difference between participants' initial level of error and their final level of error with respect to the ground truth. In the spatial reasoning task, initial and final response errors were defined as the number of cells that were different between user response and the ground truth, normalized with respect to the total number of cells (16 in our experiment task). In the count estimation task, errors were

computed as the absolute numeric difference between the initial or final count and the ground truth, normalized with respect to the ground truth value. In our results, error improvement is presented using percentages. Negative values for error improvement reflect that the accuracy of a participant's final response was lower than that of their initial response.

3.5 Study Procedure

The user interface for our study was implemented as a custom web application using the React[1] and Flask frameworks[2] and was deployed via Heroku[3]. Upon agreeing to participate in the study via informed consent within the web application, participants filled out a demographic survey, which asked for their gender, age, educational background, and familiarity with AI. Participants were randomly assigned to one of the two tasks with which to begin.

Participants were presented with the corresponding task instructions and four practice trials in which they had unlimited time to complete the task and consider the (correct) suggestions provided. Participants were told that the suggestions were from other participants who had previously completed the task. This setup was adopted to avoid users creating a mental model of the AI before reaching the main experiment. Upon completing the practice trials, participants continued to the three calibration trials detailed in Section 3.3.1 and then proceeded to the main experiment. As a screening measure for bots, a participant was only considered valid if they spent more than one second in Calibration Trial 1 and more than one second during the decision phase in Calibration Trials 2 or 3.

During the main experiment, participants were exposed to four conditions with manipulations of observation time and decision time (Table 1) in random order. Each condition consisted of two trials followed by a questionnaire regarding their experience and perception of the AI agent they had just interacted with. Before each condition, users were told explicitly that a new AI agent would assist them so that we could assess user perception in each condition.

Fig. 4 shows an example of the user interface in the main experiment. At the beginning of each trial, a countdown timer visual with the allotted observation time was presented. Once the time was up, the task image disappeared and a pop-up window prompted participants to input their initial answer. After a participant confirmed their answer, an AI suggestion was presented next to their initial answer in the same pop-up window and a second countdown timer with the allotted decision time became visible. If the decision time left was greater than three seconds, a button that allowed participants to view the task image again for three seconds was active; otherwise it was disabled. Within the decision time frame, participants could enter their final response via an input grid or box within the pop-up window, which defaulted to their initial response. They could also adopt the AI suggestion directly via a button. If participants did not perform any update, their initial answers were locked in as their final answers. We accounted for task image loading time in our implementation and only started the timers after the images had fully loaded. Participants could not pause or reset timers. The same procedure as described above was then repeated for the second task.

This study was approved by our institutional review board. On average, participants took 21 minutes to complete the study and were compensated with a $5 gift card.

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

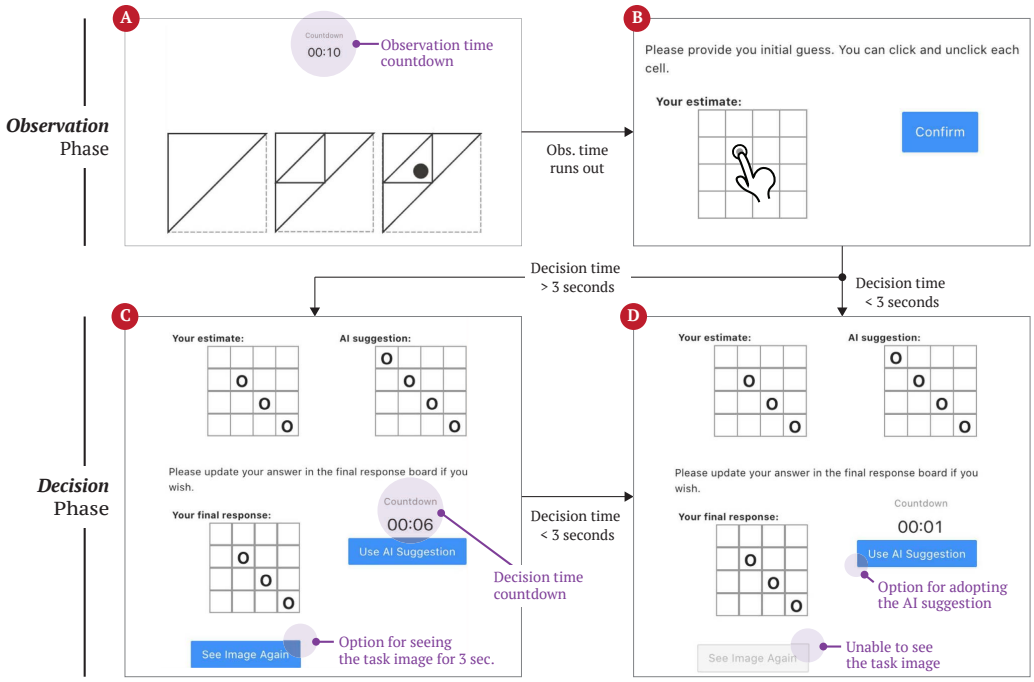

Fig. 4. Overview and steps in the user interface for each task, illustrated with the spatial reasoning task. A) Once the task loads, observation time countdown begins. Participants perform the task without assistance from the AI. B) When observation time runs out, the task image hides and participants are asked to enter an initial response. C) Once participants submit their initial response, the AI suggestion is shown and the decision time countdown begins. Participants are able to update their responses or adopt the AI suggestion if they wish. D) During the decision phase, participants have the choice of viewing the task image again for three seconds if and only if there are more than three seconds left on the decision time countdown. Once the decision countdown ends, participants are not able to make additional changes to their responses.

3.6 Participants

A total of 53 participants were recruited through convenience sampling in the local university community, 40 of whom provided valid data points according to our response screening strategy in the calibration trials (described in Section 3.5). Out of the 40 participants, 19 participants identified as male, 20 as female, and 1 as other. The valid participants' ages ranged between 20 and 35 years ($M = 24.13, SD = 3.20$). Participants self-reported to have an above-average familiarity with AI technology ($M = 3.49, SD = 1.11$) on a scale from 1 to 5, where 5 was extremely familiar. Each participant completed both the spatial reasoning and count estimation tasks.

4 RESULTS

In this study, we explored the effects of sufficient and insufficient observation time and sufficient and insufficient decision time on participants' interactions with and perceptions of an AI agent, as well as their performance in a task. Appendix A provides the distribution of observation time and decision time in the two experimental tasks. Tables 2 and 3, respectively, illustrate the descriptive statistics and statistical test results of our behavioral, subjective, and performance metrics.

Table 2. Descriptive statistics of measures arranged by observation and decision time conditions. In the columns "Final Agreement," "Perceived Trust," "Perceived Usefulness," and "Error Improvement," the group mean value is provided followed by the group standard deviation in parentheses. In the column "Switch to AI," the total number of trials varied, as the metric considered the number of trials in which participants updated their response to agree with the AI suggestion given that their initial response disagreed with the AI suggestion. "SR" denotes the spatial reasoning task and "CE" denotes the count estimation task.

Factor	Level	Behavioral Metrics		Subjective Metrics		Performance Metric
		Final Agreement	Switch to AI (#of trials switched)	Perceived Trust	Perceived Usefulness	Error Improvement (%)
Obs. Time	**Insuf.**	SR: 90.35 (12.08) CE: 86.08 (17.49)	SR: 46 out of 150 CE: 57 out of 153	SR: 3.10 (1.11) CE: 3.43 (1.03)	SR: 3.30 (1.08) CE: 3.71 (0.81)	SR: 4.73 (9.37) CE: -0.13 (12.45)
	Suf.	SR: 90.55 (8.92) CE: 86.02 (18.52)	SR: 19 out of 148 CE: 51 out of 154	SR: 2.65 (1.10) CE: 3.38 (1.01)	SR: 2.85 (1.14) CE: 3.59 (0.90)	SR: 2.11 (6.72) CE: -3.71 (14.66)
Dec. Time	**Insuf.**	SR: 88.55 (11.29) CE: 83.43 (19.29)	SR: 21 out of 148 CE: 43 out of 157	SR: 2.84 (1.14) CE: 3.28 (1.04)	SR: 2.91 (1.15) CE: 3.54 (0.95)	SR: 2.15 (6.58) CE: -3.06 (15.03)
	Suf.	SR: 92.34 (9.53) CE: 88.66 (16.23)	SR: 44 out of 150 CE: 65 out of 150	SR: 2.91 (1.12) CE: 3.53 (0.98)	SR: 3.24 (1.09) CE: 3.76 (0.73)	SR: 4.69 (9.48) CE: -0.77 (12.15)

Table 3. Statistical test results from our behavioral, subjective, and performance metrics. Significant results are highlighted in light blue. "SR" denotes the spatial reasoning task and "CE" denotes the count estimation task.

Factor	Task	Behavioral Metrics		Subjective Metrics		Performance Metric
		Final Agreement	Switch to AI	Perceived Trust	Perceived Usefulness	Error Improvement
Obs. Time	SR	$F(1, 39) = 0.02$, $p = .886$	$z(319) = 3.43$, $p < .001$	$F(1, 37) = 8.46$, $p = .006$	$F(1, 39) = 10.22$, $p = .003$	$F(1, 38) = 6.83$, $p = .013$
	CE	$F(1, 39) = 0.00$, $p = .974$	$z(319) = -0.50$, $p = .615$	$F(1, 39) = 0.08$, $p = .777$	$F(1, 39) = 0.67$, $p = .418$	$F(1, 38) = 4.35$, $p = .044$
Dec. Time	SR	$F(1, 39) = 14.43$, $p < .001$	$z(319) = -2.07$, $p = .039$	$F(1, 37) = 0.26$, $p = .615$	$F(1, 39) = 5.03$, $p = .031$	$F(1, 38) = 11.09$, $p = .002$
	CE	$F(1, 39) = 5.39$, $p = .026$	$z(319) = -2.93$, $p = .003$	$F(1, 39) = 5.57$, $p = .023$	$F(1, 39) = 4.39$, $p = .043$	$F(1, 38) = 2.33$, $p = .135$
Obs. Time × Dec. Time	SR	$F(1, 39) = 10.71$, $p = .002$	$z(319) = -0.35$, $p = .730$	$F(1, 37) = 0.03$, $p = .858$	$F(1, 39) = 0.03$, $p = .860$	$F(1, 38) = 4.67$, $p = .037$
	CE	$F(1, 39) = 0.00$, $p = .974$	$z(319) = -1.95$, $p = .052$	$F(1, 39) = 3.68$, $p = .062$	$F(1, 39) = 1.07$, $p = .308$	$F(1, 38) = 0.82$, $p = .370$

For all the statistical tests reported below, $p < .05$ was considered a significant effect. For the results related to binary-outcome-dependent variables, we used stepwise multiple logistic regression where observation time and decision time were set as the fixed effects with an interaction term between observation and decision time. The logistic regressions included participants as a random effect (to account for repeated measures) and participants' age, gender, level of familiarity with AI, average performance on the calibration trials, and whether the "See Image Again" button was used in that specific trial as potential covariates in our model. Covariates were removed by stepwise backward elimination with log-likelihood ratio as the selection criterion [70] and $p < .15$ as the stop criterion [9, 75, 79]. Similarly, for the results related to continuous dependent variables, we used a two-way repeated measures analysis of covariance (ANCOVA) where observation time and decision time were set as the fixed effects with an interaction term between observation and decision time. The ANCOVA models included participants as a random effect and participants' age, gender, level of familiarity with AI, average performance on the calibration trials, and whether

or not the "See Image Again" button was used as potential covariates in our model. Covariates were removed by stepwise backward elimination with F-statistic as the selection criterion [61, 72] and $p < .15$ as the stop criterion [15]. All post-hoc pairwise comparisons were conducted using Tukey's HSD test.

We note that *task* was not considered as a fixed effect in our analyses because the manipulations of time constraints were performed within each task. Therefore, we report results and analyses for each task separately and do not intend to draw statistical conclusions about task differences.

4.1 Behavioral Metrics

4.1.1 Final Agreement. First, we studied the effects of observation and decision time on final agreement using a two-way repeated measures ANCOVA test. Fig. 5 visualizes our results for final agreement.

Spatial Reasoning Task. Five variables were removed from the model in the following order, step-by-step: age ($F(1, 33) = 0.19, p = .663$), gender ($F(2, 34) = 0.47, p = .627$), familiarity with AI ($F(1, 36) = 0.35, p = .559$), average performance on the calibration trials ($F(1, 37) = 0.31, p = .584$), and use of the "See Image Again" option ($F(1, 38) = 0.63, p = .431$)). In the final model, the main effect of observation time on final agreement was not significant, $F(1, 39) = 0.02, p = .886$; however, there existed a significant main effect of decision time on final agreement, $F(1, 39) = 14.43, p < .001$, indicating that participants tended to agree more with the AI suggestions under sufficient decision time ($M = 92.34, SD = 9.53$) than insufficient decision time ($M = 88.55, SD = 11.29$). Moreover, a significant interaction effect of observation time and decision time on final agreement was observed, $F(1, 39) = 10.71, p = .002$. A post-hoc pairwise comparison using Tukey's HSD test revealed that participants with insufficient observation time and insufficient decision time ($M = 86.72, SD = 13.39$) had notably lower final agreement with the AI than participants with insufficient observation time and sufficient decision time ($M = 93.98, SD = 9.35$), $p < .001$. See the full results of the test in Appendix Table 7.

Count Estimation Task. Five variables were removed from the model in the following order, step-by-step: average performance on the calibration trials ($F(1, 33) = 0.25, p = .624$), gender ($F(2, 34) = 0.51, p = .603$), use of the "See Image Again" option ($F(1, 36) = 0.29, p = .596$), age ($F(1, 37) = 0.37, p = .549$), and familiarity with AI ($F(1, 38) = 1.85, p = .181$). We did not observe a significant main effect of observation time on final agreement, $F(1, 39) = 0.00, p = .974$, although there existed a significant main effect of decision time on final agreement, $F(1, 39) = 5.39, p = .026$. Participants' final responses were more similar to the AI suggestions under sufficient decision time ($M = 88.66, SD = 16.23$) than insufficient decision time ($M = 83.43, SD = 19.29$). We observed no significant interaction effect of observation time and decision time on final agreement, $F(1, 39) = 0.00, p = .974$.

4.1.2 Switch to AI. We analyzed the results of a mixed effect logistic regression model on the effects of observation and decision time on whether or not participants switched to exactly agree with the AI suggestions if their initial responses did not exactly match the suggestions in the first place. We excluded trials in which participants' initial responses exactly matched the AI suggestions in this analysis, as none of the participants updated their initial responses if they exactly matched the AI suggestions (spatial reasoning: 22 out of 22 trials, count estimation: 13 out of 13 trials). Table 4 provides details of our final logistic regression model trained for each of the two tasks.

Spatial Reasoning Task. Two variables were removed from the model in the following order, step-by-step: familiarity with AI ($F(1, 319) = 0.13, p = .718$) and gender ($F(1, 319) = 0.80, p = .669$). Our final model indicated that four variables significantly influenced whether participants switched to agree with the AI suggestion: (1) observation time ($z(319) = 3.43, p < .001$); (2)

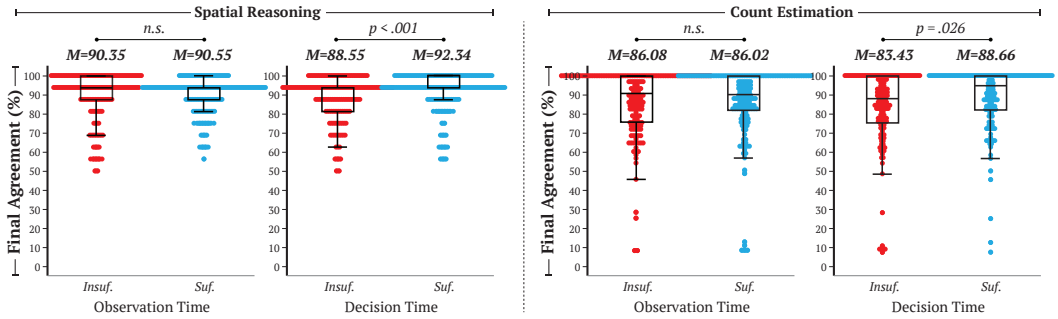

Fig. 5. Box and whisker plots of behavioral metrics showing participants' final agreement with the AI suggestions under insufficient vs. sufficient observation and decision time conditions for both the spatial reasoning (left) and count estimation (right) tasks.

Table 4. Stepwise multiple logistic regression on whether or not users switched to the AI suggestion given that their initial response disagreed with the AI suggestion. We included user ID as a random effect in each logistic regression model to account for repeated measures. We used backward elimination as the stepwise method, log-likelihood ratio as the selection criterion, and p < .05 as the stop criterion. Significant results are highlighted in light blue. The predictor "Average Calibration Performance" refers to the user's average performance on the calibration trials.

Predictor	**Spatial Reasoning**			**Count Estimation**		
	Odds Ratio	Confidence Interval (95%)	*p*	Odds Ratio	Confidence Interval (95%)	*p*
(Intercept)	0.21	0.11–0.43		0.72	0.44–1.18	
Observation Time	4.06	1.82–9.05	<.001	0.85	0.44–1.63	.615
Decision Time	0.33	0.11–0.95	.039	0.34	0.16–0.70	.003
Image Again	0.37	0.15–0.89	.026	not included in final model		
Age	1.49	1.01–2.18	.043	not included in final model		
Average Calibration Performance	1.35	0.93–1.96	.119	0.75	0.55–1.00	.052
Observation × Decision	0.80	0.22–2.92	.730	2.19	0.82–5.88	.118

decision time ($z(319) = -2.07, p = .039$); (3) whether the "See Image Again" option was used ($z(319) = -2.22, p = .026$); and (4) age ($z(319) = -2.02, p = .043$). Specifically, participants were more likely to switch to the AI suggestion under insufficient observation time (46 out of 150 trials) than sufficient observation time (19 out of 148 trials); moreover, they tended to be significantly more likely to switch to the AI suggestion under sufficient decision time (44 out of 150 trials) than insufficient decision time (21 out of 148 trials). Participants were also more likely to switch to the AI suggestion if they did not use the "See Image Again" option (54 out of 227 trials) than if they did use the option (11 out of 71 trials) or if they were older in age. No significant interaction effect of observation time and decision time was found, $z(319) = -0.35, p = .730$. Participants' average performance on the calibration trials was included in the final model, but did not have a significant main effect on participants' switch to the AI suggestion, $z(319) = 1.56, p = .119$.

Count Estimation Task. Four variables were removed from the model in the following order, step-by-step: gender ($F(1, 319) = 0.23, p = .892$), age ($F(1, 319) = 0.12, p = .732$), use of the "See Image Again" option ($F(1, 319) = 0.86, p = .354$), and familiarity with AI ($F(1, 319) = 1.16, p = .282$). After removing the insignificant covariates, our final model indicated that one variable significantly influenced whether participants switched to agree with the AI suggestion: decision time ($z(319) = -2.93, p = .003$). Participants were significantly more likely to switch to the AI suggestion under sufficient decision time (65 out of 150 trials) than insufficient decision time (43 out of 157 trials). No significant main effect of observation time ($z(319) = -0.50, p = .615$), interaction effect of observation time and decision time ($z(319) = 1.56, p = .118$), nor participants' average performance on the calibration trials ($z(319) = -1.95, p = .052$) were observed.

4.2 Subjective Metrics

4.2.1 Perceived Trust. We conducted a two-way repeated measures ANCOVA to analyze the effect of time pressure on participants' self-reported trust in the AI's suggestions. Fig. 6 presents the results for perceived trust ratings.

Spatial Reasoning Task. Three variables were removed from the model in the following order, step-by-step: average performance on the calibration trials ($F(1, 33) = 0.01, p = .945$), familiarity with AI ($F(1, 34) = 0.04, p = .952$), and gender ($F(2, 35) = 0.25, p = .784$). Our final model indicated that participants' trust ratings were significantly affected by observation time ($F(1, 37) = 8.46, p = .006$) even when controlling for participants' age and use of the "See Image Again" option. In particular, participants under insufficient observation time reported on average higher levels of trust ($M = 3.10, SD = 1.11$) than those under sufficient observation time ($M = 2.65, SD = 1.10$). Meanwhile, decision time did not significantly affect trust ratings, $F(1, 37) = 0.26, p = .615$; the interaction effect was not significant either, $F(1, 37) = 0.03, p = .858$ after controlling for participants' age and use of the "See Image Again" option. Age ($F(1, 37) = 3.00, p = .092$) and whether the "See Image Again" option was used ($F(1, 37) = 2.32, p = .136$) were considered but not significantly related to participants' perceived trust.

Count Estimation Task. Five variables were removed from the model in the following order, step-by-step: average performance on the calibration trials ($F(1, 33) = 0.01, p = .921$), familiarity with AI ($F(1, 34) = 0.03, p = .867$), use of the "See Image Again" option ($F(1, 35) = 0.06, p = .805$), gender ($F(2, 36) = 0.23, p = .793$), and age ($F(1, 38) = 0.08, p = .780$). Our final model indicated that trust ratings under insufficient observation time were not significantly different than those with sufficient observation time, $F(1, 39) = 0.08$, $p = .777$; however, trust ratings under insufficient decision time ($M = 3.28, SD = 1.04$) were on average significantly lower than those under sufficient decision time ($M = 3.53, SD = 0.98$), $F(1, 39) = 5.57$, $p = .023$. We did not observe a significant interaction effect, $F(1, 39) = 3.68$, $p = .062$.

4.2.2 Perceived AI Usefulness. We conducted a two-way repeated measures ANCOVA to analyze the effect of time pressure on participants' perceived usefulness of AI suggestions. Fig. 6 presents the results for perceived usefulness ratings.

Spatial Reasoning Task. Five variables were removed from the model in the following order, step-by-step: familiarity with AI ($F(1, 33) = 0.08, p = .786$), average performance on the calibration trials ($F(1, 34) = 0.12, p = .730$), gender ($F(2, 35) = 1.29, p = .289$), use of the "See Image Again" option ($F(1, 37) = 1.50, p = .228$), and age ($F(1, 38) = 1.51, p = .227$). Our final model indicated that participants with insufficient observation time ($M = 3.30, SD = 1.08$) rated AI suggestions to be significantly more useful than when they had sufficient observation time ($M = 2.85, SD = 1.14$), $F(1, 39) = 10.22, p = .003$. Conversely, participants with insufficient decision time ($M = 2.91, SD = 1.15$) rated AI suggestions as significantly less useful than when they had sufficient decision

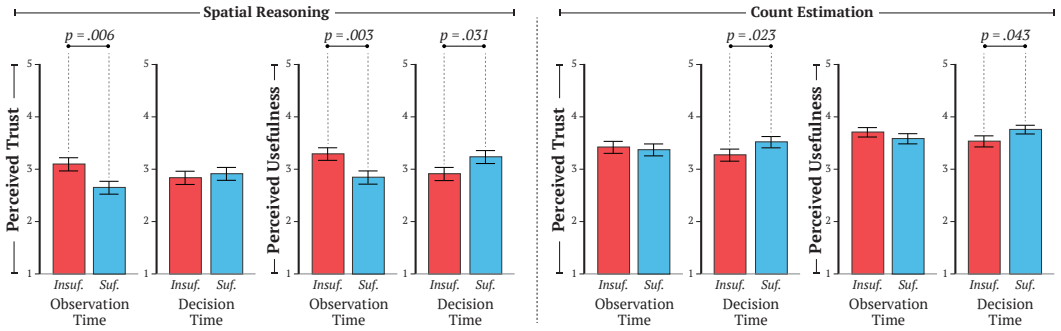

Fig. 6. Bar plots of subjective metrics showing participants' perceived trust in and perceived usefulness of AI suggestions under insufficient vs. sufficient observation and decision time conditions for both the spatial reasoning (left) and count estimation (right) tasks. The error bars shown in the plots represent the standard error and only significant results are emphasized.

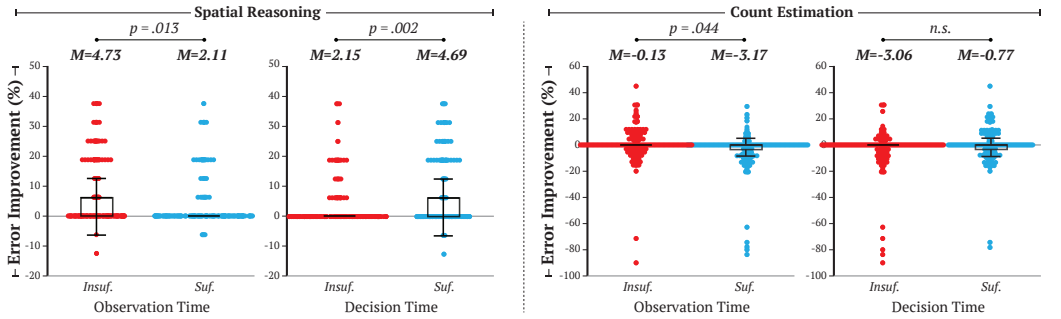

Fig. 7. Box and whisker plots demonstrating the percent error improvement under insufficient vs. sufficient observation and decision time conditions for both the spatial reasoning task (left) and the count estimation task (right). Higher positive values are more desirable.

time ($M = 3.24, SD = 1.09$), $F(1, 39) = 5.03, p = .031$. We did not observe an interaction effect, $F(1, 39) = 0.03, p = .860$.

Count Estimation Task. Five variables were removed from the model in the following order, step-by-step: gender ($F(2, 33) = 0.00, p = .998$), average performance on the calibration trials ($F(1, 35) = 0.05, p = .830$), use of the "See Image Again" option ($F(1, 36) = 0.26, p = .617$), familiarity with AI ($F(1, 37) = 0.24, p = .645$), and age ($F(1, 38) = 0.25, p = .620$). Our final model did not show significant main effects of observation time ($F(1, 39) = 0.67, p = .418$) on perceived AI usefulness. However, the main effect of decision time was significant, $F(1, 39) = 4.39, p = .043$; the average ratings of usefulness were on average lower under insufficient decision time ($M = 3.54, SD = 0.95$) than under sufficient decision time ($M = 3.76, SD = 0.73$). We did not observe an interaction effect, $F(1, 39) = 1.07, p = .308$.

4.3 Task Performance

We analyzed the change in participants' performance before and after seeing the AI suggestion under time pressure via the error improvement metric. Fig. 7 visualizes our results.

4.3.1 Error Improvement. We conducted a two-way repeated measures ANCOVA to explore the effect of time pressure on error improvement after users considered the AI's suggestions.

Spatial Reasoning Task. Four variables were removed from the model in the following order, step-by-step: familiarity with AI ($F(1, 33) = 0.12, p = .736$), gender ($F(2, 34) = 0.37, p = .695$), use of the "See Image Again" option ($F(1, 36) = 0.24, p = .630$), and age ($F(1, 37) = 0.87, p = .357$). There existed a significant main effect of observation time on participants' error improvement, $F(1, 38) = 6.83, p = .013$, indicating that error improvement under insufficient observation time ($M = 4.73\%, SD = 9.37\%$) was significantly higher than under sufficient observation time ($M = 2.11\%, SD = 6.72\%$). Likewise, there was a significant main effect of decision time on users' error improvement, $F(1, 38) = 11.09, p = .002$. When stratified by observation time, the average error improvement for insufficient decision time ($M = 2.15\%, SD = 6.58\%$) was lower than that of sufficient decision time ($M = 4.69\%, SD = 9.48\%$). Moreover, there was a significant interaction effect between observation and decision time on error improvement, $F(1, 38) = 4.67, p = .037$. Pairwise comparison using Tukey's HSD test (see Table 8 in the Appendix) found that under insufficient observation time, the difference in error improvement was significant between participants with insufficient decision time ($M = 2.50\%, SD = 6.72\%$) and those with sufficient decision time ($M = 6.95\%, SD = 11.03\%$), $p = .003$. Moreover, with sufficient decision time, participants' accuracy improved significantly more with insufficient observation time than with sufficient observation time ($M = 2.42\%, SD = 7.00\%$), $p = .008$. Participants with insufficient observation time and sufficient decision time improved significantly more than participants with sufficient observation time and insufficient decision time ($M = 1.80\%, SD = 6.46\%$), $p = .001$. Participants' average performance on the calibration trials ($F(1, 38) = 2.76, p = .105$) was included in the final model but not significantly related to error improvement.

Count Estimation Task. Four variables were removed from the model in the following order, step-by-step: gender ($F(2, 33) = 0.38, p = .687$), age ($F(1, 35) = 0.60, p = .443$), use of the "See Image Again" option ($F(1, 36) = 0.82, p = .371$), and familiarity with AI ($F(1, 37) = 1.42, p = .241$). We observed significant differences in error improvement between participants under insufficient ($M = -0.13\%, SD = 12.45\%$) and sufficient ($M = -3.71\%, SD = 14.66\%$) observation times, $F(1, 38) = 4.35, p = .044$. However, no main effect of decision time ($F(1, 38) = 2.33, p = .135$) nor interaction effect between the time variables was found on error improvement, $F(1, 38) = 0.82, p = .370$. Participants' average performance on the calibration trials ($F(1, 38) = 2.64, p = .113$) was included in the final model but not significantly related to error improvement.

4.3.2 Switch to AI and Error Improvement. We explored how whether or not participants updated their response to exactly match the AI suggestion among those whose initial response disagreed with the AI suggestion affected the accuracy of their decision outcome. Fig. 8 visualizes the results.

Spatial Reasoning Task. A Welch's t-test assuming unequal variances revealed that among participants whose initial response disagreed with the AI response, there was a significantly higher error improvement when the participant agreed with the AI suggestion ($M = 15.29, SD = 11.59$) than when the participant did not agree with the AI suggestion ($M = 0.39, SD = 2.51$), $t(65.54) = 10.30$, $p < .001$.

Count Estimation Task. A Welch's t-test assuming unequal variances revealed that there was no significant difference in error improvement when participants agreed with the AI suggestion ($M = -1.95, SD = 11.43$) than when they disagreed with the AI suggestion ($M = -1.90, SD = 14.74$), $t(267.59) = 0.04, p = .970$.

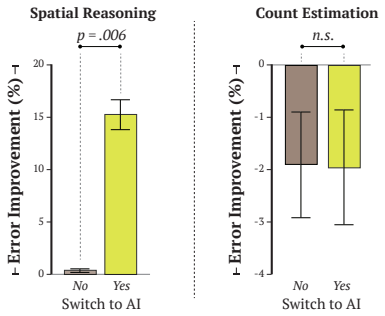

Fig. 8. Bar plots of participants' percent error improvement when they switched to the AI's suggestion vs. when they did not switch for both the spatial reasoning and count estimation tasks. The error bars shown in the plots represent the standard error.

5 DISCUSSION

5.1 Human-AI Agreement Under Time Pressure

In our study, we employed two behavioral metrics (*final agreement* and *switch to AI*) commonly used in research to retrospectively evaluate user reliance. We found that in the spatial reasoning task, observation time did not affect the level of agreement that a user's final response had with the AI suggestion; however, observation time did affect users' tendencies to adopt AI suggestions when their initial responses did not exactly match the suggestions in the first place (Table 4). These observations partially support H1 (if provided with insufficient observation time, users are more likely to agree with AI suggestions) for the spatial reasoning task. Conversely, in the count estimation task, observation time did not affect the level of agreement between the user's final response and the AI suggestion nor user tendency to switch to the AI suggestion (Table 3). These observations do not support H1 for the count estimation task.

Results of prior work suggest that insufficient observation time increases user compliance with automation when an AI suggestion is shown before the user engages with the task [54, 68]. In our study, even though the AI suggestion was shown after the user provided an initial response (a cognitive-forcing technique used to reduce over-reliance in users [8]), it was still unexpected that participants' reliance did not consistently increase with insufficient observation time. Participants under insufficient observation time had higher initial error in both tasks (spatial reasoning: M = 15.39%, SD = 14.21%; count estimation: M = 16.53%, SD = 13.10%) than participants under sufficient observation time (spatial reasoning: M = 9.53%, SD = 12.38%; count estimation: M = 9.50%, SD = 9.06%). Participants with insufficient observation time could have benefited from adopting the AI suggestion (AI error for spatial reasoning: M = 7.03%, SD = 2.07%; AI error for count estimation: M = 14.92%, SD = 3.22%) in both tasks; however, we did not observe increased reliance on the AI among participants with insufficient observation time. One possible explanation is that participants' confidence in their judgment may not have decreased with insufficient observation time [76].

Regarding decision time, our results did not support H2 (if provided with insufficient decision time, users are more likely to agree with AI suggestions); in fact, behavioral metrics suggested the opposite in both tasks: longer decision times were associated with an increased tendency to agree with AI suggestions (Fig. 5). This finding contradicts results from previous work [53], showing that allocating more time to a decision reduces anchoring bias in participants, thereby decreasing the odds of participants adopting the AI suggestion. We note that the AI suggestion was provided at different stages of the decision-making process in our study as opposed to prior work; we employed

a cognitive forcing function and showed the AI suggestion only after participants provided an initial response, whereas in prior research [53], participants were simultaneously presented with the AI suggestion and the task, causing the users to experience an initial anchoring effect on the AI suggestions before they could deliberate over the task at hand. Thus, in this case, longer (decision) time may be necessary for participants to make their own assessment first and then weigh that assessment against the AI's suggestion.

5.2 Perceptions of AI Suggestions

In this study, we employed two trust-related survey questions regarding perceived trust in AI and perceived AI usefulness. While our results show that user perceptions of an AI agent can be influenced by time pressure, our perceived trust findings did not fully agree with the results from either of the behavioral metrics, whereas the findings of perceived AI usefulness matched the results of the switch to AI metric. Specifically, in the spatial reasoning task, participants' perceived trust in and perceived usefulness of the AI were higher under insufficient observation time (Fig. 6, left); this aligns with the pattern observed in the switch to AI metric. However, participants' trust ratings were not affected by decision time, even though behavioral metrics and usefulness ratings indicated higher human-AI agreement and higher perceived AI usefulness under sufficient decision time. In the count estimation task, in agreement with findings from the behavioral metrics (Fig. 5, right), participants' perceived trust in and perceived usefulness of the AI were not affected by observation time (Fig. 6, right), and higher trust and usefulness ratings were observed under sufficient decision time than under insufficient decision time.

This result illustrates that there may be significant differences in what people consider to be trustworthy versus what they perceive as useful and therefore choose to adopt. Inspired by previous work that identified nuanced differences between trust and reliance in human-AI interaction [10] and found that trust guided reliance in human-automation interaction [37], we offer one possible explanation for why findings from perceived AI usefulness matched the behavioral metrics but not perceived trust in the spatial reasoning task: We conjecture that the AI usefulness ratings and the behavioral metrics captured user reliance on the AI, while perceived trust captured user trust in the AI. Trust and reliance, while linked, have a subtle distinction that causes them to be affected differently by time pressure.

5.3 AI Assistance in Reducing Errors

One of the main goals of integrating AI assistance into decision-making tasks is to improve human-AI team performance [4, 6, 36]. We used the error improvement metric to explore the effect of time pressure on task performance. In both tasks, error improvement was higher under insufficient observation time than sufficient observation time (Fig. 7). This outcome is expected, as the accuracy of participants' initial responses was lower under insufficient observation time, which left more room for improvement in their final responses.

On the other hand, the effect of decision time on error improvement was not consistent across the two tasks. In the spatial reasoning task, sufficient decision time led to greater error improvement than insufficient decision time, whereas in the count estimation task, decision time did not have a significant effect on participants' error improvement. From the behavioral metrics, we found that participants were more likely to follow the AI suggestion under sufficient decision time in both tasks; thus, the difference in the effect of decision time on error improvement may be explained by the variance in AI error between the two tasks. In the spatial reasoning task, the AI error was on average lower than participants' initial errors; conversely, in the count estimation task, the AI error was on average higher than participants' initial errors. Thus, in the spatial reasoning task, following the AI suggestion would likely help users improve their performance, whereas in the

count estimation task, following the AI would not be beneficial to users' task performance. To further explore this finding, we analyzed the relationship between the switch to AI metric and error improvement (Fig. 8); by comparing the error improvement of users who chose to change their response to match the AI suggestion and those who did not, we found that, in the spatial reasoning task, both groups' mean error improvement was positive. Additionally, those who changed their response to match the AI had a significantly higher error improvement than those who did not. This shows that, in the spatial reasoning task, trusting the AI suggestion was beneficial to the participant's task performance. However, in the count estimation task, both groups' mean error improvement was negative and not significantly different. A negative error improvement reflects that the participants' task performance would have been better if they had kept their initial response as their final answer.

Interestingly, we observe that a large proportion of participants had an error improvement of zero in both tasks (Fig. 7). In the spatial reasoning task, participants kept their initial and final response the same in 74% of the trials; in the count estimation task, participants anchored in their initial response in 54% of the trials. Despite the AI being more helpful in the spatial reasoning task, participants in this task demonstrated higher agency in their decisions than in the count estimation task. This result is likely due to the difference in task nature as described in Section 5.1. In the spatial reasoning task, participants were more logically involved, particularly in the final decision phase of the task, than they were in the count estimation task; thus, they might be more attached to their initial response as they had logic supporting their decision-making. In comparison, in the count estimation task, participants likely had more difficulty gauging the correctness of their own initial response, as well as that of the AI's suggestion—especially under time pressure. Thus, even though participants tended to anchor in their initial response in both tasks, they showed even more agency in the logic-based spatial reasoning task.

5.4 Designing for Human-AI Collaboration Under Time Pressure

Our findings have important implications for the design of human-AI collaboration under time pressure. Decisions in high-risk domains—such as the handling of icing encounters in aviation and interpreting CT scans in the emergency room —are often made under intense time pressure; AI assistants are increasingly being called upon to facilitate human decision-makers in these stressful situations. However, appropriate reliance and trust is fundamental to successful human-AI interaction. Our results show that observation time, decision time, and their interactions can significantly impact user reliance on and trust in an AI assistant. Thus, human-AI collaboration designs must adapt to changes in user reliance and trust patterns induced by time pressure. For instance, expert radiologists in a rush may have sufficient observation time to systematically read through a CT scan, but may have left themselves with insufficient decision time when moving on to the next reading. In this case, according to our spatial reasoning task result, the radiologist is less likely to rely on AI assistance; therefore, AI systems should incorporate ways to increase participants' trust and reliance without slowing them down in each individual case [27]—i.e., show evidence of model performance at the beginning of the interaction [27].

We additionally highlight the importance of considering task context when designing for collaboration under time pressure, as the effects of insufficient observation time and insufficient decision time and their interactions can vary depending on the task. Prior research has demonstrated that time allocation strategies can be employed to help reduce anchoring bias [53]; moreover, previous work has found that delaying the presentation of an AI suggestion (increasing observation time) gives users more time to reflect on the task and improves their ability to assess the accuracy of the AI's suggestion [47]. However, these works only considered time pressure from a single phase of decision-making, and for some tasks, there may not be unlimited task time. Our findings show the

potential for a strategic distribution of task time into initial observation time and final decision time to help users achieve more optimal decisions when strict time constraints are unavoidable. For instance, prior work showed that users who are very familiar with a task tend to rely less on AI assistance; in such a scenario, an AI suggestion should be shown earlier for an experienced user than it should for a user who is less familiar with the task, such that some of the observation time within the trial may be reallocated as decision time to help account for the former user's lower reliance on the AI.

5.5 Limitations and Future Work

This current work has a number of limitations that warrant future investigation.

First, our study had a relatively small sample size and our participants were recruited from a homogeneous population; as a result, all of our participants were young, well-educated, and somewhat familiar with AI technology. Accordingly, our ability to identify the effect of demographics-related covariates was limited; thus, while age was identified to have significantly influenced whether or not participants switched to exactly agree with the AI suggestions in our analysis, we cannot provide further analysis nor discussion of this supposed effect. Moreover, we note that additional research is required to determine whether other user factors may actually be of significance.

Second, the tasks employed in this work were low-stakes in nature. Although we sought to introduce and simulate time pressure into both tasks by experimentally manipulating observation and decision time, participants may not necessarily have felt the pressure typically associated in high-stakes or time-sensitive tasks in the real world. Our results, along with findings from previous works, indicate that user behavior in AI-assisted decision-making varies with task nature; further research is needed to systematically characterize the impact of task nature on human-AI decision-making.

Third, although a pilot study was conducted to gauge the difficulty level of and range of participant performance on the experimental tasks, participants in the main experiment performed unexpectedly well on the count estimation task. This caused the AI error to be on average higher than participants' initial error in the count estimation task, particularly when participants had sufficient observation time, which may have affected participants' interactions with the AI.

Fourth, we contextualized our study on the effect of time pressure on people's behavior when interacting with a simulated AI agent in a simulated environment. Having a simulated setup limits experimental fidelity given that participants could perform poorly on the tasks and that there were no consequences associated with poor performance.

Finally, time pressure is only one of the stressors in real-world decision-making. As we continue to develop AI systems to assist human decision-making, it is important to obtain a comprehensive, profound understanding of how different factors—such as the amount of information, complexity and consequences of a decision, uncertainty associated with the AI models in question, and human experience and domain knowledge—may shape decision quality and human trust in and reliance on AI in assisted decision-making.

6 CONCLUSION

In this paper, we present empirical findings from a user study investigating human decision-making behavior and consequent task performance under time pressure. Our results show that time pressure induced by limited initial observation and final decision time has different effects on user decision-making behavior and task performance; specifically, we found that the more decision time users had, the more likely they were to be influenced by an AI suggestion in their final response. Furthermore, task nature also shaped how time pressure affected participants; our findings suggest that users tended to have more agency when they were more logically involved in a task. This work provides a

nuanced understanding of how time pressure in different phases of a collaborative decision-making task may influence human decision-making behavior and joint human-AI team performance.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation award #1840088. We would like to acknowledge Chenghao Sun for contributing to the initial ideation of this work. We thank Jaimie Patterson for her feedback and assistance in this work.

REFERENCES

[1] Robert Alexander, Stephen Waite, Michael A Bruno, Elizabeth A Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. 2022. Mandating limits on workload, duty, and speed in radiology. *Radiology* 304, 2 (2022), 274–282.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[3] Carlos Arteta and Andrew Lempitsky, Victor andZisserman. 2016. Counting in the Wild. In *European Conference on Computer Vision*.

[4] 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.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[6] 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.

[7] RJM Bruls and RM Kwee. 2020. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into imaging* 11, 1 (2020), 1–7.

[8] 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.

[9] Zoran Bursac, C Heath Gauss, David Keith Williams, and David W Hosmer. 2008. Purposeful selection of variables in logistic regression. *Source code for biology and medicine* 3, 1 (2008), 1–8.

[10] 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.

[11] 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.

[12] 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.

[13] Plainsight Corp. 2021. Use Vision AI for Accurate Livestock Monitoring at Scale. https://plainsight.ai/use-vision-ai-for-accurate-livestock-monitoring-at-scale/

[14] Constance de Margerie-Mellon and Guillaume Chassagnon. 2022. Artificial intelligence: A critical review of applications for lung nodule and lung cancer. *Diagnostic and Interventional Imaging* (2022).

[15] Shelley Derksen and Harvey J Keselman. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *Brit. J. Math. Statist. Psych.* 45, 2 (1992), 265–282.

[16] Ruth Fernandez, Itiel E Dror, and Claire Smith. 2011. Spatial abilities of expert clinical anatomists: Comparison of abilities between novices, intermediates, and experts in anatomy. *Anatomical sciences education* 4, 1 (2011), 1–8.

[17] Howard E Gardner. 2011. *Frames of mind: The theory of multiple intelligences*. Hachette Uk.

[18] 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), 1–8.

[19] Jinchao Lin Gerald Matthews, Ryan W. Wohleber. 2019. Stress, Skilled Performance, and Expertise: Overload and Beyond. *The Oxford Handbook of Expertise* (May 2019). https://doi.org/10.1093/oxfordhb/9780198795872.013.22

[20] 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.
[21] Adam S Goodie and C.L. Crooks. 2004. Time-pressure effects on performance in a base-rate task. *The Journal of General Psychology* 131, 1 (2004), 18–28.
[22] Jill Graham, A.J. Ramirez, Stuart Field, and M.A. Richards. 2000. Job stress and satisfaction among clinical radiologists. *Clinical Radiology* 55, 3 (2000), 182–185.
[23] Oai Ha and Ning Fang. 2016. Spatial ability in learning engineering mechanics: Critical review. *Journal of Professional Issues in Engineering Education and Practice* 142, 2 (2016), 04015014.
[24] Zacks RT. Hasher L. 1979. Automatic and effortful processes in memory. *Journal of Experimental Psychology: General* 108, 3 (1979), 356–388. https://doi.org/108:356\T1\textendash388
[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] H Jachmann. 2002. Comparison of aerial counts with ground counts for large African herbivores. *Journal of Applied Ecology* 39, 5 (2002), 841–852.
[27] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[28] 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.
[29] Irving L. Janis and Leon Mann. 1977. Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment. *The ANNALS of the American Academy of Political and Social Science* 488 (1977).
[30] Daniel Kahneman. 1973. Attention and effort. *Prentice-Hall series in experimental psychology* (1973).
[31] Jeffry S Kellogg, Derek R Hopko, and Mark H Ashcraft. 1999. The effects of time pressure on arithmetic performance. *Journal of Anxiety disorders* 13, 6 (1999), 591–600.
[32] Michael Kirchler, David Andersson, Caroline Bonn, Magnus Johannesson, Erik Ø Sørensen, Matthias Stefan, Gustav Tinghög, and Daniel Västfjäll. 2017. The effect of fast and slow decisions on risk taking. *Journal of Risk and Uncertainty* 54, 1 (2017), 37–59.
[33] Martin G Kocher and Matthias Sutter. 2006. Time is money—Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization* 61, 3 (2006), 375–392.
[34] Lydia Kogler, Veronika I. Müller, Amy Chang, Simon B. Eickhoff, Peter T. Fox, Ruben C. Gur, and Birgit Derntl. 2015. Psychosocial versus physiological stress — Meta-analyses on deactivations and activations of the neural correlates of stress reactions. *NeuroImage* 119 (2015), 235–251. https://doi.org/10.1016/j.neuroimage.2015.06.059
[35] Igor Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* 23, 1 (2001), 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X
[36] 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.
[37] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[38] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[39] Chih-Wei Liu, Ai-Yun Hsieh, Shao-Kang Lo, and Yujong Hwang. 2017. What consumers see when time is running out: Consumers' browsing behaviors on online shopping websites when under time pressure. *Computers in Human Behavior* 70 (2017), 391–397.
[40] Sonia J. Lupien and S. Gaudreau. 1997. Stress-induced declarative memory impairment in healthy elderly subjects: relationship to cortisol reactivity. *The Journal of clinical endocrinology and metabolism* 82, 7 (1997), 2070–2075. https://doi.org/10.1210/jcem.82.7.4075
[41] Sonia J Lupien and Martin Lepage. 2001. Stress, memory, and the hippocampus: can't live with it, can't live without it. *Behavioural Brain Research* 127, 1 (2001), 137–158. https://doi.org/10.1016/S0166-4328(01)00361-8
[42] Mather M. and Lighthall N. R. 2012. Both Risk and Reward are Processed Differently in Decisions Made Under Stress. *Current directions in psychological science* 21, 2 (2012), 36–41. https://doi.org/10.1177/0963721411429452
[43] Suet Mui Ma. 2016. Working memory in Spanish-English and Chinese-English bilinguals. *Psychology and Behavioral Sciences* 5, 4 (2016), 104–112.

[44] Francesca Gioia Maria De Paola. 2016. Who performs better under time pressure? Results from a field experiment. *Journal of Economic Psychology* 53 (2016), 37–53. Issue C.

[45] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[46] Lisa Ordóñez and Lehman Benson. 1997. Decisions under Time Pressure: How Time Constraint Affects Risky Decision Making. *Organizational Behavior and Human Decision Processes* 71, 2 (1997), 121–140. https://doi.org/10.1006/obhd.1997.2717

[47] 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.

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

[49] Marcus Paul, RK Lech, Juliane Scheil, AM Dierolf, Boris Suchan, and OT Wolf. 2016. Acute stress influences the discrimination of complex scenes and complex faces in young healthy men. *Psychoneuroendocrinology* 66 (2016), 125–129.

[50] Nathalie Pettorelli, Steeve D Côté, André Gingras, François Potvin, and Jean Huot. 2007. Aerial surveys vs hunting statistics to monitor deer density: the example of Anticosti Island, Quebec, Canada. *Wildlife Biology* 13, 3 (2007), 321–327.

[51] Gloria Phillips-Wren and Monica Adya. 2020. Decision making under stress: the role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems* 0, 0 (2020), 1–13. https://doi.org/10.1080/12460125.2020.1768680 arXiv:https://doi.org/10.1080/12460125.2020.1768680

[52] Jens C. Pruessner, Katarina Dedovic, et al. 2008. Deactivation of the limbic system during acute psychosocial stress: Evidence from positron emission tomography and functional magnetic resonance imaging studies. *Biological Psychiatry* 63, 2 (2008), 234–240. https://doi.org/10.1016/j.biopsych.2007.04.041

[53] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22.

[54] Stephen Rice, David Keller, Gayle Hunt, and David Trafimow. 2009. Automation Dependency Under Time Pressure. In *2009 International Symposium on Aviation Psychology*. 611.

[55] Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications* 11, 3923 (2020). https://doi.org/10.1038/s41467-020-17419-7

[56] Tobias Rieger and Dietrich Manzey. 2022. Understanding the impact of time pressure and automation support in a visual search task. *Human Factors* (2022), 00187208221111236.

[57] Ericka Rovira, Kathleen McGarry, and Raja Parasuraman. 2007. Effects of imperfect automation on decision making in a simulated command and control task. *Human factors* 49, 1 (2007), 76–87.

[58] Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi, and Haidi Ibrahim. 2015. Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence* 41 (2015), 103–114.

[59] Nadine B Sarter and Beth Schroeder. 2001. Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human factors* 43, 4 (2001), 573–583.

[60] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[61] Robert A Scott, Mark ES Bailey, Colin N Moran, Richard H Wilson, Noriyuki Fuku, Masashi Tanaka, Athanasios Tsiokanos, Athanasios Z Jamurtas, Evangelia Grammatikaki, George Moschonis, et al. 2010. FTO genotype and adiposity in children: physical activity levels influence the effect of the risk genotype in adolescent males. *European Journal of Human Genetics* 18, 12 (2010), 1339–1343.

[62] Sonia Jawaid Shaikh and Ignacio F Cruz. 2022. AI in human teams: effects on technology use, members' interactions, and creative performance under time scarcity. *AI & SOCIETY* (2022), 1–14.

[63] Martin J. Sliwinski, Joshua M. Smyth, Scott M. Hofer, and Robert S. Stawski. 2006. Intraindividual coupling of daily stress and cognition. *Psychology and aging* 21, 3 (2006), 545–557. https://doi.org/10.1037/0882-7974.21.3.545

[64] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[65] Robert S Stawski et al. 2006. Stress-related cognitive interference predicts cognitive function in old age. *Psychology and aging* 21, 3 (2006), 535–544. https://doi.org/10.1037/0882-7974.21.3.535

[66] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.

[67] Ola Svenson and Anne Edland. 1987. Change of Preferences Under Time Pressure. *Scandinavian Journal of Psychology* 28 (06 1987), 322 – 330. https://doi.org/10.1111/j.1467-9450.1987.tb00769.x
[68] Casey Tunstall, Stephen Rice, Rian Mehta, Victoria Dunbar, and Korhan Oyman. 2014. Time pressure has limited benefits for human-automation performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 1043–1046.
[69] Jessica Phuong Uy and Adriana Galván. 2017. Sleep duration moderates the association between insula activation and risky decisions under stress in adolescents and adults. *Neuropsychologia* 95 (2017), 119–129. https://doi.org/10.1016/j.neuropsychologia.2016.12.018
[70] 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.
[71] Xinru Wang and Ming Yin. 2021. *Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650
[72] Luc A Wauters, Simon A de Crombrugghe, Nadia Nour, and Erik Matthysen. 1995. Do female roe deer in good condition produce more sons than daughters. *Behavioral ecology and sociobiology* 37, 3 (1995), 189–193.
[73] Peter Wright. 1974. The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology* 59, 5 (1974), 555–561. https://doi.org/10.1037/h0037186
[74] Zhuoran Xiong, Xiao-Yang Liu, Shan Zhong, Hongyang Yang, and Anwar Walid. 2018. Practical deep reinforcement learning approach for stock trading. *arXiv preprint arXiv:1811.07522* (2018).
[75] Lu Xu and Wen-Jun Zhang. 2001. Comparison of different methods for variable selection. *Analytica Chimica Acta* 446, 1-2 (2001), 475–481.
[76] Huiqin Yang, Carl Thompson, and Martin Bland. 2012. The effect of clinical experience, judgment task difficulty and time pressure on nurses' confidence calibration in a high fidelity clinical simulation. *BMC medical informatics and decision making* 12, 1 (2012), 1–9.
[77] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[78] 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.
[79] Dietmar Zellner, Frieder Keller, and Günter E Zellner. 2004. Variable selection in logistic regression models. *Communications in Statistics-Simulation and Computation* 33, 3 (2004), 787–805.
[80] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI Conference on Human Factors in Computing Systems*. 1–28.
[81] 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 DISTRIBUTION OF OBSERVATION TIME AND DECISION TIME

Table 5. Distribution of observation time and decision time in seconds as a result of time pressure manipulation for the two tasks.

Time Condition	**Insufficient**	**Sufficient**
Observation Time (Spatial Reasoning)	$M = 5.10s, SD = 2.91s$	$M = 14.73s, SD = 8.57s$
Decision Time (Spatial Reasoning)	$M = 3.18s, SD = 2.50s$	$M = 9.03s, SD = 7.53s$
Observation Time (Count Estimation)	$M = 6.72s, SD = 2.71s$	$M = 19.65s, SD = 8.03s$
Decision Time (Count Estimation)	$M = 2.85s, SD = 2.41s$	$M = 7.99s, SD = 7.14s$

B COMPARISON OF INITIAL ERROR AMONG PARTICIPANTS WHO SWITCHED VERSUS NOT SWITCH

Table 6. Distribution of initial error of participants who switched to the AI suggestion and participants who did not switch to the AI suggestion for the spatial reasoning and count estimation tasks.

Switch to AI	Spatial Reasoning	Count Estimation
Yes	$M = 22.50\%, SD = 11.79\%$	$M = 12.83\%, SD = 12.15\%$
No	$M = 9.90\%, SD = 12.88\%$	$M = 12.82\%, SD = 10.72\%$

C ADDITIONAL PAIRWISE COMPARISON RESULTS

Table 7. Results from pairwise comparisons using Tukey's HSD test for interaction effect of observation time and decision time on final agreement for the spatial reasoning task.

Observation Time	Decision Time	-Observation Time	-Decision Time	*p*-value	Significant
insufficient	insufficient	sufficient	insufficient	.118	no
insufficient	insufficient	insufficient	sufficient	<.001	yes
insufficient	insufficient	sufficient	sufficient	.076	no
sufficient	insufficient	insufficient	sufficient	.131	no
sufficient	insufficient	sufficient	sufficient	.998	no
insufficient	sufficient	sufficient	sufficient	.193	no

Table 8. Results from pairwise comparisons using Tukey's HSD test for interaction effect of observation time and decision time on error improvement for the spatial reasoning task.

Observation Time	Decision Time	-Observation Time	-Decision Time	*p*-value	Significant
insufficient	insufficient	insufficient	sufficient	.003	yes
insufficient	insufficient	sufficient	insufficient	.952	no
insufficient	insufficient	sufficient	sufficient	1.000	no
insufficient	sufficient	sufficient	insufficient	.001	yes
insufficient	sufficient	sufficient	sufficient	.008	yes
sufficient	insufficient	sufficient	sufficient	.950	no

Received July 2022; revised January 2023; accepted March 2023