Explanations <u>Can</u> Reduce Overreliance on AI Systems During Decision-Making

HELENA VASCONCELOS, Stanford University, USA

MATTHEW JÖRKE, Stanford University, USA

MADELEINE GRUNDE-MCLAUGHLIN, University of Washington, USA

TOBIAS GERSTENBERG, Stanford University, USA

MICHAEL S. BERNSTEIN, Stanford University, USA

RANJAY KRISHNA, University of Washington, USA

Prior work has identified a resilient phenomenon that threatens the performance of human-AI decision-making teams: *overreliance*, when people agree with an AI, even when it is incorrect. Surprisingly, overreliance does not reduce when the AI produces explanations for its predictions, compared to only providing predictions. Some have argued that overreliance results from cognitive biases or uncalibrated trust, attributing overreliance to an inevitability of human cognition. By contrast, our paper argues that people strategically choose whether or not to engage with an AI explanation, demonstrating empirically that there *are* scenarios where AI explanations reduce overreliance. To achieve this, we formalize this strategic choice in a cost-benefit framework, where the costs and benefits of engaging with the task are weighed against the costs and benefits of relying on the AI. We manipulate the costs and benefits in a maze task, where participants collaborate with a simulated AI to find the exit of a maze. Through 5 studies ($N = 731$), we find that costs such as task difficulty (Study 1), explanation difficulty (Study 2, 3), and benefits such as monetary compensation (Study 4) affect overreliance. Finally, Study 5 adapts the Cognitive Effort Discounting paradigm to quantify the utility of different explanations, providing further support for our framework. Our results suggest that some of the null effects found in literature could be due in part to the explanation not sufficiently reducing the costs of verifying the AI's prediction.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **HCI theory, concepts and models**; *Empirical studies in collaborative and social computing*; **Empirical studies in collaborative and social computing**; *Empirical studies in HCI*.

Additional Key Words and Phrases: human-AI collaboration, explainable AI, decision-making, cost-benefit analysis

1 INTRODUCTION

Human-AI decision-making has fallen short of its goal of achieving complementarity: Human-AI teams ought to perform better, and therefore make better decisions, than either humans or AIs

Authors' addresses: Helena Vasconcelos, helenav@cs.stanford.edu, Stanford University, USA; Matthew Jörke, joerke@cs.stanford.edu, Stanford University, USA; Madeleine Grunde-McLaughlin, mgrunde@cs.washington.edu, University of Washington, USA; Tobias Gerstenberg, gerstenberg@stanford.edu, Stanford University, USA; Michael S. Bernstein, msb@cs.stanford.edu, Stanford University, USA; Ranjay Krishna, ranjay@cs.washington.edu, University of Washington, USA.
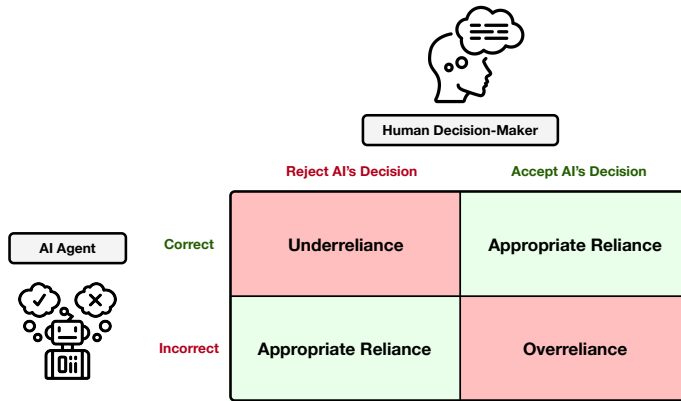
Fig. 1. A simplified diagram outlining the different ways people rely on AI systems. Provided with a prediction from an AI system, a human decision-maker has the choice to either accept or reject the AI's prediction. Appropriate reliance occurs when the human accepts a correct AI prediction or corrects an incorrect AI prediction. Underreliance occurs when the human fails to accept an correct AI prediction. Overreliance occurs when the human fails to correct an incorrect AI prediction.

alone [4]. Unfortunately, the literature has repeatedly established that people are influenced by the AI and often accept incorrect decisions without verifying whether the AI was correct. This phenomenon is known as *overreliance* [9]. Of the possible error types in human-AI decision-making (Figure 1), overreliance is the most frequent outcome in empirical AI studies [4, 8, 9, 53, 97]. Overreliance can be seen as foregoing agency and accountability to the AI when making final decisions [10, 38, 55], and is of particular concern in high-stakes domains, running the risk of reinforcing machine bias [1, 30] under the guise of human agency.

Explainable AI (XAI) has been proposed as a mitigation, but experimental evidence has so far failed to find support that explanations reduce overreliance. The intuition runs that, if people see an explanation that is incorrect, they will more carefully scrutinize the prediction and be less likely to overrely [6, 10, 11]. However, empirically, explanations have failed to improve levels of overreliance [4, 9]. Potential cognitive mechanisms that underlie these results suggest that the mere presence of an explanation increases trust [96, 97], and that explanations anchor humans to the prediction [9, 91]. Scaffolds such as cognitive forcing functions, which force people to engage more thoughtfully with the AI's explanations, can reduce overreliance [9]. Taken together, existing literature puts forth a resilient finding: that explanations exacerbate or do not alter overreliance unless intervention methods in the form of forcing functions compel people to engage with the task.

In this paper, we produce evidence that explanations can in fact reduce overreliance without the need for forcing functions. To demonstrate this, we begin by observing that real-world tasks and explanations can vary greatly in their cognitive costs [9]: tasks can be cognitively costly (difficult) to complete, or cognitively cheap (easy); checking explanations can likewise be cognitively costly or cheap. We argue that the null effects found in the literature—that explanations do not reduce overreliance—occur because prior work has focused on a narrow slice of this cost space where these two costs are nearly equal and both nontrivially costly. In this situation, where both the task and the explanation are costly, we might expect people to just rely on the AI's prediction without carefully checking it. For example, answering a reading comprehension question requires reading a passage and incurs a similar cognitive effort as reading the AI generated natural language

explanation [4]; similarly, guessing the calories in a photo of food attracts a similar cost as guessing and summing up the calories of the individual ingredients produced as an AI explanation [9]. If it takes roughly as much effort to verify an AI's explanation as it does to complete the task manually, then we should expect people to sometimes forego verifying the explanation (either by overrelying or by completing the task alone). But, if we were to make the task much more challenging and thus more costly, while the explanation remains equally costly as it was before, the explanation would suddenly become strongly preferable and overreliance would reduce. This is because, relatively speaking, verifying the AI's explanation has become very easy to do, compared to doing the task alone (very costly) or relying on the AI (undesirable because of possible errors). Likewise, if the explanation were to become substantially easier to check than it was before, it would become preferable and overreliance would reduce.

We introduce a theoretical framework to reason about the conditions under which explanations will reduce overreliance. We then use this framework to predict new conditions under which explanations will in fact reduce overreliance. Our framework, in detail, is a cost-benefit analysis wherein people strategically weight the cognitive effort required to complete the task alone or verify the AI against the effortless strategy of overreliance. It suggests that when tasks require more effort and the explanations sufficiently reduce the cost of verifying with the AI, or when the explanations are easier to understand, explanations will reduce overreliance.

We test this framework through a series of five experiments ($N = 731$) with a simulated AI in a maze-solving task. Study 1 ($N = 340$) manipulates task difficulty. It replicates prior results that receiving an AI's prediction versus receiving an AI's prediction with an explanation have equal levels of overreliance—but only for an easy task. As the task becomes more effortful, and engaging with the explanation becomes less relative cognitive effort, XAI overreliance becomes lesser than AI overreliance. Study 2 ($N = 340$) and Study 3 ($N = 286$) manipulate the cognitive cost of understanding and engaging with the AI's explanations, and find, again, that when the cognitive effort is lower, the reduction in overreliance is larger. Study 4 ($N = 114$) tests whether the cognitive effort (costs) are balanced against the incentives (benefits) accrued for expending those costs—the benefit piece of our cost-benefit framework—by manipulating the financial incentive for a correct answer. We verify that, as the monetary reward for accuracy increases, overreliance decreases. Study 5 ($N = 76$) adapts the Cognitive Effort Discounting paradigm [94] to synthesize the costs (Studies 1-3) with benefits (Study 4) through the lens of economic utility. This final study measures the financial value (utility) attached to explanations and validates our hypotheses, finding that people attach more utility to explanations that are easier to understand and attach more utility to explanations in harder tasks.

Our results indicate that overreliance is not an inevitability of cognition but a strategic decision where people are responsive to the costs and benefits. Beyond the immediate results of our study, our work suggests that designers of collaborative human-AI systems should carefully consider how explanations reduce the cognitive effort of verifying an AI in comparison to a low-effort strategy of overreliance.

2 BACKGROUND

Empirical HCI studies of human-AI decision-making find little evidence that human-AI teams actually achieve complementary performance in which teams perform better than a human or an AI alone, even when explanations are provided [4]. Since AIs outperform humans in many task domains, a strategy which blindly relies on model predictions is likely to achieve similar performance to one which utilizes explanations to evaluate model outputs. Human-AI performance generally only exceeds human-only performance [8, 26, 30, 55, 65, 78, 97]; stark performance differences between humans and AIs make it such that one would achieve the highest performance

by removing humans from the loop entirely. Therefore, among studies that report that performance improves when an AI gives explanations [8, 36, 53, 55], it is unclear whether explanations truly improve understanding of model capabilities, or if they instead serve as a signal for blind trust in a AI that is more accurate than the human baseline [4, 9, 24].

In light of this conundrum, recent studies evaluate human-AI decision-making in domains with roughly equal human and AI performance. Unfortunately, they do not find that explanations improve performance above a prediction-only baseline [4, 15], offering evidence for the latter theory—that explanations exacerbate overreliance. In such complementary domains, explanations (even those generated by humans [4]) appear to increase reliance on both correct and incorrect AI predictions. Among inputs incorrectly predicted by the AI, humans achieve worse performance than if they had completed the task without any AI partner. In light of growing concerns surrounding machine bias [1, 30], such *overreliance* on incorrect model predictions is particularly concerning and is the primary focus of our work. We investigate the conditions under which people engage in effortful thinking to verify AI predictions, reducing overreliance. We are guided by the following research question:

> RESEARCH QUESTION: *Under what conditions do people engage with explanations to verify AI predictions, reducing overreliance?*

2.1 Cognitive biases in decision-making

Research to this point focuses on elucidating the cognitive mechanisms and biases that cause overreliance. For decades, cognitive psychology literature on mental effort and advice utilization has attempted to understand the cognitive processes involved in decision-making. Pioneered by Tversky & Kahneman [40, 41], *dual process theory* accounts for reasoning by means of two distinct cognitive subsystems: one that is unconscious, automatic, fast, and intuitive (System 1) and one that is deliberate, effortful, slow, and analytical (System 2). System 1 processes employ heuristics to arrive at fast and efficient conclusions, without which all thinking would be slow and arduous. While System 1 thinking can greatly speed up decision-making, its reliance on cognitive shortcuts is also the source of many known cognitive biases [41].

Systems 1 and 2 can interact with and override each other throughout decision-making processes, leading to judgements that are often hybrids of intuitive and analytical thinking [17]. As a result, cognitive biases are pervasive in most forms of decision-making, even in high-stakes tasks with domain experts. For example, the role of cognitive biases is well-documented in medicine [16, 17, 56], with survey studies indicating a diagnostic error rate of around 10-15%, with 75% of these errors resulting from cognitive factors. Other work has investigated the role of cognitive biases in foreign policy [58], the initiation of war [90], judicial sentencing decisions [31], national intelligence [66], or catastrophe risk assessment [89]. These cognitive biases manifest themselves in the way people process information during decision-making.

Based on dual process theory, a recent study investigates the use of cognitive forcing functions—interventions designed to encourage effortful, analytical thinking (e.g., forcing the user to wait for a certain period of time before they receive an explanation from the AI) —as a means for reducing overreliance [9]. The authors find that forcing functions successfully reduce overreliance on incorrect model predictions. However, they also reduce reliance on correct model predictions, yielding no significant differences when humans are not aided by an AI. The authors also identified an interesting trade-off: participants performed best in the conditions they preferred and trusted the least. We provide another lens through which to view this line of work: one that uses a cost-benefit framework where cognitive forcing functions can be viewed as manipulations for cost, i.e. by necessitating cognitive effort and analytical thinking.
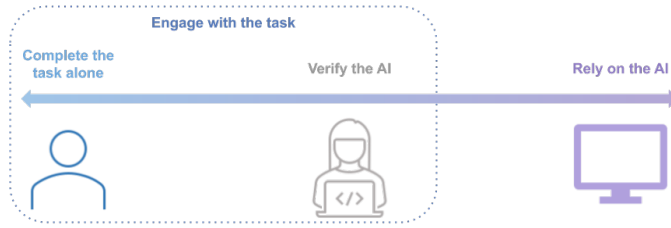
Fig. 2. We define two prototype strategies: (1) engage with the task (in the form of completing the task alone or verifying the AI's prediction and/or explanation), and (2) rely on the AI. Of course, decision-makers can lie on this continuum, especially as they interact with AI over long periods of time, but these strategies help ground our framework.

2.2 Decision-making in behavioral economics

Dual process theory makes clear that System 2 can override default biases and heuristics [25]. However, human decision-makers are "satisficers", devoting only as much effort as is minimally satisfactory rather than aiming for optimal decision-making outcomes [83]. To frame this aversion to engaging in effortful thinking in economic terms: people weigh the potential benefits of cognitive effort against its perceived costs [46, 70]. This economic framing of cognitive effort allocation is supported by evidence from behavioral experiments [45, 48, 94], neuroscientific studies [7, 68, 81], and computer simulations [39, 75]. In this economic framing, effort-based decision-making is modeled as a selection among different cognitive strategies, each associated with its own perceived benefits (in terms of task performance) and costs (in terms of cognitive effort). Decision-makers navigate this effort-accuracy trade-off [39] by adopting the strategy which maximizes subjective utility as a function of costs and benefits [46]. In the context of computer-based decision aids, prior work has found that decision aids which reduced the cognitive effort associated with a particular strategy induced behavior associated with that strategy [84], implying that decision-aid designers can influence cognitive strategy selection by manipulating costs. Similar studies [47, 80] also provide support for the influence of benefits: when higher rewards are at stake, people engage in more effortful strategies. Our study investigates the effect of XAI methods by manipulating the costs associated with a high-effort strategy which verifies an AI's predictions versus the low-effort strategy of overreliance.

3 FRAMEWORK

Theories proposed in prior work—from anchoring biases [9, 91] to humans' innate adversity to cognitive effort [9, 46, 70, 97]—suggest that overreliance, or the frequency with which people agree with an AI when it is incorrect [9], is the default and inevitable state of human performance when presented with an explanation. However, clearly sometimes people do engage with automated systems AIs in calibrated ways. For example, people rarely choose a suggested auto-reply to an important email to their boss [32], and will override AI suggestions if a virtual assistant misunderstands a request [63]. Furthermore, these frameworks do not account for research in cognitive psychology in which explanations improve performance in human-human contexts [5, 14, 49, 62, 98].

To account for this contradiction, we propose that overreliance is not an inevitability of cognition, but instead the result of a strategic choice. For example, a person may be less likely to overrely when the *benefit* of getting the right answer is high (e.g., not sending an unprofessional email to a boss). Alternatively, a person might be less likely to overrely when the *cost* associated with extracting the right answer is low (e.g., deciding whether a tweet has positive sentiment).

Prior work in behavioral economics uses cost-benefit frameworks [46, 70] to explain which decision making strategies people choose. In this framework, people subconsciously weigh the costs (e.g., cognitive effort, time) and benefits (e.g., task performance, rewards) of performing different cognitive strategies. The person forms an opinion on the value of each strategy, called the *subjective utility* of that strategy, by judging its ability to minimize costs and maximize benefits. They enact the strategy that they value to have the highest subjective utility. Utility if often defined to a function of perceived costs and benefits.

We propose that this cost-benefit framework from behavioral economics also applies to the way people collaborate with AI predictions and explanations. In the context of human-AI decision making, this framework suggests that people compare the potential benefits of engaging with the task (i.e. professional accomplishment, monetary reward) weighed against its inherent costs (i.e. cognitive effort, time)[1]. People are likely to overrely if engaging with the task is not the optimal strategy. When people choose to engage with the task, they are adopting one of two strategies: either they do the task alone, ignoring any AI support, or they verify the AI's prediction, sometimes using AI generated explanations if they are present.

Perceptions of costs in the form of cognitive effort, time, etc., as well as benefits in the form of monetary reward, stakes, etc. are subject to individual differences [48, 87]. However, we predict that the following principle holds: *if an element of a task (e.g., an explanation) reduces the cost (via reducing the difficulty of the task or other means) and/or increases the benefits (via increasing monetary reward or other means) of a particular decision-making strategy relative to other strategies, a decision-maker will be more inclined to employ that strategy.*[2] Therefore, when aggregating across many decision-makers, we expect that conditions which support a particular strategy will induce behavior consistent with that strategy.

This framework offers a possible explanation for why XAIs have not decreased overreliance compared to traditional AIs in prior work: we have been focused on a narrow slice of the cost-benefit space. In particular, receiving the AI's explanation does not reduce the costs of engaging with the task. This is to say that, in these cases, the cost of engaging with the task using the AI's explanation is roughly equal to engaging via using the prediction or by completing the task alone. In this case, we do not expect to see differences between overreliance levels with explanations, since there is no difference in cost. Therefore, the strategy to engage with the task does not have a change in utility.

However, other parts of the cost-benefit space would predict different outcomes. Consider the following example: if the task were more difficult and if the AI's explanation reduced the cost of engaging with the task, the strategy to engage with the task would seem relatively much more attractive. This is because relying remains undesirable due to the prospect of the AI making mistakes (lowering the benefits of this strategy) and engaging via completing the task alone remains very costly. In this case, engaging with the task (likely by engaging with the explanations) is desirable and we can moderately assume that overreliance will decrease.

3.1 Predictions of our cost-benefit framework

Applying the cost-benefit framework to human-AI decision making, we see that the cost (in this case, cognitive effort) required to engage with the task depends on the difficulty of the underlying task and the difficulty of using the explanation. Given these two factors, we predict that:

> PREDICTION 1: As tasks increase in difficulty, there will be greater reductions in overreliance with explanations, compared to only getting predictions.

---

[1] We note that there are effects of other factors, like trust and intrinsic rewards, but we do not explore these in this paper. See: Section 10.4.

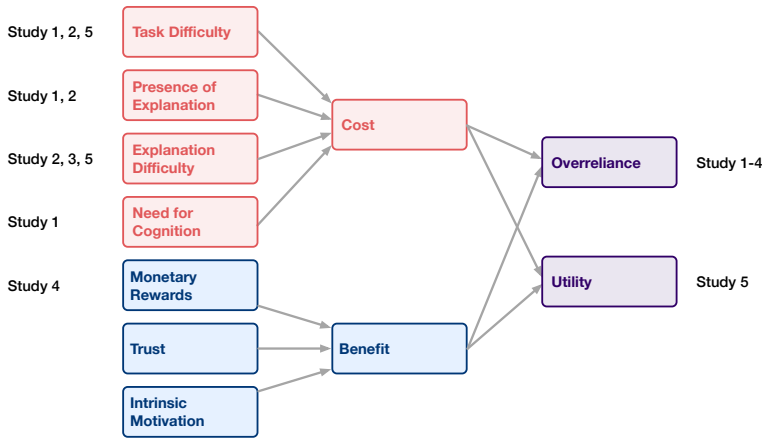[2]This principle is adapted from [84, p. 50-51].

Fig. 3. Outline of some of the factors that affect the costs and benefits of engaging with the explanations, as well as our associated studies that manipulate or measure the factors.

> *This is because explanations are relatively more "useful," or have more utility, in harder tasks, since they greatly reduce the cognitive effort required to engage with the task.*
>
> PREDICTION 2: As the effort to understand the explanation decreases, overreliance decreases.
> *Intuitively, explanations that are easier to parse are thus easier to discern when they are correct or incorrect.*

Applying the cost-benefit framework, we can also see that benefit required to engage with the task can depend on factors such as monetary reward (i.e. in a crowdsourced task) or stakes (i.e. in real-world medical decision-making). As such, we predict that:

> PREDICTION 3: As the monetary benefit of correctly completing the task increases, overreliance decreases.
> *People are more inclined to expend effort into a task if there are higher stakes attached to it.*

Finally, our theory predicts that an individual's *subjective utility* of the AI is associated with the cognitive effort reduction given by the explanation. In the cost-benefit framework, people estimate the value of a strategy, called its subjective utility, by assessing the degree to which it maximizes benefits and minimizes costs. To understand the cost-benefit framework's relevance in Human-AI interaction, we need to address the connection between cost, benefit, and subjective utility in the Human-AI interaction domain. Therefore, we predict that:

> PREDICTION 4: As tasks increase in difficulty, people will attach higher subjective utility to an AI's explanation.
> *This is because, in harder tasks, the explanation greatly reduces the cost (therefore increasing the utility) of engaging with the task.*
>
> PREDICTION 5: As the effort to understand the explanation decreases, people will attach higher subjective utility to an AI.
> *This is because explanations that are easier to parse reduces the cost of (therefore increasing the utility) engaging with the task.*

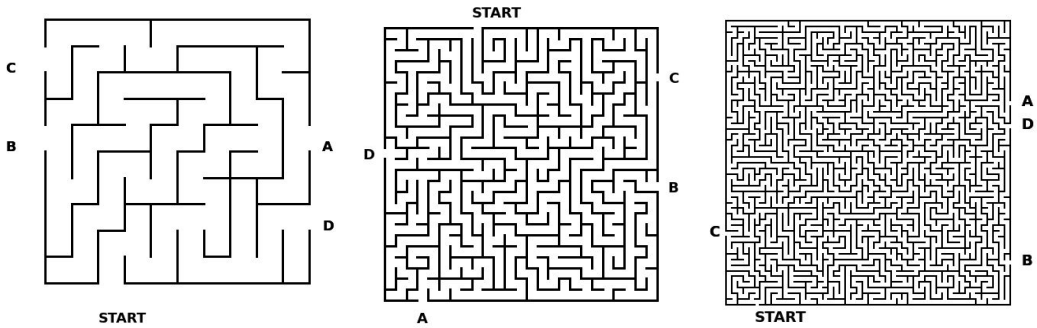We summarize these factors in Figure 3.

Fig. 4. Examples of the maze task that participants are asked to collaborate with the AI. **Left:** An example of our easy maze, which is a 10x10 grid. **Middle:** An example of our medium-difficulty maze, which is a 25x25 grid. **Right:** An example of our hard maze, which is a 50x50 grid.

4 METHODS

Our research aims to study the effects of explanations on human-AI decision-making. In the following section, we outline our experimental setup in which participants accomplish a task in collaboration with an AI system, with and without explanations, in different task difficulty settings. We situate our work in the context of human-AI decision making tasks. In such a task, a human is presented with some input and asked to make a decision, which may be an open ended answer or a multi-choice answer. To aid them with the task, the decision maker is partnered with an AI, which provides a prediction for the answer. In some conditions, the AI can also provide an explanation for their prediction. The decision maker is ultimately responsible for producing the final answer and has the option of utilizing or ignoring the AI's prediction and explanation.

4.1 Human-AI collaboration task

For the purposes of our study, we needed a task that fulfills the following requirements:

(1) *Multi-class classification*: In order to measure overreliance, we require that a random guessing strategy not produce the same answer as the AI very often: that is, we require a decision of higher cardinality than binary classification.

(2) *Non-trivial for humans to complete alone*: We require a task that presupposes the need for an AI's help. The AI should be able to help on the task by giving predictions; the explanations should reduce the cost of engaging with the task.

(3) *Lack of domain knowledge*: Since we are not testing our framework on a specific use-case, we needed a task that did not require specialized domain knowledge.

(4) *Enables different levels of task difficulty*: We require a task that can yield different levels of difficulty so that we can test their respective effects on overreliance.

(5) *Enables multiple types of explanations*: We require a task that can yield different explanation modalities so that we can test their respective effects on overreliance.

Prior XAI literature has used a variety of tasks, including binary classification [3, 4, 8, 54] and question answering about images [12, 15] and text [4, 26]. However, these tasks do not allow for the manipulations necessary for our experiments. Binary classification tasks, such as sentiment classification [4], would lead to 50% overreliance by random guessing, so it would be more difficult to measure the effects of task conditions on the participant's unconscious decision to use an overreliance strategy. Other tasks, such as answering questions about text [79] or predicting the

calorie count in images of food [9], are not amenable to controlling task difficulty (where "easy" is very little cognitive effort, and "hard" is very high cognitive effort).

In our paper, we use a visual search maze solving task, which meets all of our stated needs:

(1) *Multi-class classification*: Participants are shown a maze with a start position with four possible exits. Participants are asked to determine which of the four possible exits is the true exit. With 4 possible answers, the random chance of choosing the correct answer is 25%. When the AI is incorrect, overreliance with random guessing is also 25%.

(2) *Non-trivial for humans to complete alone*: Mazes, especially large mazes, can be difficult to complete alone; so an AI could conceivably help by making a prediction of which exit is correct. Furthermore, engaging with the task is difficult as the participant still must solve the maze to check. Therefore, including (easy to understand) explanations reduces the cost of engaging with the task.

(3) *Lack of domain knowledge*: Mazes do not require expertise or domain knowledge to complete.

(4) *Enables different levels of task difficulty*: We can manipulate the difficulty of the task by changing the dimensions of the maze.

(5) *Enables multiple types of explanations*: The maze task supports multiple explanation modalities, including highlight explanations, which are easy to check, and written explanations, which are more difficult to check. We visualize examples of these explanations in Figure 5.

Ecological Validity: Selecting maze solving trade s off some ecological validity compared to other tasks, as few human-XAI tasks outside of the lab will include solving mazes, but a rich literature in cognitive psychology and cognitive science involves solving puzzles of this sort (e.g., [2, 60]).

Even though AI are not often used to solve maze tasks in a collaborative setting, we argue that our maze task is comparable to game playing settings, some of which, like Chess or Go, are often used in human-AI collaborative settings [13]. Though these situations are not identical in nature (e.g., maze tasks do not have the same societal importance as a game like Chess, which may affect peoples' behavior), we argue that predictions and explanations made by an AI in the maze task have a similar structure as those made by an AI for a game like Chess. Our prediction in the maze task can be likened to the chess example of predicting the next move; our explanation in the maze task can be likened to the chess example of showing future roll-outs of the game (to justify its initial prediction).

Maze Selection: We programmatically generate mazes to be 10×10, 25×25, and 50×50 dimensions in height and width, corresponding to easy, medium, and hard tasks, respectively. To ensure that the mazes are of an appropriate difficulty, we pre-tested the questions on crowdworkers. We ran a study in which six crowdworkers on the Prolific platform answered the questions with no help from the AI. Questions from the easy mazes that did not yield 100% human accuracy were omitted (because we wanted the easy questions to truly be easy). And questions from the medium and hard mazes that did yield 100% human accuracy were omitted (because we wanted the medium and hard questions to truly reflect these difficulties).

4.2 AI Model

Simulated AI: To study overreliance in human-AI decision making, it is important that the AI accuracy is comparable to human accuracy [4]. When humans and AIs perform roughly equally for a given task, there is no additional incentive for the human decision-maker to over- or underrely on the AI's predictions. In pilot studies, the human accuracy was 83.5% across all difficulty conditions. Therefore, we control the number of incorrect AI predictions each participant sees such that AI accuracy is exactly 80% for all studies (unless specified in the study design section of a particular study). For the questions the AI gets incorrect, about 1/3 of them come from mazes with lengthy
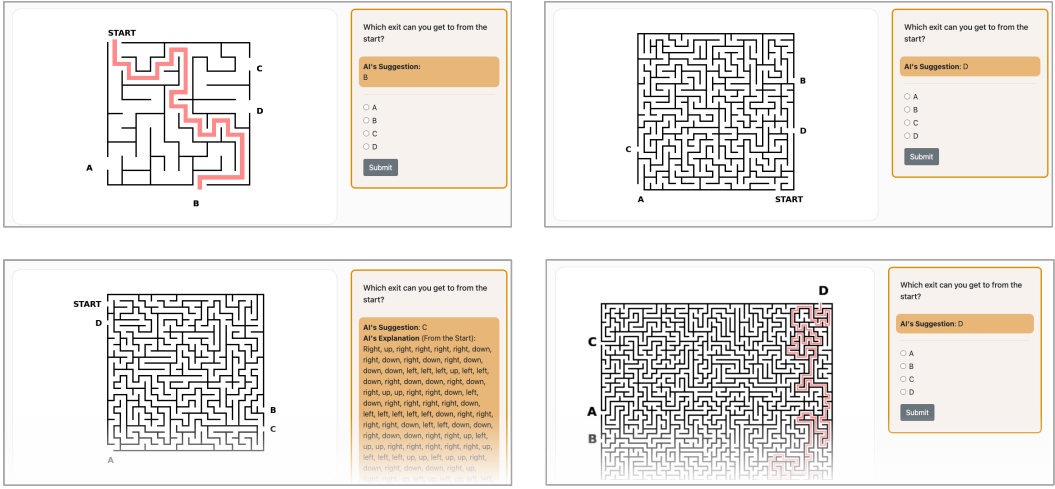
Fig. 5. An example of a maze solving task in which the AI's suggestion is correct. We visualize three different task difficulty conditions and two different explanation conditions. **Top Left:** The *easy task, highlight explanation* condition. The maze is 10×10 and features highlights of the path the AI suggests. The AI's suggestion is provided above the answer choices. **Top Right:** The *medium task, prediction* condition. The maze is 25×25 and does not show explanation highlights. **Bottom Left** The *medium task, written explanation* condition. The maze is 25x25 and features a written version of the path the AI suggests. **Bottom Right:** The *hard task, highlight explanation* condition. The maze is 50×50 and features highlights of the path the AI suggests.

path solutions, 1/3 from mazes with medium-length path solutions, and 1/3 from mazes with short path solutions. This ensures that participants in our study do not notice a bias in the type of mazes the AI gets incorrect.

Simulated Explanations (And why they're considered explanations): For our main studies , we generate two types of explanations: highlight and writing explanations (Figure 5). Although what constitutes an explanation is debated by philosophers [76], we employ the following definition from [34] that characterizes explanations as describing how the "explanandum" (that which is being explained, or, in our case, the AI's predictions) is logical consequence of given conditions or premises. The explanations we created for this work, much like similar work in XAI, are local explanations in a decision-making context [69]. These explanations attempt to provide people with insights of *how* the AI came to its decision. While our explanations give this decision-making information by indicating the path the AI took, other explanations, like SHAP [64], provide features that contributed most to the AI's outcome.

Highlight Explanations: Our highlight explanations indicate the path between the start point and the predicted exist. Highlight explanations have commonly been used in prior XAI studies [4, 26, 33, 53, 71, 97]. We argue that our highlight explanations constitute as an explanation given by the definition from [34]. The conclusion (the exit to take to get out of a maze) follows from the premise (the steps taken to an exit in a maze). Since the highlight explanation describes visually the premise, this is an explanation.

Our two other, highlight based explanations, *incomplete* and *salient* show the highlights in a similar manner. However, in these cases, if the AI has crossed a wall (an illegal move in the maze setting) to reach its answer, the incomplete explanation does not show the path after the wall and

the salient explanation highlights the path after the wall in blue. This is designed to make the AI's errors more obvious.

Written Explanations: For written explanation, we translated the ground-truth path into words (e.g. "left, up, right, up..."). Our written explanations are more analogous to logical written arguments explaining the process of arriving at the answer [4], and in our context, *increase* the cost of engaging with the task (using the explanations) relative to the highlight explanation, as participants have to parse the writing and map it onto the maze. We argue that our written explanations also constitute a viable explanation, since it describes, in a written format, the premise (the steps taken to an exit in a maze) which logically leads to the conclusion (the exit to take to get out of the maze).

Generation of explanations: We generate "accurate" explanations when an AI's prediction is correct. When the AI is incorrect, our explanations indicate a path that goes through a maze wall. Evaluating with "accurate" explanations ensures that our results are not confounded by faulty or uninformative AI-generated explanations.

4.3 General Study Design

All of our five studies follow a similar template (unless otherwise noted in the particular study's procedure section). The number of trials and exact details are also noted in the procedure section of every study.

All participants are onboarded, where they are asked demographic questions (age and gender). Then, they complete a training phase where they complete the mazes of their task difficulty condition alone. Then, they complete a training phase with the AI in their task difficulty and AI condition. In both training phases, they are given immediate feedback on whether their answers were correct or not. We chose to include this training phase as preliminary pilot participants displayed unstable behavior before familiarizing themselves with the AI. This allowed participants to become familiar with the task, the AI, and the AI's explanations, if applicable. After both training phases, they are placed in a collaboration phase with the AI in their task difficulty and AI condition. They are not given feedback on whether their answers were correct or not. We removed feedback on answer correctness as we did not want their judgements of the AI's ability to change across the collaboration phase.

Finally, we collect subjective self-report measures on how much participants trusted the AI, on how they interacted with the AI, and on their need for cognition (NFC) score [18], a stable personality trait measuring people's propensity towards completing cognitively difficult tasks. We measured this as it might be indicative of person-dependent differences in cognitive engagement in our task. Unless otherwise noted, each self-report question is assessed using a 7-point Likert scale. The full set of self-report questions and NFC questions are provided in the appendix.

Mitigating effects of trust: In light of trust's well-documented relation to reliance [4, 9, 10, 55, 59, 96, 97], we aim to neutralize any effects of trust in our task such that any differences in over-reliance are attributable to our cost-benefit manipulations and not to individual differences in trust. To do so, we remove any task or interface components with known relations to trust. In particular, we abstain from anthropomorphizing the AI, always referring to the AI as "the AI", as Anthropomorphization has been linked to increased trust [35]. We always refer to AI predictions as "suggestions" [74] to counteract perfect automation schemas [21] and emphasize that this suggestion can be rejected. As participants complete a series of maze solving tasks, we do not provide immediate feedback since humans are particularly sensitive to seeing algorithms err [19]. Lastly, we
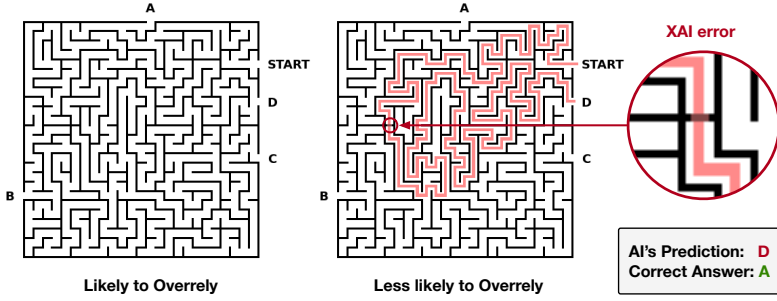
Fig. 6. **Left:** The *medium-difficulty task, prediction only* condition. The maze is 25 × 25 and features the prediction only. **Right:** The *medium-difficulty task, explanation* condition. The maze is 25 × 25 and features explanations in the form of inline highlights. An example of when the AI gives an incorrect prediction and explanation.

do not provide any information about the model's accuracy, as making this explicit also modulates trust [55, 95–97].

4.4 Participant Selection

All of our participants are recruited via Prolific (prolific.co), an online crowdsourcing platform. We only included participants that had at least 50 submissions, were located in the United States, were native English speakers, had an approval rating of at least 95% on Prolific, and that had not completed our task before. All participants received a payment of $4 for 20 minutes of their time with no bonuses, unless otherwise noted in a specific study's procedure section. We estimated the time required to complete the tasks through pilot experiments.

4.5 Measured Variables

We collect the following objective performance measures for our hypotheses:

(1) *Overreliance:* We define overreliance as the percentage of decisions when a human decision maker accepts an AI's incorrect prediction [9]. For example, if a decision-maker agreed with the AI's prediction 80% of the time when the AI was incorrect, then the overreliance is 80%.

(2) *Need for Cognition (NFC) Score [18] :* NFC is a stable personality trait measuring people's propensity towards completing cognitively difficult tasks. We average the responses to the six NFC questions (e.g., "I would prefer complex to simple problems."), which are available in the appendix, to compute a person's NFC score.

5 STUDY 1 – MANIPULATING COSTS VIA TASK DIFFICULTY

In our first study, we explore PREDICTION 1 by manipulating how difficult the task is to do in the presence or absence of explanations. The presence of an explanation (e.g., highlighting of relevant information) can decrease the cognitive effort required to engage with the task. However, explanations may not dramatically reduce the effort required to solve a task if the task is easy for a human to solve on their own. We pre-registered this study (osf.io/vp749 and osf.io/g6tjh).

5.1 Conditions

We adopt a two-factor mixed between and within subjects study design in which each participant sees one of two *AI conditions* in multiple task *difficulty* conditions. Half of the participants see both easy and medium task difficulty conditions, and the other half of participants see only the hard

condition, since the amount of time required to complete the task was approximately the same in these two study configurations. The differences between conditions are depicted in Figure 4 and 5.

Task difficulty conditions: The task difficulty condition manipulates the maze difficulty. This manipulation modifies cognitive effort because it requires more cognitive effort to search for the correct path in a longer maze [67]. Note that quality of questions, predictions, and explanations remain the same in both task difficulty conditions.

(1) *Easy:* In the easy condition, participants see a 10×10 maze.
(2) *Medium:* In the medium condition, participants see a 25×25 maze.
(3) *Hard:* In the hard condition, participants see a 50×50 maze.

AI conditions: The AI condition manipulates the presence of explanations for an AI's prediction.

(1) *Prediction:* In the prediction condition, participants are only provided with the AI's prediction. The AI's predicted answer is displayed directly below the question.
(2) *Explanation:* In the explanation condition, participants are provided with a highlight explanation in addition to the AI's prediction. Highlights are displayed directly on the maze.

5.2 Hypotheses

Our first prediction expects that explanations will reduce overreliance when they reduce the cost of engaging with the task; similarly, explanations will not reduce overreliance when the reduction of cost in engaging with the task is small (e.g. in the easy task). Additionally, we include a hypothesis to test if there are individual differences in propensity to engage with the task (H1E).

Therefore Study 1 has 5 hypotheses:

Hypothesis 1A: In the prediction only case, overreliance is **lower** in the easy task than in the medium-difficulty task.

Hypothesis 1B: In the easy task, overreliance is **the same** in the prediction only case and prediction with explanation case.

Hypothesis 1C: In the medium-difficulty task, overreliance is **lower** in the prediction with an explanation case than in the prediction only case.

Hypothesis 1D: In the hard task, overreliance is **lower** in the prediction with an explanation case than in the prediction only case.

Hypothesis 1E: In the hard task, overreliance is affected by an **interaction effect** between participants' Need for Cognition (NFC) scores and the AI Condition.

5.3 Procedure

For participants in the easy and medium condition, during the initial human-only phase, participants are asked to do two easy mazes and two medium mazes alone. During the initial AI training phase, participants are prompted to complete ten questions with an AI aid to experience using it, where half are easy and half are medium. The AI agent always offers a prediction for each question; in the explanation condition, it also highlights the path in the maze. We ensure that the AI errs on two of the ten training questions (one in easy and one in medium) and randomize the order in which participants encounter these ten questions.

After the training phase, participants are asked to complete thirty test questions in collaboration with the same AI agent. Exactly six of the thirty questions have incorrect AI predictions and their positions are randomized.

For participants in the hard condition, given how long it takes to complete the task, we decrease the total number of trials. Therefore, we cut the number of mazes in half in each phase of the study
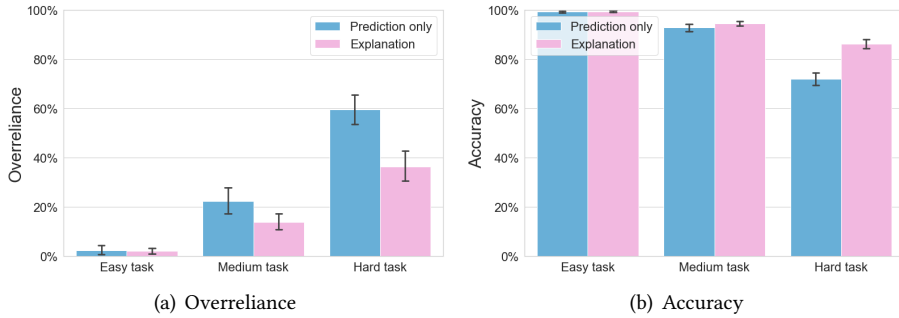
Fig. 7. **(a)** Overreliance levels in study 1. In the easy and medium-difficulty tasks, we find no differences between prediction and explanation conditions. We find that explanations reduce overreliance in the hard task condition. **(b)** Accuracy levels in study 1. In exploratory analysis, we find that explanations increase decision-making accuracy in the hard task condition. These findings together suggest that the gains provided by explanations are more likely to be found when tasks are hard for people to do.

and only showed participants the hard condition. This, however, does not affect how hard people perceived the task to be, since (1) people did not have access to how many trials they had left, and (2) qualitative data shows people found the hard task to be more difficult than the other two tasks.

5.4 Participants

We recruited a total of $N = 340$ participants ($N = 170$ in prediction only, $N = 170$ in explanation, half of which are in the easy and medium condition, and the other half of which are in the hard condition). Our sample size comes from a power analysis done on pilot data.

5.5 Results

We analyze our results using a Bayesian[3] linear mixed effects model (see Table 1, 2). We visualize our results in Figure 7. We adopt the convention of saying that a comparison is notable if the 95% credible interval of the posterior distribution excludes 0 and the estimate of the mean is positive or negative (depending on the direction of the hypothesis). *Note*: We provide the model specification in the appendix.

5.6 Summary

We find support for H1A, which provides empirical evidence for the following claim: as tasks increase in difficulty, overreliance on AI's predictions increases. This can be explained in the following way through our cost-benefit framework: an increase in task difficulty increases the cost of following the strategy to engage with the task, as it generally takes more cognitive effort for people to either (a) ensure that the AI has gotten the answer correct when provided with only predictions or (b) complete the task alone. However, the strategy to rely on the AI does not incur the same costs, as people are not engaging with how hard the task is. Since the strategy to rely on the AI inherently results in overreliance, overreliance will increase, as seen here.

---

[3] We adopt Bayesian statistics for its advantages in being able to hypothesize null effects [27, 73]. Bayesian statistics has also been proposed as a better alternative to frequentist statistics for HCI research [43]. For more information on our analysis, please see the appendix section.

Table 1. Results from study 1. We find that overreliance increases as tasks get harder; overreliance does not change with explanations in easy tasks; and overreliance decreases with explanations in hard tasks. This supports our hypothesis that explanations are more likely to provide assistance, and therefore reduce overreliance, when tasks are hard. This fits into our cost-benefit framework, as explanations in hard tasks sufficiently reduce the cost of doing a task properly.

Parameter	Mean	Lower 95% CI	Upper 95% CI	Notable	Hypothesis	Confirmed?
Easy Pred - Medium Pred	-3.967	-5.480	-2.53	*	H1A	**yes**
Easy Pred - Easy XAI	0.164	-1.678	2.17		H1B	**yes**
Medium Pred - Medium XAI	0.916	-0.305	2.08		H1C	no
Hard Pred - Hard XAI	1.74	0.889	2.76	*	H1D	**yes**

Table 2. Results from study 1. We do not find support for our hypothesis that there is an interaction effect of AI Condition and a participants' Need for Cognition (NFC) score. We hypothesized that people who are more likely to engage in cognitively demanding tasks would have little differences in overreliance levels regardless of whether AI explanations were provided or not. However, we did not find support for this, which we hypothesize could be due to the fact that, in the hard task, even participants with high propensities to engage in effortful thinking are likely to overrely. An initial exploratory analysis of an interaction between AI Condition and NFC in the medium task condition provides some evidence to support this claim.

Parameter	Mean	Lower 95% CI	Upper 95% CI	Notable	Hypothesis	Confirmed?
AI Condition:NFC	0.35	-0.76	1.49		H1E	no

We find support for H1B, which provides empirical evidence for the following claim: in easy tasks, predictions and explanations will provide similar levels of overreliance. This can be explained in the following way through our cost-benefit framework: since, in the easy task, explanations do not make engaging with the task any easier (i.e. they do not provide any reduction in costs), the utility of the two strategies do not change. Therefore, overreliance levels do not change.

We do not find support for H1C; therefore not providing empirical evidence for the following claim: in medium-difficulty tasks, there is more overreliance in the prediction condition than in the explanation condition. This may be in due part to the fact that this condition is not hard enough to see the reductions in cost that we expect to see with explanations.

We find support for H1D, which provides empirical evidence for the following claim: in hard tasks, there is more overreliance in the prediction condition than in the explanation condition. This can be explained in the following way through our cost-benefit framework: in hard tasks, explanations provide extra assistance to completing the task over predictions only, greatly reducing the cost of engaging with the task. Since the costs have been greatly reduced and are relative to the costs of what would otherwise be incurred (e.g. by completing the task alone or just using an AI's prediction), the explanation looks far more "useful" as a means of reducing effort. Therefore, overreliance levels reduce in this case since people are more likely to use the strategy to engage in the task.

We do not find support for H1E; therefore, not providing empirical evidence for the following claim: there is an interaction effect between participants' Need for Cognition (NFC) scores and the type of explanation modality (highlights of lackthereof) when measuring overreliance. This could be
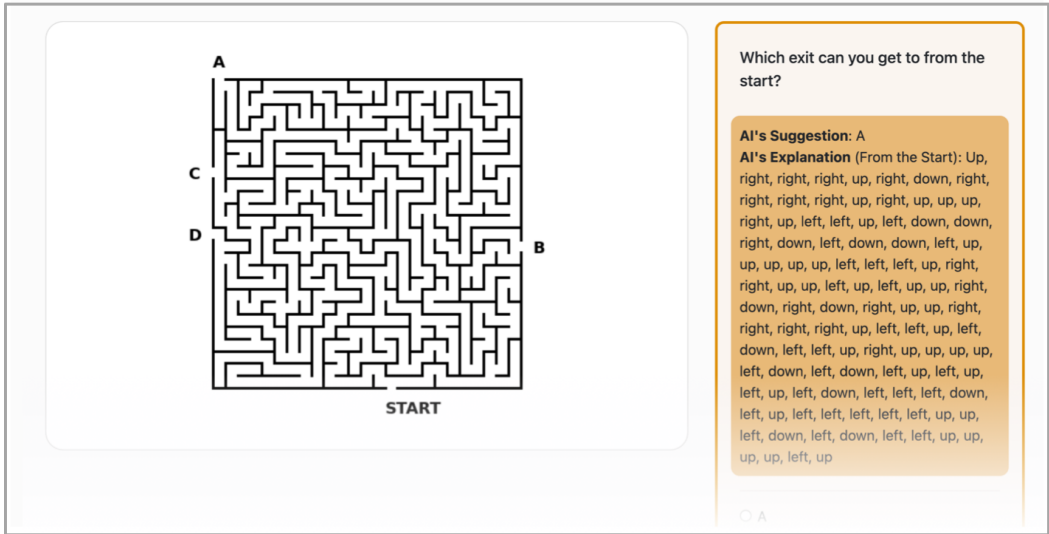
Fig. 8. The *medium task, written explanation* condition. The maze is 25 × 25 and features a written version of the path the AI suggests. In the written explanation, each word corresponds to one step in the maze. We find that these written explanations do not reduce overreliance, compared to not receiving an explanation.

because the hard task is too difficult to demonstrate differences in behavior across peoples' NFC, since most people are likely to overrely on the AI's prediction anyway. In an exploratory analysis, we find a interaction between AI condition and Need for Cognition (NFC) in the medium task. We also find that, as NFC scores increase, the difference between overreliance in prediction and explanation diminishes. The medium task could be a sufficiently difficult enough task to demonstrate differences in behavior across NFC.

Additionally, we ran an exploratory analysis to measure the effects of explanations on human-AI decision-making accuracy. Our exploratory analysis uncovers an instance where explanations increase human-AI accuracy. In the hard task, explanations increase the accuracy compared to the prediction only condition (see Figure 7b). This finding stands in contrast against prior work [4, 9], which has failed to find differences between the conditions.

In summary, explanations do not produce an observable reduction in overreliance in easy or medium tasks, but do reduce overreliance in the hard task. This result demonstrates that overreliance can in fact be mitigated by explanations—and that this occurs precisely when tasks are substantially complex and effortful that reviewing the explanation yields a substantial cognitive effort benefit.

6 STUDY 2 – MANIPULATING COSTS VIA INCREASING EXPLANATION DIFFICULTY

In our second study, we explore PREDICTION 2 by manipulating how difficult the explanation is to parse. We hypothesize that explanations that are easier to understand will have lower levels of overreliance. We pre-registered this study (osf.io/4dbqp) .

6.1 Conditions

We adopt a two-factor mixed between and within subjects study design in which each participant sees one of two *explanation difficulty* conditions in multiple task *difficulty* conditions. Half of the participants see both easy and medium task difficulty conditions, and the other half of participants see only the hard condition, since the amount of time required to complete the task was approximately the same in these two study configurations.

Since we wanted to manipulate the difficulty of the explanation in order to measure its effects on overreliance, we created an easy-to-understand explanation and a hard-to-understand explanation. We did so by creating visually easy-to-parse highlight explanations overlaid on the maze. These easy explanations are contrasted with written explanations that incur more cognitive effort, since participants need to follow along these written explanations that are on the side of the maze (see Figure 8).

Explanation conditions: The explanation condition manipulates the difficulty of the explanations provided to people.

(1) *Highlight explanation:* Participants see the predictions as well as explanations, which are displayed by directly highlighting the proposed path to the AI's chosen exit.
(2) *Written explanation:* Participants see the predictions as well as explanations, which are displayed by writing the path alongside the maze (e.g. "left, left, up, right, down...") that the AI takes from the start.

6.2 Hypothesis

Our second set of hypotheses further addresses the connection between overreliance and the cognitive effort required to engage with a task by manipulating explanation costs. Specifically, in our second study, we adjust the cost of engaging with the task by manipulating explanation difficulty. Explanations can decrease the cognitive effort of engaging with the task, insofar as the content of the explanation makes doing the task sufficiently easier. PREDICTION 2 expects that explanations that are easier to parse have less overreliance. Therefore STUDY 2 has 2 hypotheses:

> HYPOTHESIS 2A: In the medium-difficulty task, overreliance is **lower** in the highlight explanation case than in the written explanation case.

> HYPOTHESIS 2B: In the hard task, overreliance is **lower** in the highlight explanation case than in the written explanation case.

6.3 Procedure

We adopt the same study procedure for Study 1.

6.4 Participants

We recruited an additional $N = 170$ participants from Profilic, using the same exclusion criteria as in Study 1. All were placed in the written explanation condition. Half were placed in the easy and medium task difficulty condition; half were placed in the hard task difficulty condition. The highlight explanation condition has 170 participants, from the data collected in Study 1. This gives us a total of $N = 340$ participants in Study 2. Our sample size comes from a power analysis done on pilot data.

6.5 Results

We analyze our results using two Bayesian linear mixed effects models (for H2A and H2B, respectively; see Table 3). We visualize our results in Figure 9. We adopt the convention of saying that a comparison is notable if the 95% credible interval of the posterior distribution excludes 0 and the
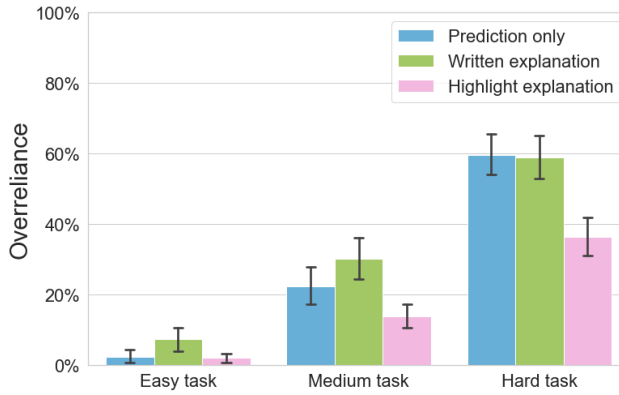
Fig. 9. Overreliance levels in study 2. We find that overreliance is greater in the written explanations condition than in the highlight explanation condition. This supports our hypothesis that explanations that are easier to parse will generally yield lower levels of overreliance.

Table 3. Results from study 2. We find support for our hypotheses that there are lower levels of overreliance when provided an explanation in the form of highlights (easy to parse), as opposed to in the form of written language (hard to parse).

Parameter	Mean	Lower 95% CI	Upper 95% CI	Notable	Hypothesis	Confirmed?
Highlight - Written in Medium	-1.63	-2.78	-0.559	*	H2A	**yes**
Highlight - Written in Hard	-1.632	-2.570	-0.694	*	H2B	**yes**

estimate of the mean is positive or negative (depending on the direction of the hypothesis). *Note*: We provide the model specification in the appendix.

6.6 Summary

We find support for H2A, which provides empirical evidence for the following claim: as explanations increase in difficulty to understand, overreliance on the AI increases. This can be explained in the following way through our cost-benefit framework: an increase in explanation difficulty increases the cost (via cognitive effort or time) to engage with the task. This increase in cost makes the strategy to engage with the task seem less appealing or "useful." As people engage less with this strategy, they might follow the other strategy: rely on the AI. As people are more likely to choose the strategy to rely on the AI in this case, they will have higher levels of overreliance.

We find support for H2B, which provides empirical evidence for the following claim: as explanations increase in difficulty to understand, overreliance on the AI increases. This can be explained in the same way as H2A can.

Additionally, we ran exploratory analyses which showed, across all task difficulty (easy, medium, and hard) conditions, there was no difference between the prediction only case and the hard-to-understand, written explanation case. This suggests that these hard-to-understand explanations do not act as an additional signal for trust (which would increase overreliance).

In summary, Study 2 demonstrates that overreliance is responsive not just to the cognitive effort of completing the task, but also to the effort to understand an explanation.

7 STUDY 3 – MANIPULATING COSTS VIA DECREASING EXPLANATION DIFFICULTY

In our third study, we explore how decreasing the difficulty of understanding explanations can reduce overreliance. In accordance with the results of study 2, we postulate that explanations with lower levels of difficulty (to understand) will have lower levels of overreliance. This experiment tries to push the boundaries of displaying *extremely obvious* explanations to see if there is a discernible floor effect to overreliance (e.g. will there always be *some* overreliance, or can we get it to zero?) This experiment is exploratory and therefore does not have pre-registration.

7.1 Conditions

We adopt a between-subjects experiment, wherein participants are placed in one of five explanation conditions. Participants are placed in the hard task difficulty condition, since we wanted to see how much we could reduce overreliance (and the hard task has the highest levels of overreliance in the prediction only case).

AI conditions: The AI condition manipulates the presence of and difficulty of the explanations provided to people.

(1) *Prediction:* Participants are only provided with the AI's prediction.
(2) *Highlight explanation:* Participants see the predictions as well as explanations, which are displayed by directly highlighting the proposed path to the AI's chosen exit.
(3) *Written explanation:* Participants see the predictions as well as explanations, which are displayed by writing the path alongside the maze (e.g. "left, left, up, right, down...") that the AI takes from the start.
(4) *Incomplete explanation:* Participants see the predictions as well as a highlighted path to the AI's chosen exit. If the AI crosses a wall in the maze, the rest of the path is not shown.
(5) *Salient explanation:* Participants see the predictions as well as a highlighted path to the AI's chosen exit. If the AI crosses a wall in the maze, the rest of the path is highlighted in blue.

7.2 Hypothesis

We did not have pre-registered hypotheses for two reasons: (1) this is an exploratory study to find if there is a floor effect and (2) our hypotheses relating explanation difficulty and overreliance have been confirmed in STUDY 2. However, we were guided by the following question: *how much can we reduce overreliance by having explanations that make it easy to spot mistakes?*

7.3 Procedure

We adopt the same study procedure for Study 1 in the hard task.

7.4 Participants

We recruited an additional $N = 31$ participants from Profilic, using the same exclusion criteria as in Study 1. All were placed in the hard condition. 16 were in the incomplete explanation condition; 15 were in the salient explanation condition. (Our sample size comes from a power analysis done on pilot data.) There were $N = 85$ in each of the prediction only, highlight explanation, and written explanation conditions from study 1 and 2. In total, there were $N = 286$ participants in this study.
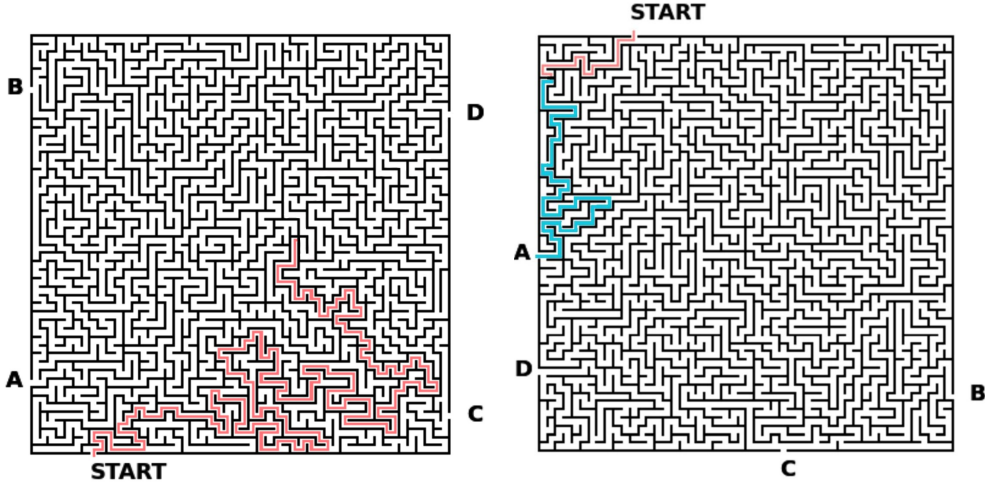
Fig. 10. An example in which the model is incorrect in (a) the incomplete explanation, where the the explanation does not continue after the AI goes over a wall, and (b) salient explanation, where the explanation is highlighted in blue when the AI goes over a wall. We designed these explanations to push the boundaries of explanations' saliency. In study 3, we validated that, the more salient an explanation makes an AI's mistake, the less overreliance there is.
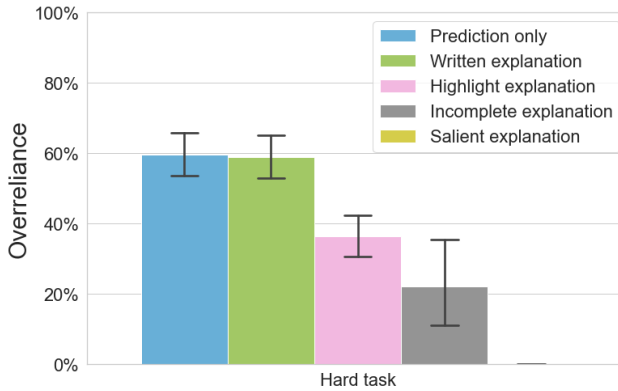
Fig. 11. Overreliance levels in our exploratory study, study 3. We find that (1) more salient explanations reduce overreliance over less salient explanations and (2) there is no "floor" effect on overreliance; that is, we observe that people will not overrely if they *know* the AI is wrong.

7.5 Results

We analyze our results using a Bayesian linear mixed effects model (see Table 4). We visualize our results in Figure 11. Because we didn't pre-register any hypotheses for this study, we report the means and credible intervals for all the pairwise comparisons here, but don't indicate whether the differences are credible. *Note*: We provide the model specification in the appendix.

Table 4. Results from our exploratory study, study 3. We find that, the more salient an explanation is, the less overreliance there is. Additionally, we did not observe a "floor" effect on overreliance; that is, we observe that people will not overrely if they *know* the AI is wrong.

Comparison	Mean	Lower 95% CI	Upper 95% CI
Prediction - Incomplete Explanation	3.336	1.309	5.36
Written Explanation - Incomplete Explanation	3.205	1.236	5.38
Highlight Explanation - Incomplete Explanation	1.479	-0.490	3.60
Prediction - Salient Explanation	41.698	5.055	153.76
Written Explanation - Salient Explanation	41.570	5.316	153.95
Highlight Explanation - Salient Explanation	39.816	3.843	152.03
Incomplete Explanation - Salient Explanation	38.232	1.971	150.90

7.6 Summary

Our exploratory analysis finds that the salient explanation condition has less overreliance than all other conditions (prediction, highlight explanation, written explanation, and incomplete explanation). We also find that the incomplete explanation condition has less overreliance than the prediction only and written explanation conditions, more overreliance than the salient explanation condition and finds no difference in overreliance in the highlight explanation condition. Complementary to Study 2, Study 3 provides further support for our hypotheses.

Furthermore, we do not find a "floor" effect, i.e. an effect in which we cannot reduce overreliance to 0%. The salient explanation condition, our most most obvious explanation condition, has an average overreliance rate of 0%. Although there is more work to be done whether this is the case with more participants, these initial findings suggest that people are unlikely to agree with an AI if they *know* it is incorrect.

8 STUDY 4 – MANIPULATING BENEFIT VIA MONETARY BONUS

In our fourth study, we explore PREDICTION 3 to validate the "benefit" aspect of our cost-benefit framework. In this study, we aim to show that increasing the benefit of doing the task properly (i.e. getting the answer correct) reduces overreliance. We pre-registered this study (https://osf.io/hgz2x).

8.1 Conditions

We adopt a two-factor mixed between and within-subjects experiment in which each participant sees both *bonus conditions* and sees one of two *AI conditions*. All participants are placed in the hard task difficulty condition. Half of the participants see the prediction only condition; half of the participants see the highlight explanation condition.

AI conditions: The AI condition manipulates the presence of explanations for an AI's prediction.
  (1) *Prediction:* Participants are only provided with the AI's prediction. The AI's predicted answer is displayed directly below the question.
  (2) *Explanation:* Participants are provided with a highlight explanation in addition to the AI's prediction. Highlights are displayed directly on the maze.
Bonus conditions: The bonus condition manipulates how much bonus each maze completed is worth.
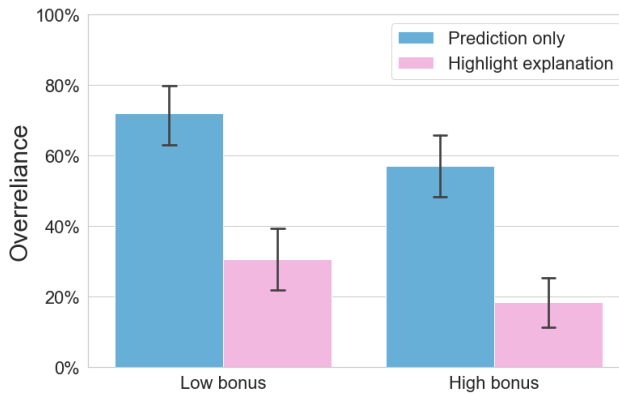
Fig. 12. Overreliance levels in study 4. We hypothesize and validate that, the more bonus provided for completing the task properly, the less overreliance there is.

(1) *Low Bonus:* Participants are given $0.01 for each maze done correctly.
(2) *High Bonus:* Participants are given $0.50 for each maze done correctly.

8.2 Hypothesis

Our third set of hypotheses addresses the connection between overreliance and benefit of completing the task. Specifically, in our fourth study, we adjust the benefit of each strategy by controlling the amount of bonus given. This influences how desirable a strategy such as rely on the AI looks to a participant, given that they will not get the bonus if the AI is wrong and they overrely. Therefore STUDY 4 has two hypotheses:

HYPOTHESIS 3A: Overreliance is **lower** in the high bonus case than in the low bonus case.

HYPOTHESIS 3B: Overreliance is **lower** in the highlight explanation case than in the prediction case. *Same hypothesis as H1D; we retest this to ensure that our study changes in bonus and amount of questions do not affect our previous finding.*

8.3 Procedure

We follow the same study procedure as study 1 in the hard task, with two exceptions. First, we adopted a blocked, within-subjects study design in which participants are randomly placed in one bonus condition for the first half of the study and the other condition for the second half[4]. We chose this study design because prior work [86] shows that people calculate their judgements of monetary gains and losses relative to a certain reference point, which was indicative to us that we ought to have within-subjects comparisons of monetary value. Second, we add one question where the AI is incorrect to the collaboration phase as to have an even amount of questions with the low- and high- bonus.

8.4 Participants

We recruit ed $N = 114$ participants for our study, half of whom (57 participants) are in the prediction condition, the other half (57 participants) are in the explanation condition. Our sample size comes from a power analysis done on pilot data.

---

[4]The pre-registration does not specify a blocked study design.

Table 5. Results from study 4. We find that overreliance decreases with more bonus provided for completing the task properly. This affirms the "benefit" element of our cost-benefit framework, showing that changes in reward or incentive structures affect peoples' behavior. Additionally, we reproduce a finding from study 1 that there are lower levels of overreliance when given a highlight explanation in a hard task, even when bonuses are added.

Comparison	Mean	Lower 95% CI	Upper 95% CI	Notable	Hypothesis	Confirmed?
Low Bonus - High Bonus	0.93	0.414	1.46	*	H3A	**yes**
Prediction - Highlight XAI	2.47	1.72	3.36	*	H3B	**yes**

8.5 Results

We analyze our results using two Bayesian linear mixed effects models (for H3A and H3B, respectively; see Table 5). We visualize our results in Figure 12. We adopt the convention of saying that a comparison is notable if the 95% credible interval of the posterior distribution excludes 0 and the estimate of the mean is positive or negative (depending on the direction of the hypothesis). *Note*: We provide the model specification in the appendix.

8.6 Summary

We find support for H3A, which provides empirical evidence for the following claim: when the benefit (in this case, via monetary bonus) of doing a task properly increases, overreliance decreases. This can be explained by our cost-benefit framework in that increasing the benefit of completing the task properly makes the utility of the strategy to engage with the task higher than the strategy to rely on the AI, since the latter strategy is likely to lead to more mistakes. Since the strategy to rely on the AI leads to the most overreliance and this is the strategy of which the utility is the lowest when benefits (of completing the task properly) increase, we can then say that the overreliance is decreased because of this utility comparison.

We find support for H3B, which provides empirical evidence for the following claim: in hard tasks, highlight explanations reduce overreliance compared to only receiving the AI's prediction. This replicates the finding in H1D.

Taken together, Study 3 demonstrates that overreliance is responsive not just to costs, but also to the benefit accrued for expending these costs. These findings also have implications on the designs of crowd-sourced studies, for example, as findings may differ when providing a bonus or not in a study.

9 STUDY 5 – MEASURING THE UTILITY OF XAI METHODS

In our fifth and final study, we measure the effect of task difficulty and explanation difficulty on the subjective utility participants assign to the AI. We explore PREDICTION 4 and PREDICTION 5 to validate our final two claims: (1) people attach higher utility to an AI's explanation when tasks are harder and (2) people attach higher utility to an AI's explanation when it is easier to understand. This study aims to make explicit the choice people make between strategies (e.g. whether to rely on the AI or not). This final study validates our framework by showing that peoples' utilities of an AI are predicted by it. We pre-registered this study (https://osf.io/cskvb).
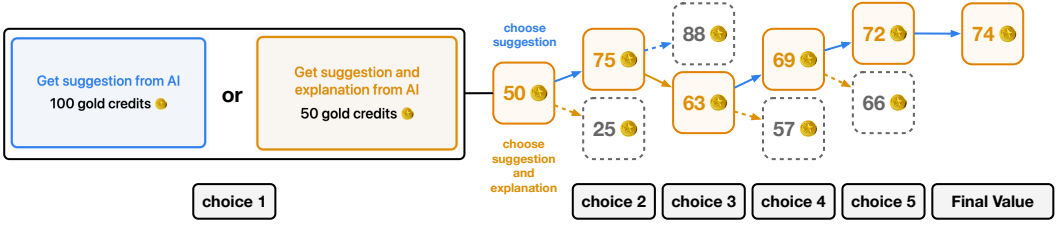
Fig. 13. An overview of the cognitive effort discounting (COG-ED) task in Study 5. Participants are asked to choose between completing with the AI's suggestion for 100 gold credits or completing with the AI's suggestion and explanation for a dynamically shifting price.

9.1 Cognitive Effort Discounting (COG-ED)

To quantify participant's subjective utility of adopting various cognitive strategies when collaborating with an AI, we adapt the Cognitive Effort Discounting (COG-ED) paradigm [94] to human-AI collaboration (Figure 13). During the standard COG-ED protocol, as originally designed [94], participants are presented with a series of choices between a high-effort task for a fixed monetary reward and a low-effort task for a lower monetary reward. The COG-ED paradigm takes in this series of decisions and outputs a number representing the participant's perceived value of completing the low effort task.

COG-ED reaches this output number by adjusting the reward of the low effort tasks depending on the task participants choose (Figure 13). Consider an example in which a participant makes the first choice between completing hard task for a $1 reward or an easier task for a $0.50 reward. If the participant chooses the harder task, then the reward for the easier task increases for the next choice. If the participant chooses the easier task, then the reward for the easy task decreases for the next choice. The specific formula is detailed below. This process continues iteratively. This process is equivalent to a binary search that converges after a logarithmic number of choices. At this converged point, the subjective utility of both tasks are assumed to be equal, allowing the difference in reward to quantify the subjective utility of the low-effort task relative to the high-effort task.

In two extreme conditions, if the participant always chooses the harder task, the reward for the easier task will increase and leave a small difference in reward between the harder and easier task. This small difference indicates that participant does not see the low-effort task as reducing the effort enough to be worth the lower reward. In contrast, if the participant always chooses the easier task, the reward for the easier task will decrease and leave a large difference in reward between the harder and easier task. This large difference indicates that the participant sees the low-effort task as reducing the effort enough to be worth receiving lower pay. Thus, the COG-ED paradigm indicates the person's valuation of the reduced effort in the low-effort task.

In behavioral economic terms, the COG-ED paradigm is an example of a choice experiment, a common method for inferring subjective utilities [93]. The assumption underlying such *choice experiments* is that the indifference point, i.e. the point at which a decision-maker freely chooses between two conditions at roughly equal rates, implies that the decision-maker associates both conditions with equal utilities [93]. In human-computer interaction, similar choice experiments have previously been employed to assess the utility of user interfaces [85].

9.2 Conditions

We modify existing COG-ED experiments for human-AI teams by asking participants to choose between completing the task alone (high-effort) versus completing the task with the AI (low-effort).

Participants are placed in one of two within-subject scenarios.

Within-Subjects Scenarios:

(1) *Comparison between AI's utility in different task difficulty conditions.* Participants work with two AI agents– one of which can generate highlight explanations in the medium task, the other of which can generate highlight explanations in the easy task.
We presume that these results can also extrapolate to comparisons between easy and hard, as well as medium and hard. This is because the predicted effect size between easy and medium is smaller than the other comparisons.

(2) *Comparison between AI's utility in different explanation conditions.* Participants participate with two AI agents – one of which can generate highlight explanations in the medium task, the other of which can generate written explanations in the medium task.
We test these in the medium task since (1) the easy task was unlikely to find any differences in behavior and (2) we presume any results can also extrapolate to the hard task, since in prior studies we found that effect sizes of manipulations were much smaller in the medium task than the hard task.

9.3 Hypothesis

PREDICTION 4 states that people will attach higher subjective utility to an AI's explanation in harder tasks. This can be adapted to our study with the following hypothesis:

> HYPOTHESIS 4A: The subjective utility of the AI is **lower** in the easy task than in the medium-difficulty task.

PREDICTION 5 states that people will attach higher subjective utility to an AI's explanation when it is easier to understand. This can be adapted to our study with the following hypothesis:

> HYPOTHESIS 5A: The subjective utility of the AI is **lower** in the written explanation case than in the highlight explanation case.

9.4 Procedure

Similar to the first two studies, this study is divided up in an AI training and a testing phase (there is no human-only training) in each within-subjects condition. We made only one change to the study as described in the preregistration: instead of completing two questions without the AI at the beginning, the participant completes two questions with the AI (prediction only) in the beginning. Then, participants complete five questions with the AI (explanation) in the AI training phase and five questions with the AI (explanation) in the testing phase.

Unlike previous studies, this study has a COG-ED phase after the testing phase. During the COG-ED experiments, participants are asked to complete 5 questions, during which, they are asked to choose between completing the task in a high-effort prediction-only condition for a fixed financial reward (R_{max} credits, where $R_{max} = 100$ in our experiments) or completing the task in a low-effort explanation condition for a smaller reward (initially R_1 credits, where $R_1 = 50$ in our experiments). The rewards are initialized such that $R_1 < R_{max}$. The high-effort condition, in this case, is opting to complete the task alone. The low-effort condition, in this case, is opting to complete the task with the AI. Each time a participant chooses the low-effort option, the reward for the low-effort option shifts to adjust for the participant's preferences. Specifically, assuming the reward for the low-effort condition is R_t, the new reward is dynamically updated such that:

$$R_{t+1} = \begin{cases} \frac{R_t + R_{max}}{2}, & \text{if high-effort is chosen} \\ \frac{R_t + R_{t-1}}{2}, & \text{otherwise} \end{cases} \quad (1)$$

(a) Effect of task difficulty   (b) Effect of explanation modality
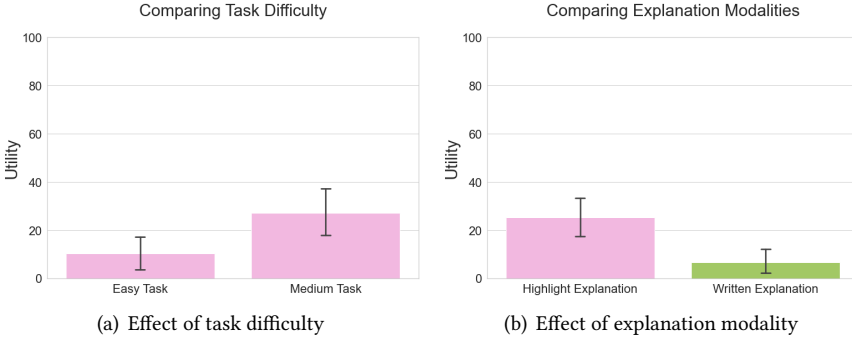
Fig. 14. **(a)** Utility levels in study 5. We find that highlight explanations have a higher subjective utility in the medium task than in the easy task. **(b)** Utility levels in study 5. We find that highlight explanations have a higher subjective utility than written explanations. Taken with the findings of studies 1 and 2, this study shows that people tend to attach higher utility to AI Conditions that generally decrease overreliance. This study validates that people are indeed computing a "cost-benefit" utility and that it is highly predictive of general trends in overreliance.

where $R_0 = 0$. The monetary reward for choosing the high-effort prediction-only condition does not change.

After the participants have completed the five questions in the COG-ED phase, the monetary reward R_6 is the final value on which the participants have converged. $R_{max} - R_6$ represents how much reward a participant is willing to forgo in order to do an easier task. This quantifies the *subjective utility* of the AI in the easier task.

9.5 Participants

We recruited $N = 76$ participants from Profilic, half of which are placed in condition 1, half of which are placed in condition 2, using the same exclusion criteria as in Study 1. Our sample size comes from a power analysis done on pilot data.

Participants receive a base payment of $4 for 20 minutes of their time, the same as Study 1 and Study 2. However, participants are also given opportunities for bonuses in this study when they answer a question correctly. To equalize participant perception of monetary reward and mitigate income effects, we use a credit system, where participants earn "credits" instead of dollar amounts. Participants are informed during the study that 100 credits are equivalent to $0.10 in the medium task and 100 credits are equivalent to $0.05 in the easy task (for participants who see the easy task).

9.6 Results

We analyze our results for H4 and H5 using two Bayesian linear mixed effects models (see Table 6). We visualize these results in Figure 14. We adopt the convention of saying that a comparison is notable if the 95% credible interval of the posterior distribution excludes 0 and the estimate of the mean is positive or negative (depending on the direction of the hypothesis). *Note*: We provide the model specification in the appendix.

9.7 Summary

We find support for H4A, which provides empirical evidence for the following claim: people find AI's explanations to have higher utility in harder tasks. This fits into our cost-benefit framework:

Table 6. Results from study 5. We find that people place a higher utility level on the AI's explanations when the task is harder. We also find that people place a higher utility level on the AI's explanations when it is easier to parse. These results, taken together with the results of study 1 and 2, provide evidence for our cost-benefit framework in overreliance, as people place higher utility on the AI Conditions that generally yield lower levels of overreliance.

Comparison	Mean	Lower 95% CI	Upper 95% CI	Notable	Hypothesis	Confirmed?
Medium task - Easy task	16.96	5.22	28.7	*	H4A	**yes**
Highlight XAI - Written XAI	18.61	9.51	27.79	*	H5A	**yes**

explanations are used with the strategy to engage with the task. In easier conditions, this strategy is about as much work as doing the task alone, since the AI provides little additional help; however, as the task gets harder, this strategy allows people to (1) be more likely to be correct with (2) less cost (via time and cognitive effort). This strategy therefore greatly reduces the cost needed to complete the task properly (compared to doing it alone) in hard tasks.

We find support for H5A, which provides empirical evidence for the following claim: people find AI's explanations to have higher utility when they are easier to understand. This fits into our cost-benefit framework: explanations are used with the strategy to engage with the task. Anyone who engages with this strategy would want to also minimize the cost to do so. Since explanations that are harder to understand have higher costs (in the form of cognitive effort and time) than those that are easier to understand, they have lower subjective utility.

This study demonstrates that people are indeed making a strategic decision to overrly on the AI. Our findings match with the outcomes of the cost-benefit framework, which predicts overreliance on the AI.

10 DISCUSSION

Taken together, our results identify scenarios in which explanations reduce overreliance. In line with prior work examining cost-benefit analyses in the allocation of mental effort [46], we provide evidence that humans engage in cost-benefit analyses when choosing *how* to collaborate with an AI when making decisions.

10.1 Summary of results

Our five studies demonstrate the validity of our cost-benefit framework. We find that explanations reduce overreliance in more difficult tasks. We also find that the *type* of explanation matters: explanations that make the AI's mistake more obvious reduce overreliance more. This supports our "cost" element of the framework: people will engage with the task when elements (e.g. the explanation by an AI) reduce cognitive costs (e.g. in hard tasks or if the explanations are very easy to verify the AI's predictions). We also find that overreliance is reduced when people see a benefit to not overrelying: when people are given a larger monetary benefit to complete the task, they overrely less. Finally, we validate our claim that people strategically choose when to engage with the AI. We do this by calculating the subjective utility and find that, overwhelmingly, people attach higher utility to explanations in harder tasks and that are easier to understand. We summarize our primary findings on overreliance (studies 1-4) in Table 7.

Table 7. Summary of primary findings. We find that there is an overall increase in overreliance as tasks are more difficult; that in easier tasks, overreliance decreases with explanations; that in hard tasks, overreliance decreases with explanations; that given explanations that are easier to parse or that make the AIs mistake more salient, overreliance decreases; and that when the benefit to getting the answer correct is higher, overreliance decreases.

Hypothesis	Task Difficulty	AI Condition	Monetary Bonus	Overreliance
H1A	Easy → Medium	Prediction		↑
H1B	Easy	Prediction → Highlight XAI		same
H1C	Medium	Prediction → Highlight XAI		same
H1D & H3B	Hard	Prediction → Highlight XAI		↓
H2A	Medium	Written XAI → Highlight XAI		↓
H2B	Hard	Written XAI → Highlight XAI		↓
Exploratory	Hard	XAI becomes more salient		↓
H3A	Hard	Prediction & Highlight XAI	Low → High	↓

10.2 Analysis of prior work

Conflicting expectations (of explanations reducing overreliance) and empirical results (of explanations not reducing overreliance) largely inspired this paper. As such, in this section, we aim to address how the findings of prior work fit into our framework. Bansal et al. [4] find that explanations do not decrease overreliance; however, we postulate that the task (sentiment analysis) could have been too easy to complete alone and that the explanation modality (logical explanations) was too complicated to understand when participants were placed in a higher-effort LSAT task. Buçinca et al. [9] find that cognitive forcing functions decrease overreliance compared to simple explanations. We postulate that this could be due to cognitive forcing functions increasing the difficulty of doing the task. [28] find that some explanations do help in error detection. The study's task was in open-domain question-answering, which is a more difficult, open-ended task, where explanations can reduce costs of engaging with the task.

10.3 Implications

Our results find instances in which explanations reduce overreliance over a prediction only baseline. We identify some of the metaphorical knobs to be turned to modulate levels of overreliance: task difficulty, explanation difficulty, and monetary reward. These, of course, were proxies for other elements of human-AI interaction; for example, monetary reward aimed to target increasing benefits, which could also be targeted by increasing the stakes of a task. In the following section, we discuss in more detail how designers can manipulate the task to reduce overreliance.

Manipulating costs: Here, we discuss some design details that can influence overreliance by manipulating the associated costs (via cognitive effort, time, or other factors) of the task:

(1) *Cognitive forcing functions:* As [9] found, these cognitive forcing functions decrease overreliance. We postulate that these forcing functions increase the cost of following the strategy to rely on the AI.

(2) *Applying easy-to-understand explanations to hard tasks:* Our work finds that explanations are especially beneficial when tasks are hard. Tasks that are already difficult may be good places to apply explanations.

(3) *Artificially manipulating tasks to be harder:* As in our experiments, designers can also manipulate the relative costs of engaging with the task by increasing the baseline task difficulty. This outcome may arise naturally as AI performance improves and AI systems tackle more challenging tasks. However, in many settings, increasing the difficulty of the task may not be feasible.

(4) *Creating easier-to-understand explanations:* Our work finds that, if explanations are not easy-to-understand, then they do not provide any assistance over prediction only baselines. Designs should include explanations that are easy to map to the AI's predictions for verification. However, generating human-interpretable explanations continues to be a challenge for XAI [52].

Manipulating Benefits: Here, we discuss some design details that can influence overreliance by manipulating the associated benefits (via reward, stakes, or other factors) of the task:

(1) *Monetary reward:* Although this is not true of all tasks, some tasks, like crowd-sourcing, could benefit from reward systems where people are bonused for their joint work with AIs. Since we found an impact of rewards on overreliance, there may also be a difference in the findings of work that bonuses crowd-workers versus work that does not.

(2) *Stakes:* The benefit of the task may also include professional, personal, or other benefits in the form of stakes. High-stakes domains, for example, may be a case where the benefits of doing tasks properly is really high, thereby discouraging people from relying on the AI.

(3) *Enjoyability:* Tasks that are inherently enjoyable also have high benefit attached to them, as people naturally want to engage with the task [82]. Though not possible for all tasks, increasing the enjoyability of a task may increase peoples' propensity for engaging with the task.

(4) *AI design:* Although increasing benefits of the following may result in increases in overreliance, designers are also able to manipulate (1) perceived model accuracy [95], (2) AI anthropromorphization [50], and (3) the competence associated with metaphors describing the system [44]. For example, anthropromorphizing the AI may lead to an increase sense of trust in the AI's accuracy, which may lead people to falsely assume that such an AI has a greater benefit (via overall accuracy) than it actually does.

10.4 Limitations and Future Work

Task: As previously mentioned, our task was in a very controlled, low-stakes, maze-solving environment. As such, we do not know how much of our work will extend to the following settings (and others):

(1) *High-stakes environments:* Our study focused on a low-stakes maze solving task. Future work should explore our theory on high-stakes tasks such as recidivism prediction [92], where the consequences of a wrong decision are much higher. High stakes tasks may have different levels of costs and benefits, in which case our findings may not generalize. Such an experiment is out of scope for the current paper given the complexity that might arise due to Prospect theory [87]. Prospect theory explains that people value gains and losses differently, placing more weight on perceived gains versus perceived losses. We hypothesize that the perception of loss by making an incorrect decision in high-stakes tasks will modify the costs associated with the task. This in turn might affect overreliance independent of task difficulty and explanation conditions.

(2) *Generative tasks:* Generative tasks often lack a "correct," ground truth answer because of their open-ended nature. Future work remains to see how much the results of this work would translate to more open-ended tasks.

(3) *Non-game-playing tasks:* As noted by a few participants, our task was, at least to some, inherently "enjoyable." As enjoyability may affect how much people want to engage with tasks, future work remains on how this work translates to tasks less gamified.

(4) *Uncertain environments:* Some decision-making tasks, such as toxicity prediction, lack a "correct" answer [29]. In this case, we do not know how our framework in understanding when explanations reduce overreliance would translate, since there is inherent uncertainty in the decision being made.

(5) *Prediction tasks with incomplete explanations*: Many common applications, including common image classification tasks and natural language inference tasks, utilize highlight explanations that identify patches of images or phrases in sentences that were primarily responsible for a prediction. These types of highlight explanations provide incomplete information about a model's underlying behavior, unlike the procedural highlight explanations we used in the maze solving tasks. Future work should explore how well our theory translates to such prediction tasks.

Explanations: Our explanation modalities were primarily limited to those in the form of highlights and written explanations. The following are some crucial limitations to the explanations we chose:

(1) *Explanation quality:* In our work, we specifically chose "perfect" explanations so that explanation quality would not affect our results. However, in practice, AIs do not provide these ideal explanations [20]. Because study two shows that explanation understandability affects levels of overreliance, we hypothesize, but do not empirically validate, that lower explanation quality affects overreliance similarly. However, future work remains to assess the impact of explanation quality on our results, since there may be factors such as a change in trust that may arise when evaluating imperfect explanations.

(2) *Explanation format:* In our work, our explanations specifically helped users understand the AI's reasoning process to the end of the maze. However, some explanations, such as Shapley Additive Explanations [64, 72] or contrastive example-based explanations [88], only provide users a rough understanding to the AI's biases and output. These do not necessarily tell users about the veracity of the AI's prediction.

Users: Our studies were primarily done with crowd-sourced workers. As mentioned in the *Task* section, our study may not be reflective of high-stakes target domains of interest (such as medicine, law, risk assessment, etc.). The following outlines limitations and future work on users studied:

(1) *Different practitioners:* While prior work has established ample evidence of cognitive biases in high-stakes domain [16, 17, 31, 56, 58, 89], practitioners in these domains may have vastly different needs and values, yielding different assessments of costs and benefits. In light of recent calls for sociotechnical approaches to XAI [22, 23, 37], in which the human decision-maker's values and social context are centered when designing explanations, we emphasize the need for co-designing decision support tools with relevant stakeholders and evaluating these tools in real contexts. For instance, a recent study found that clinicians simply did not have time to engage with feature-based explanations for individual patient predictions and instead requested evidence that the tool had been clinically validated [37]. In this example, the time pressures doctors face when treating patients inflated the costs of engaging with an explanation. As such, future work is necessary to test our framework on different practitioners and how different levels of expertise may affect overreliance.

(2) *Long-term users:* Our crowd-sourced study lacks long-term user engagement,. Future work remains to study how long-term users engage with these systems and how overreliance might change over time.

(3) *Need for Cognition:* Our study found mixed results whether a person's propensity for effortful thinking influenced their overreliance. As such, future work is necessary to tease apart the individual factors that contribute to this phenomena.

Interface Design: People moderate their perceptions of the costs and benefits of using an AI using a wide variety of sources. Through pilot studies, we found that small changes in the interface affected the results: including a progress bar, the total number of questions, naming the AI, providing feedback on what answers were correct, and explicitly stating the model's accuracy. We chose the most "neutral" setting with each of these design choices and limited the number of questions asked for each user to avoid fatigue. Future work can explore how changes in these interface considerations affect behavior.

Trust: This paper does not address the causes and effects of AI explanations on trust. Previous work has suggested that trust increases someone's likelihood of using a machine [96, 97]. However, XAI studies found conflicting results about the effect of explanations on trust. Some studies found that explanations increase trust [4, 8], some found no significant effect [15, 42, 97], while others found varying effects under different conditions [49, 51]. Explanations intend to improve performance in part by helping a user calibrate their trust according to a more accurate mental model of the AI's ability and accuracy [4]. The effect of a mental model of AI accuracy is unclear, as some studies found that any accuracy score increases trust [55], others found high accuracy scores increased trust and low accuracy scores decreased trust [96, 97], while another found that these effects are mediated by an individual's innate likelihood of trusting automation [77].

It also remains unclear whether explanation-induced trust is caused by explanations increasing understanding of the underlying system or explanations serving as a blind signal for trust and reliance. A recent study found that placebic, uninformative explanations increased trust in a recommender system nearly as much as real explanations [24], mirroring work in social psychology demonstrating adding a redundant explanation to a request ("May I use the xerox machine, *because I have to make copies*?") increases compliance [57].

Finally, when researching the effects of trust on behavior, trust is difficult to measure accurately. Some studies use reliance as a proxy measure of trust [4, 15, 55, 97], while others subjectively measure an individual's trust in the system and innate disposition to trust [8–10, 42, 51, 77]. It will also be necessary to address how individuals in different contexts and occupations build trust in the systems they use. Although measuring and modulating trust was outside the scope of this paper, we see a rich opportunity for future work further investigating the effects of AI explanations on trust and the effects of trust on a cost-benefit analysis and resulting behavior.

11 CONCLUSION

Prior work in human-AI decision-making found that explanations do not empirically reduce overreliance. Inspired by cost-benefit analysis in behavioral economics, we outline a framework that aims to identify when explanations reduce overreliance. Our studies validate this framework. We find that when we increase the cost of doing the task alone or with just the AI's predictions (via increasing task difficulty), people utilize explanations, which leads to a reduction in overreliance, compared to a prediction only baseline. We also find that when we increase the cost of engaging with the task (via decreasing the understandability of an explanation), people are less likely to utilize the explanation, which leads to an increase in overreliance. We validate the benefit component and

find that, as the benefit of completing the task properly increases (via monetary bonus), overreliance decreases. Through our studies using the Cognitive Effort Discounting paradigm, we quantify the utility of different explanations in different settings. We find that people explicitly forego monetary rewards for an AI in a harder task or for an AI that gives more understandable explanations. We observe that people assign high utility to the conditions that had the biggest reductions in overreliance (compared to the prediction only baseline) in prior experiments. Our experiments suggest that overreliance is not merely an immutable inevitability of cognition but at least, in part, a strategic choice.

ACKNOWLEDGMENTS

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[2] Yuichiro Anzai and Herbert A Simon. 1979. The theory of learning by doing. *Psychological review* 86, 2 (1979), 124.

[3] 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.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. arXiv:2006.14779 [cs.AI]

[5] Bernard Barber. 1983. The logic and limits of trust. (1983).

[6] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M Mack, George Ruiz, Mark S Smith, and Eric Horvitz. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one* 9, 10 (2014), e109264.

[7] Matthew M Botvinick and Jonathan D Cohen. 2014. The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive science* 38, 6 (2014), 1249–1285.

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[9] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

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

[11] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[12] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It Takes Two to Tango: Towards Theory of AI's Mind. arXiv:1704.00717 [cs.CV]

[13] Jim X Chen. 2016. The evolution of computing: AlphaGo. *Computing in Science & Engineering* 18, 4 (2016), 4–7.

[14] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[15] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. arXiv:2007.12248 [cs.LG]

[16] Pat Croskerry. 2009. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in health sciences education* 14, 1 (2009), 27–35.

[17] Pat Croskerry. 2009. A universal model of diagnostic reasoning. *Academic medicine* 84, 8 (2009), 1022–1028.

[18] Gabriel Lins de Holanda Coelho, Paul H. P. Hanel, and Lukas J. Wolf. 2020. The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment* 27, 8 (2020), 1870–1885. https://doi.org/10.1177/1073191118793208 arXiv:https://doi.org/10.1177/1073191118793208 PMID: 30095000.

[19] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[20] Jeff Druce, James Niehaus, Vanessa Moody, David Jensen, and Michael L. Littman. 2021. Brittle AI, Causal Confusion, and Bad Mental Models: Challenges and Successes in the XAI Program. https://doi.org/10.48550/ARXIV.2106.05506

[21] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.

[22] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[23] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[24] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[25] Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences* 7, 10 (2003), 454–459.

[26] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play. arXiv:1810.09648 [cs.AI]

[27] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. 1995. *Bayesian data analysis*. Chapman and Hall/CRC.

[28] Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human Evaluation of Spoken vs. Visual Explanations for Open-Domain QA. arXiv:2012.15075 [cs.CL]

[29] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[30] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[31] Chris Guthrie and Andrew J Wistrich. 2007. Blinking on the Bench: How Judges Decide Cases. *Cornell Law Review* 93 (2007), 1.

[32] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-mediated communication: definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100.

[33] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *arXiv preprint arXiv:2005.01831* (2020).

[34] Carl G. Hempel and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of Science* 15, 2 (1948), 135–175. http://www.jstor.org/stable/185169

[35] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[36] Benjamin D. Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adali. 2019. Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone? arXiv:1904.01531 [cs.CY]

[37] 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.

[38] 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.

[39] Eric J Johnson and John W Payne. 1985. Effort and accuracy in choice. *Management science* 31, 4 (1985), 395–414.

[40] Daniel Kahneman. 2003. A perspective on judgment and choice: mapping bounded rationality. *American psychologist* 58, 9 (2003), 697.

[41] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

[42] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[43] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4521–4532.

[44] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[45] Wouter Kool and Matthew Botvinick. 2014. A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General* 143, 1 (2014), 131.

[46] Wouter Kool and Matthew Botvinick. 2018. Mental labour. *Nature human behaviour* 2, 12 (2018), 899–908.

[47] Wouter Kool, Samuel J Gershman, and Fiery A Cushman. 2017. Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science* 28, 9 (2017), 1321–1333.

[48] Wouter Kool, Joseph T McGuire, Zev B Rosen, and Matthew M Botvinick. 2010. Decision making and the avoidance of cognitive demand. *Journal of experimental psychology: general* 139, 4 (2010), 665.

[49] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[50] Philipp Kulms and Stefan Kopp. 2019. More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation. In *Proceedings of mensch und computer 2019*. 31–42.

[51] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[52] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An Evaluation of the Human-Interpretability of Explanation. https://doi.org/10.48550/ARXIV.1902.00006

[53] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[54] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan 2019). https://doi.org/10.1145/3287560.3287590

[55] 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* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[56] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.

[57] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction. *Journal of personality and social psychology* 36, 6 (1978), 635.

[58] James H. Lebovic. 2014. National Security Through a Cockeyed Lens: How Cognitive Bias Impacts U.S. Foreign Policy by Steve A. Yetiv. Baltimore, MD, Johns Hopkins University Press, 2013. 168 pp. $24.95. *Political Science Quarterly* 129, 3 (2014), 534–536. https://doi.org/10.1002/polq.12235 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/polq.12235

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

[60] NY Louis Lee, Geoffrey P Goodwin, and Philip N Johnson-Laird. 2008. The psychological puzzle of Sudoku. *Thinking & Reasoning* 14, 4 (2008), 342–364.

[61] Russell V. Lenth. 2021. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans R package version 1.6.0.

[62] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[63] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[64] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/ARXIV.1705.07874

[65] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.

[66] David R Mandel and Philip E Tetlock. 2018. Correcting judgment correctives in national security intelligence. *Frontiers in Psychology* 9 (2018), 2640.

[67] Michael Scott McClendon et al. 2001. The complexity and difficulty of a maze. In *Bridges: Mathematical Connections in Art, Music, and Science*. Citeseer, 213–222.

[68] Joseph T McGuire and Matthew M Botvinick. 2010. Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the national academy of sciences* 107, 17 (2010), 7922–7926.

[69] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. arXiv:1706.07269 [cs.AI]

[70] David Navon and Daniel Gopher. 1977. *On the Economy of the Human Processing System: A Model of Multiple Capacity.* Technical Report. TECHNION-ISRAEL INST OF TECH HAIFA FACULTY OF INDUSTRIAL AND MANAGEMENT ….

[71] Dong Nguyen. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* Association for Computational Linguistics, New Orleans, Louisiana, 1069–1078. https://doi.org/10.18653/v1/N18-1097

[72] Yasunobu Nohara, Koutarou Matsumoto, Hidehisa Soejima, and Naoki Nakashima. 2019. Explanation of machine learning models using improved Shapley Additive Explanation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics.* 546–546.

[73] Alonso Ortega and Gorka Navarrete. 2017. Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Significance Testing (NHST) in Psychology and Social Sciences. In *Bayesian Inference*, Javier Prieto Tejedor (Ed.). IntechOpen, Rijeka, Chapter 12. https://doi.org/10.5772/intechopen.70230

[74] 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.

[75] John W Payne, James R Bettman, and Eric J Johnson. 1988. Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition* 14, 3 (1988), 534.

[76] Joseph C Pitt. 1988. Theories of explanation. (1988).

[77] Vlad L Pop, Alex Shrewsbury, and Francis T Durso. 2015. Individual differences in the calibration of trust in automation. *Human factors* 57, 4 (2015), 545–556.

[78] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. arXiv:1802.07810 [cs.AI]

[79] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250 [cs.CL]

[80] Dasha A Sandra and A Ross Otto. 2018. Cognitive capacity limitations and Need for Cognition differentially predict reward-induced cognitive effort expenditure. *Cognition* 172 (2018), 101–106.

[81] Amitai Shenhav, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L Griffiths, Jonathan D Cohen, and Matthew M Botvinick. 2017. Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience* 40 (2017), 99–124.

[82] Paul J. Silvia and Todd B. Kashdan. 2009. Interesting Things and Curious People: Exploration and Engagement as Transient States and Enduring Strengths. *Social and Personality Psychology Compass* 3, 5 (2009), 785–797. https://doi.org/10.1111/j.1751-9004.2009.00210.x arXiv:https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2009.00210.x

[83] Herbert A Simon. 1956. Rational choice and the structure of the environment. *Psychological review* 63, 2 (1956), 129.

[84] Peter Todd and Izak Benbasat. 1994. The influence of decision aids on choice strategies: an experimental analysis of the role of cognitive effort. *Organizational behavior and human decision processes* 60, 1 (1994), 36–74.

[85] Michael Toomim, Travis Kriplean, Claus Pörtner, and James Landay. 2011. Utility of human-computer interactions: Toward a science of preference measurement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2275–2284.

[86] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[87] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty* 5, 4 (1992), 297–323.

[88] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[89] Milica Vasiljevic, Mario Weick, Peter Taylor-Gooby, Dominic Abrams, and Tim Hopthrow. 2013. Reasoning about extreme events: A review of behavioural biases in relation to catastrophe risks. (2013).

[90] Valerie M Vaughn, Vineet Chopra, and Joel D Howell. 2016. War games and diagnostic errors. *BMJ: British Medical Journal* 355 (2016).

[91] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[92] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21).* Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[93] Andrew Westbrook and Todd S Braver. 2015. Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience* 15, 2 (2015), 395–415.

[94] Andrew Westbrook, Daria Kester, and Todd S Braver. 2013. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS one* 8, 7 (2013), e68210.

[95] 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.

[96] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[97] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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* (Barcelona, Spain) *(FAT* *'20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[98] Shoshana Zuboff. 1988. In the age of the smart machine. (1988).

12 APPENDIX

12.1 Self-report Questions

6-Question Trust Survey:

(1) I believe the AI is a competent performer. (on a 7 point Likert scale)
(2) I trust the AI. (on a 7 point Likert scale)
(3) I have confidence in the advice given by the AI. (on a 7 point Likert scale)
(4) I can depend on the AI. (on a 7 point Likert scale)
(5) I can rely on the AI to behave in consistent ways. (on a 7 point Likert scale)
(6) I can rely on the AI to do its best every time I take its advice. (on a 7 point Likert scale)

Task Survey:

(1) I found this task interesting. (on a 7 point Likert scale)
(2) I found this task difficult without the AI. (on a 7 point Likert scale)
(3) I found this task difficult even with the AI. (on a 7 point Likert scale)
(4) I would prefer to complete this task with the AI's suggestions than to complete it by myself. (on a 7 point Likert scale)
(5) Please select the option below which best represents how you used the AI. (4 options, listed below)
  (a) I did not use the AI and completed the task by myself.
  (b) I first completed the task myself and then verified my response with the AI's.
  (c) I first looked at the AI's suggestion and then verified it was correct.
  (d) I always chose the AI's suggestion.
(6) Approximately, how accurate do you think the AI is? Please indicate using the slider below. (on a 100% slider)
(7) In the box below, please describe how you used the AI when its suggestions were given to you. (free form answer box)
(8) In the box below, please describe how you chose between using the AI and doing the task yourself. (free form answer box)

Need for Cognition Survey:

(1) I would prefer complex to simple problems. (on a 5 point Likert scale)
(2) I like to have the responsibility of handling a situation that requires a lot of thinking. (on a 5 point Likert scale)
(3) Thinking is not my idea of fun. (on a 5 point Likert scale , reverse coded)

(4) I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. (on a 5 point Likert scale , reverse coded)

(5) I really enjoy a task that involves coming up with new solutions to problems. (on a 5 point Likert scale)

(6) I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought. (on a 5 point Likert scale)

12.2 Models for each hypothesis

We use Bayesian models for all our analyses. We fit random intercepts of participants and mazes, as opposed to random slopes, as our fixed effects of interest (AI Condition, task difficulty, and explanation condition) do not vary within a group. We do account for random effects such as participant and maze. We ran post-hoc tests using the emmeans package [61] on all models to find our specific effects.

H1A, H1B, H1C: overreliance $\sim 1 +$ Task difficulty $*$ AI Condition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$
H1D: overreliance $\sim 1 +$ AI Condition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$
H1E: overreliance $\sim 1 +$ AI Condition $*$ Need For Cognition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$
H2A: overreliance $\sim 1 +$ Explanation Condition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$
H2B: overreliance $\sim 1 +$ Explanation Condition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$
H3A: overreliance $\sim 1 +$ Bonus Condition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$
H3B: overreliance $\sim 1 +$ AI Condition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$
H4A: utility $\sim 1 +$ Task Difficulty Condition $+ (1 \mid$ participant$)$
H5A: utility $\sim 1 +$ Explanation Condition $+ (1 \mid$ participant$)$
STUDY 3 (NON-PRE-REGISTERED): overreliance $\sim 1 +$ AI Condition $+ (1 \mid$ participant$) + (1 \mid$ maze$)$

12.3 Previous pre-registrations

We previously pre-registered two studies[5] that made similar hypotheses using a popular Wikipedia Question Answering Task [79].

12.3.1 Study Design. We had a similar study design as outlined in our main paper. The main difference is that, instead of the maze task, we use the Wikipedia Question-Answering Task, where our easy condition was one paragraph and our hard condition was five paragraphs. Our explanation condition was in the form of "perfect" inline highlights, where the answer to the model's prediction was highlighted in the text, as common in prior work [4, 26, 53, 55].

12.3.2 Hypotheses and results. Our specific hypotheses can be found in the preregistration links.
Our hypotheses with the prior pre-registration did not appear to have a large effect, so we ended the study early at a number of participants far lower than pre-registered; therefore, we do not report these values.

12.3.3 Understanding the results. We posit that our studies using the Wikipedia Question Answering Task were (1) not suitable for detecting task difficulty differences and (2) that we previously had a conceptual error in our hypotheses. The former may be beause modulating the paragraph length to five paragraphs does not incur any significant extra cognitive cost, as it is easy to skim all five paragraphs to find the relevant key words. We find evidence for this in our qualitative analysis. At least 7/17 ($\sim 41.1\%$) participants in the hard prediction condition wrote in the free-form text box (which asked people something along the lines of, "how did you use the AI?") that they skimmed, scanned, or searched; for the second study, at least 17/57 ($\sim 29.8\%$) participants in the hard prediction condition wrote something about skimming, scanning, or searching. This also

---

[5]osf.io/h9256 and osf.io/5q7r6.

indicated that the hard condition was too easy to do. The conceptual error regarding the hypothesis is due to initially thinking that explanations increase cost in the easy task, where we finally settled on the framing that explanations do not decrease the cost in the easy task. This prompted our switch from frequentist to Bayesian Statistics, where we are able to have "no difference" hypotheses.