When Do XAI Methods Work? A Cost-Benefit Approach to Human-AI Collaboration

HELENA VASCONCELOS, Stanford University, USA
MATTHEW JÖRKE, Stanford University, USA
MADELEINE GRUNDE-MCLAUGHLIN, University of Washington, USA
RANJAY KRISHNA, Stanford University, USA
TOBIAS GERSTENBERG, Stanford University, USA
MICHAEL S. BERNSTEIN, Stanford University, USA

Recent work has explored the surprising result that people often do not improve their decision making when supported by an explainable AI (XAI). A main cause of this result is *overreliance*: people accept the XAI's answer even when the explanation makes clear that the XAI is wrong. Prior literature ties overreliance to cognitive biases or uncalibrated trust, finding that overreliance is extremely challenging to reduce. In this paper, we produce evidence that humans strategically decide when to engage with an XAI. We find that overreliance *can* be reduced, because overreliance arises specifically when the effort required to verify the XAI's answer is high compared to the effort required to complete the task manually. Our study ($N = 340$) manipulates task difficulty, finding that XAI does not reduce overreliance for easier tasks, where the effort required to complete the task without the XAI is low, but does reduce overreliance for difficult tasks, where manual verification is effortful and so verifying the XAI's explanation is less relative effort. These results suggest that the absence of performance improvements in previous XAI studies may be in part due to the XAI not providing a substantial enough effort reduction.

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

Additional Key Words and Phrases: explanation, decision making, cost-benefit analysis

ACM Reference Format:
Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Ranjay Krishna, Tobias Gerstenberg, and Michael S. Bernstein. 2022. When Do XAI Methods Work? A Cost-Benefit Approach to Human-AI Collaboration. In *CHI '22: Conference on Human Factors in Computing Systems, Apr 30–May 6, 2022, New Orleans, LA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Support from explainable artificial intelligence (XAI) offers the potential for high performing human-XAI teams [31]. In fact, it is frequently assumed that human-XAI teams will achieve complementary performance on challenging decision-making tasks: that human-XAI teams will perform better than either humans or XAIs alone [1, 5, 31, 32]. From medical diagnosis [11, 42], to recidivism prediction [2, 24], employee hiring [25], and loan risk assessment [20], XAI decision-making systems have been deployed in the hopes that this complementarity will result in better outcomes.

Manuscript submitted to ACM

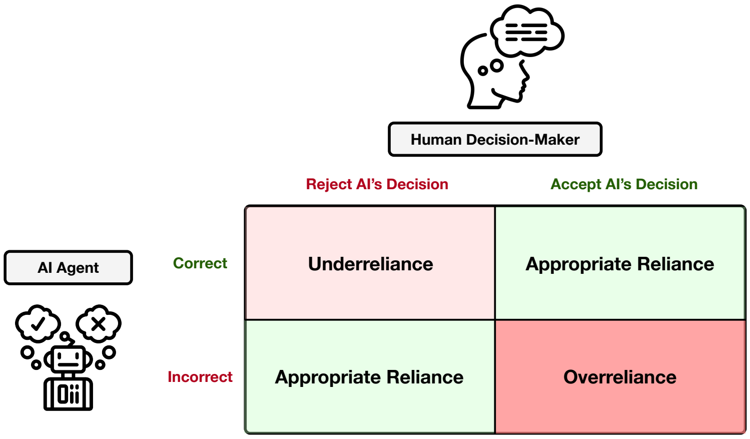


Fig. 1. Types of reliance on AI systems. Provided with a prediction from an AI system, a human decision-maker has the choice to either accept or reject the AI's prediction. Appropriate reliance occurs when the human accepts a correct AI prediction or corrects an incorrect AI prediction. Underreliance occurs when the human fails to accept an correct AI prediction. Overreliance occurs when the human fails to correct an incorrect AI prediction.

Unfortunately, this promised human-XAI team potential has yet to be realized. The only situation where human-XAI teams outperform people alone is when the AI model's accuracy is higher than human-only accuracy [6]; but in these situations, human-XAI teams also perform worse than the XAI alone [9]. A central reason for this divergence between promise and reality is *overreliance*: instead of combining their insights with the XAI's predictions and vetting the XAI's explanation, people accept the XAI's prediction when the XAI is incorrect. Of the possible error types in human-AI decision-making (overreliance and underreliance; see Figure 1), overreliance has been observed most frequently in empirical XAI studies [6, 8, 9, 37, 54]. Overreliance results in foregoing agency and accountability to the AI when making final decisions [10, 26, 38], and is of particular concern in high-stakes domains, running the risk of reinforcing machine bias [2, 20] under the guise of human agency. In principle, explanations should help humans identify when the AI's reasoning is incorrect, and thus reduce overreliance [7, 10, 12]. Despite this, XAI methods do not empirically improve peoples' ability to discern correct from incorrect model predictions [6, 9]. As such, current theories suggest that overreliance is an inevitable result of biases in human cognition, as explanations increase trust [53, 54], anchor humans to the prediction [9, 51], or require cognitive effort to verify [6, 9, 54].

In this paper, we demonstrate that overreliance can be reduced and produce evidence that overreliance arises in situations where the effort required to verify the XAI is high relative to the effort required to complete the task manually. We claim that overreliance is not an immutable inevitability of cognition but a strategic choice where people are responsive to the costs and benefits of engaging in the cognitive effort that is required to reduce their overreliance. This paper tests whether manipulating the costs of an effortful strategy—verifying the AI—has the effect of reducing overreliance, opening an avenue to improving human-AI team performance. Specifically, we hypothesize that the overreliance will decrease in conditions where explanations substantially reduce the cognitive effort, i.e. the costs, incurred to verify the AI's predictions.

We test this cost-benefit framework through an ($N = 340$) online experiment where we measure levels of overreliance as well as utility. Our study modifies the cognitive *costs of the task*, thus modifying the effort required to verify AI predictions. We study overreliance in the context of a maze solving task.

The study results validate our hypothesis. We replicate prior work in finding no differences between levels of overreliance on predictions with and without explanations when the cost conditions are similar to prior studies; we then extend this prior work by manipulating the cost and benefit of using the XAI and find changes to levels of overreliance. In high-effort tasks, explanations significantly reduce the cognitive effort of verifying the AI compared to doing the task without explanations. However, overreliance does not change for low-effort tasks even when explanations are present, since explanations do not induce a significant reduction in cognitive effort for the low-effort task condition. Our results suggest that some of the null effects found in literature could be due in part to the XAI not reducing the costs of verifying the AI's prediction far enough.

2 BACKGROUND

2.1 Human-AI decision making

Empirical HCI studies of human-AI collaborative 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 [6]. Recent studies do not find that explanations improve performance above a prediction-only baseline [6, 13], offering evidence for the following theory: that explanations serve as yet another signal to blindly overrely. In such complementary domains, explanations (even those generated by humans [6]) appear to increase reliance on both correct and incorrect model predictions. Among inputs the model predicts incorrectly, humans achieve worse performance than if they had completed the task by themselves. In light of growing concerns surrounding machine bias [2, 20], 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?*

2.2 Cognitive biases in decision-making

Based on dual process theory [29, 30], a recent study investigates the use of cognitive forcing functions—interventions designed to encourage effortful, analytical thinking—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 in overall performance. The authors also identified an interesting trade-off: participants performed best in the conditions they preferred and trusted the least. We continue this line of work using a cost-benefit framework instead of cognitive forcing functions to encourage effortful and analytical thinking.

2.3 Decision-making in behavioral economics

Human decision-makers are "satisficers", devoting only as much effort as is minimally satisfactory rather than aiming for optimal decision-making outcomes [47]. To frame this aversion to engaging in effortful thinking in economic terms: people weigh the potential benefits of cognitive effort against its perceived costs [34, 44]. 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 [28] by adopting the strategy which maximizes subjective utility as a function of costs and benefits [34]. 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 [49], implying that decision-aid designers can influence cognitive strategy selection by manipulating costs. 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 A COST-BENEFIT FRAMEWORK FOR HUMAN-AI COLLABORATION

Frameworks proposed in prior work–from anchoring biases [9, 51] to humans' innate adversity to cognitive effort [9, 34, 44, 54]–suggest that overreliance is the default and inevitable state of human performance when given an explanation. These frameworks do not account for other findings where people choose cognitively difficult strategies when interacting with AI without cognitive forcing functions [22, 41]. To account for this contradiction in findings, we propose that overreliance is not a default cognitive state, as suggested by previous frameworks in XAI research, but instead the result of an unconscious strategic choice among strategies. In the field of behavioral economics, the cost-benefit framework [34, 44] can help explain which decision making strategies people choose. In this framework, people subconsciously weigh the costs (e.g., cognitive effort) and benefits (e.g., task performance) 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.

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 verifying an AI's predictions (i.e. professional accomplishment or monetary reward) weighed against its inherent costs (i.e. cognitive effort). The cost of cognitive effort includes the difficulty of the task itself. Applying the cost-benefit framework to human-AI decision making, we see that the cognitive effort (cost) required to verify the AI's prediction depends on the difficulty of the underlying task. Therefore, we hypothesize that:

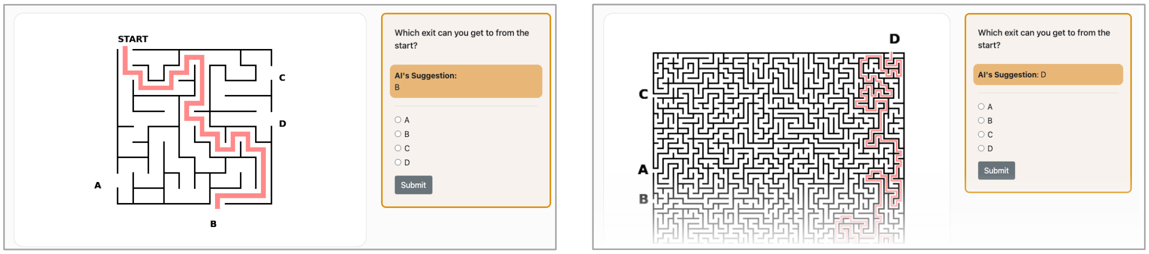
> Increasing the effort required to complete the task reduces overreliance when explanations are present.

4 METHOD

Our research aims to study the effects of explanations on human-AI decision-making. We seek an experimental setup in which participants accomplish a task in collaboration with an AI system, with and without explanations. To study these effects through the cost-benefit framework, we create experimental scaffolds that shift the relative costs of verifying the AI. In the context of this work, costs are the time and cognitive effort required to verify a prediction. We thus developed an experiment to investigate how AI explanations (or lack thereof) affect participants' use of different strategies—and thus overreliance on the AI—across different task difficulties.

4.1 Designing a human-AI collaboration task

In our paper, we use a visual search maze solving task which meets all of our stated needs. 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%. We can manipulate the difficulty of the task by changing the dimensions of the maze. 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, verifying



Which exit can you get to from the start?

AI's Suggestion:
B

○ A
○ B
○ C
○ D

Submit

Which exit can you get to from the start?

AI's Suggestion: D

○ A
○ B
○ C
○ D

Submit

Fig. 2. An example of a maze solving task in which the AI's suggestion is correct. We visualize two different task difficulty conditions. **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. **Right:** The *hard task, prediction* condition. The maze is 50 × 50 and shows explanation highlights.

the prediction is difficult as the participant still must solve the maze to check. Therefore, including explanations reduces the cost of verifying the AI's prediction.

The maze task supports multiple explanation modalities, including highlight explanations, which are easy to check. We visualize examples of our explanations in Figure 3. 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.

4.2 Designing an AI and Explanation

Our study paradigm simulates an AI, with predefined AI predictions for each of our generated mazes. To study overreliance in human-AI decision making, it is important that the AI accuracy is comparable to human accuracy [6]. 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 83.54% across all difficulty conditions, so we control the number of incorrect AI predictions each participant sees such that model accuracy is exactly 80%.

We generate one type of explanation: highlight explanations. (Figure 3). Our highlight explanations indicate the path between the start point and the predicted exist. Highlight explanations have commonly been used in prior XAI studies [6, 18, 23, 37, 45, 54].

4.3 Mitigating effects of trust

As participants complete a series of maze solving tasks, we do not provide immediate feedback since humans are particularly sensitive to seeing algorithms err [17]. Lastly, we do not provide any information about the model's accuracy, as making this explicit also modulates trust [38, 52–54].

4.4 Measures

We collect the following objective performance measures:

- *Overreliance.* We define overreliance as the percentage of answers for which participants select the same answer as the AI among the four questions the AI predicts incorrectly in the collaboration phase.
- *Need for Cognition*: Each participant's Need for Cognition (NFC) score [16] is based on a six question survey scale. Questions 3 and 4 in the scale are reverse coded.

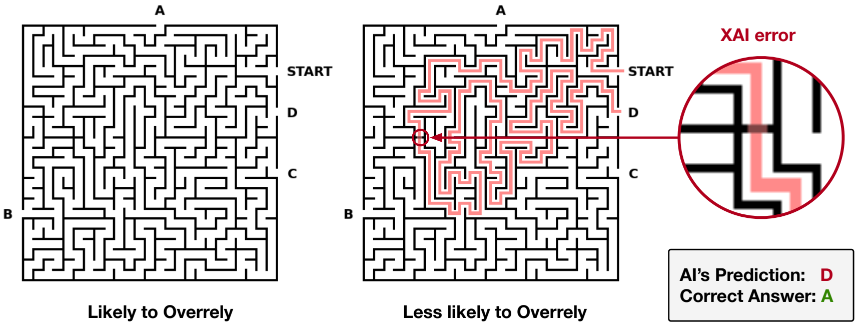


Fig. 3. **Left:** The *medium task, prediction only* condition. The maze is 25 × 25 and features the prediction only. **Right:** The *medium task, explanation* condition. The maze is 25 × 25 and features explanations in the form of inline highlights. An example of when the AI gives an incorrect prediction and explanation.

We additionally collect subjective self-report measures after the testing phase. 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.

5 MAIN STUDY – MANIPULATING TASK COSTS

In our study, we explore our hypothesis (H1) by manipulating how difficulty 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 verify a AI's prediction. 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[1].

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.

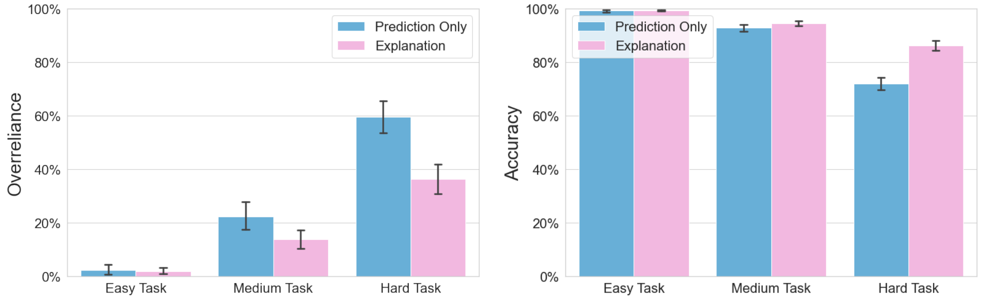
Task difficulty conditions: The task difficulty condition manipulates the maze difficulty. This manipulation modifies cognitive effort because it requires more cognitive effort to search for the correct path in a longer maze [43]. Note that quality of questions, predictions, and explanations remain the 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.

[1] osf.io/vp749



(a) (b)

Fig. 4. **(a)** Overreliance levels in study 1. We find that explanations reduce overreliance in the medium and hard task conditions. In the easy task, we find no differences between prediction and explanation conditions. **(b)** Accuracy levels in study 1. In exploratory analysis, we find that explanations increase decision-making accuracy in the medium and hard task conditions.

(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 Procedure

The study consists of a human-only training phase, in which participants are given a chance to familiarize themselves with the task. Once familiarized, participants are placed in a training phase to acquaint themselves with the AI agent and its explanations (if they are in the explanation condition). The training phase is followed by a testing phase where participants work together with the AI to answer questions. Specific details on the number of trials and feedback given to participants are located in the preregistration.

5.3 Hypotheses

H1 predicts that explanations, when they are present, will reduce overreliance when they reduce cognitive effort to complete a hard task; likewise, explanations will not decrease overreliance when the reduction in cognitive effort is small (e.g., in the easy task). Therefore H1 has 5 sub-hypotheses:

Hypothesis 1a: When given a prediction but no explanation (i.e. in the absence of any explanations), overreliance will be lower in an easy task than a medium difficult task.

Hypothesis 1b: In an easy task, there will be no measurable difference between overreliance whether explanations are present or absent.

Hypothesis 1c: In a medium difficulty task, overreliance will be higher when using a prediction without explanations than using a prediction with an explanation (in the form of highlights).

Hypothesis 1d: In a hard task, overreliance will be higher when using a prediction without explanations than using a prediction with an explanation (in the form of highlights).

Hypothesis 1e: In a hard task, there will be an interaction effect between participants' Need for Cognition (NFC) scores and the type of explanation modality (highlights or lackthereof) when measuring overrreliance.

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) from Profilic (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.

5.5 Results

We analyze our results for Hypothesis 1a, 1b, and 1c using the following Bayesian linear mixed effects model[2]:

$$\text{overreliance} \sim 1 + \text{Task difficulty} * \text{AI condition} + (1 \mid \text{participant}) + (1 \mid \text{maze})$$

Differences	Estimate	Lower 95% CI	Upper 95% CI
Easy Prediction - Medium Prediction	-3.967	-5.480	-2.53
Easy Prediction - Easy Explanation	0.164	-1.678	2.17
Medium Prediction - Medium Explanation	0.916	-0.305	2.08

We do a post-hoc Tukey Test to to test for pairwise differences. We find that overreliance was higher in medium prediction than in easy prediction, supporting H1a. We find that there is no difference between overreliance rates in easy prediction and easy explanation, supporting H1b. We find that medium prediction does not have more overreliance than medium explanation, not supporting H1c. We visualize these results in 4(a).

We analyze our results for Hypothesis 1d using the following Bayesian linear effects model:

$$\text{overreliance} \sim 1 + \text{AI condition} + (1 \mid \text{participant}) + (1 \mid \text{maze})$$

Differences	Estimate	Lower 95% CI	Upper 95% CI
Hard Prediction - Hard Explanation	1.74	0.889	2.76

We find that hard prediction has higher overreliance than hard explanation, supporting H1d.

We analyze our results for Hypothesis 1e using the following Bayesian linear mixed effects model:

$$\text{overreliance} \sim 1 + \text{AI condition} * \text{Need For Cognition} + (1 \mid \text{participant}) + (1 \mid \text{maze})$$

Differences	Estimate	Lower 95% CI	Upper 95% CI
AI Condition:NFC	0.35	-0.76	1.49

We find that there is no interaction effect between NFC and AI Condition in the hard task, finding no support for H1e.

5.6 Summary.

Our results show support for H1a, H1b, and H1d. We find that the easy task, compared to the medium task, has lower levels of overreliance in the prediction condition. This validates our hypothesis H1a, supporting our conjecture that overreliance increases with task difficulty, barring any explanations.

[2] For all studies, we say that our hypothesis is supported if the 95% credible interval of the posterior distribution for a learned coefficient is greater (or less) than 0 (depending on the direction of the hypothesis).

We find that, in the easy task, there is no difference between overreliance in prediction and highlight explanation conditions, validating our hypothesis H1B and replicating the effect found in prior work. This supports our conjecture that explanations do not sufficiently reduce the cognitive cost required to complete a task when the task is easy.

While the difference is more marked in the medium task, the explanation does not cross the threshold to refuting the null hypothesis. Therefore, H1C was not validated. We posit that this occurred because the medium task difficulty is not sufficiently difficult enough to elicit differences. This postulation is supported by the finding that, in the hard task, there exist differences between overreliance in prediction and highlight explanations. As such, H1D was validated.

In the hard task, our results do not show an interaction effect between NFC and AI Condition, so H1E was not validated. This could be because the task is too difficult to demonstrate differences in behavior across peoples' NFC, since most people are likely to overrely anyway. In an exploratory analysis, we find a interaction between AI condition and Need for Cognition (NFC) in the medium task. 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. In the hard task, explanations increase the accuracy compared to the prediction only condition (see Figure 4(b)).

In summary, explanations did not produce an observable reduction in overreliance in easy or medium tasks, but did reduce overreliance in the hard task. This suggests that tasks may need to be substantially complex and effortful to yield benefits with XAIs.

6 DISCUSSION

In line with prior work examining cost-benefit analyses in the allocation of mental effort [34], our results provide evidence that humans engage in cost-benefit analyses when choosing *how* to collaborate with an AI when making decisions. Prior work found that explanations, which reduce the cost of verifying an AI's prediction, did not prompt changes in overreliance. We replicate this result for the easy task condition, and we extend it by demonstrating that explanations do reduce overreliance in a hard task condition, in which the utility of verifying the AI significantly increases when people have access to explanations.

6.1 Analysis of Prior Work

As we have studied the effect of task difficulty and explanation difficulty, we can dissect the study designs and postulate the reason for mixed results in some prior work. For example, one study's authors find that explanations do not decrease overreliance [6]. This could be due to the task being sentiment analysis (low-effort) or the explanation modality being complicated to understand (logical explanations) in a higher-effort LSAT task. On the other hand, another study's authors find that cognitive forcing functions decrease overreliance compared to simple explanations [9]. This could be due to cognitive forcing functions increasing the difficulty of doing the task.

6.2 Implications

Our results demonstrate one of the first approaches that successfully uses explanations to reduce overreliance and thus increase human-AI performance [19]. The results imply not that XAIs contain the information needed for effective human-AI collaboration, but that in many contexts, they do not increase the subjective utility enough to prompt the decision-maker to override a strategy of overreliance. In this section, we discuss how designers can manipulate costs to increase the utility of verifying the AI.

Manipulating costs. One clear design implication is that explanations should strive to reduce the cognitive effort of verifying a model's predictions as much as possible. Our results suggest that a major reason prior work did not observe reductions in overreliance was because this reduction in cost was insufficient to sway users' strategies. If this reduction is not possible or if the nature of the task is already effortless, XAI techniques may continue to trigger overreliance. As in our experiments, designers can also manipulate the relative costs of verifying the AI 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.

Strategies for decreasing overreliance. Different users subjectively place different weights on costs and benefits, causing them to adopt different strategies. Given our observed differences among individual users both in our studies and in prior work [9], future work may also consider explanation strategies that adapt to an individuals' perception of costs and benefits.

6.3 Limitations and future work

Explanation Format. One crucial limitation of our work is the evaluation of explanation utilization only in the domain of maze-solving tasks. Future work is needed to investigate the extent to which our findings generalize to different task domains and explanation modalities. We also evaluated our study with simulated explanations and further work is needed to assess the impact of explanation quality on our results. Recent work has cast doubt on the explanatory power of attention weights in natural language processing models [27] or the analogous saliency maps in computer vision [3, 4, 48], suggesting that imperfect explanation modalities ought to be tested in themselves.

User needs and values. Our study was evaluated with crowdworkers in a low-stakes setting, and thus may not be reflective of high-stakes target domains of interest (such as medicine, law, risk assessment, etc.). While prior work has established ample evidence of cognitive biases in high-stakes domain [14, 15, 21, 39, 40, 50], practitioners in these domains may have vastly different needs and values, yielding different assessments of costs and benefits.

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 [53, 54]. However, XAI studies found conflicting results about the effect of explanations on trust. Some studies found that explanations increase trust [6, 8], some found no significant effect [13, 33, 54], while others found varying effects under different conditions [35, 36].

7 CONCLUSION

In human-AI decision-making, prior literature finds that XAI methods do not empirically improve peoples' ability to discern correct from incorrect model predictions. By studying human overreliance on an AI's prediction through the lens of a cost-benefit framework, we identify factors when explanations reduce overreliance. When a task is difficult and explanations are easy to use to verify an AI, people are less likely to overrely. We find that people are willing to forego a portion of their monetary reward to attain explanations when these conditions are met. Our experiments suggest that overreliance is not an immutable inevitability of cognition but a strategic choice where people are responsive to the costs and benefits of engaging in the cognitive effort that is required to reduce their overreliance.

REFERENCES

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

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There’s software across the country to predict future criminals and it’s biased against blacks. (2016).

[3] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. 2020. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv preprint arXiv:2008.02766* (2020).

[4] Akanksha Atrey, Kaleigh Clary, and David Jensen. 2019. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv preprint arXiv:1912.05743* (2019).

[5] 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.

[6] 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]

[7] 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.

[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] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[12] 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.

[13] 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]

[14] 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.

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

[16] 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.

[17] 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.

[18] 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]

[19] 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]

[20] 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.

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

[22] 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.

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

[24] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI become reliable source to support human decision making in a court scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 195–198.

[25] Rebecca Heilweil. 2019. Artificial intelligence will help determine if you get your next job. https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen

[26] 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.

[27] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

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

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

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

[31] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.

[32] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing.. In *AAMAS*, Vol. 12. 467–474.

[33] 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

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

[35] 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.

[36] 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

[37] 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

[38] 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

[39] 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.

[40] 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

[41] 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.

[42] 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.

[43] 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.

[44] 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 ....

[45] 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

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

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

[48] Suraj Srinivas and François Fleuret. 2020. Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability. *arXiv preprint arXiv:2006.09128* (2020).

[49] 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.

[50] 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).

[51] 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

[52] 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.

[53] 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

[54] 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

A APPENDIX

The following is the contents of the appendix: (1) Self-report Questions, (2) Frequentist Analysis, (3) Summary of Previous pre-registrations, and (4) Previous Pre-registrations.

A.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)
(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)
(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)

A.2 Frequentist Statistics Analysis

As frequentist statistics are more common in HCI literature, we add the frequentist analysis of our hypotheses here.

Study 1. We analyze our results for H1A, H1B, and H1C using a linear mixed effects model with a logistic linking function. We include overreliance as the dependent variable; We add fixed effects of AI condition, task difficulty, as well as the interaction between the two. We include participant and maze as two random effects. As a criterion for statistical significance, we adopt an alpha value of 0.05 (two-sided). Specifically, the model specification was as follows:

$$\text{overreliance} \sim 1 + \text{Task difficulty} * \text{AI condition} + (1 \mid \text{participant}) + (1 \mid \text{maze})$$

We do a post-hoc Tukey Test to validate the differences in pairs. We find that hard prediction has more overreliance than easy prediction ($p < 0.0001$). We find that there is no significant difference between overreliance rates in easy prediction and easy XAI ($p = 0.9999$). We find that hard prediction does not have significantly more overreliance than hard XAI ($p = 0.6985$). Coefficients for our model are as follows: Intercept (-5.81), Hard Task (4.40701), Explanation Condition (0.07381), Interaction (-0.93186).

We analyze our results for H1D and H1E using a linear effects model with a logistic linking function. We include overreliance as the dependent variable; We add a fixed effect of AI condition. We include participant and maze as two random effect. As a criterion for statistical significance, we adopt an alpha value of 0.05 (two-sided). Specifically, the model specification was as follows:

$$\text{overreliance} \sim 1 + \text{AI condition} + (1 \mid \text{participant}) + (1 \mid \text{maze})$$

We analyze our results using a linear mixed effects model with a logistic linking function. We include overreliance as the dependent variable; We add two fixed effects of AI condition and Need for Cognition (NFC) and an interaction between the two. We include participant and maze as two random effect. As a criterion for statistical significance, we adopt an alpha value of 0.05 (two-sided). Specifically, the model specification was as follows:

$$\text{overreliance} \sim 1 + \text{AI condition} * \text{Need For Cognition} + (1 \mid \text{participant}) + (1 \mid \text{maze})$$

We find that prediction has more overreliance than highlight explanation ($p < 0.0001$). We find that there is not a significant interaction between NFC and AI condition ($p = 0.4713$). Coefficients for our first model are as follows: Intercept (0.6861) and Highlight Explanations (-1.6785). Coefficients for our second model are as follows: Intercept (3.0483), Highlight Explanations (-2.9422), NFC (-0.6644), and Interaction (0.3536).

Study 2. We analyze our results using a linear model with a logistic linking function. We include overreliance as the dependent variable; We add fixed effects of the explanation condition. We include participant and maze as two random effects. As a criterion for statistical significance, we adopt an alpha value of 0.05 (two-sided). Specifically, the model specification was as follows:

$$\text{overreliance} \sim 1 + \text{Explanation Condition} + (1 \mid \text{participant}) + (1 \mid \text{maze})$$

We do a post-hoc Tukey Test to validate the differences in pairs. We find that written explanations have more overreliance than highlight explanations ($p = 0.0027$). Coefficients for our model are as follows: Intercept (-3.4346) and Written Explanations (1.6813).

Study 3. We analyze our results using a linear model. We include utility as the dependent variable. In the first condition, we add fixed effects of the explanation. In the second condition, we add fixed effects of the task difficulty. We include participant as a random effect. As a criterion for statistical significance, we adopt an alpha value of 0.05 (two-sided). Specifically, the model specification was as follows for the first condition:

$$\text{utility} \sim 1 + \text{Explanation Condition} + (1 \mid \text{participant}))$$

Specifically, the model specification was as follows for the second condition:

$$\text{utility} \sim 1 + \text{Task Difficulty Condition} + (1 \mid \text{participant}))$$

We do a post-hoc Tukey Test to validate the differences in pairs. We find that highlight explanations have more utility than written explanations ($p < .0001$). Coefficients for our model are as follows: Intercept (25.42) and Written Explanation (-18.68).

We find that highlight explanations have a higher utility in the hard task than in the easy task ($p = 0.0070$). Coefficients for our model are as follows: Intercept (27.13) and Easy Task (-16.82).

A.3 Discussion of Previous Pre-registrations

We previously pre-registered two studies that made similar hypotheses using a popular Wikipedia Question Answering Task [46]. We did not validate our hypotheses with these studies.

Since the primary differences between these studies and our final study lie in the task and statistical analysis, we posit that the Wikipedia Question Answering Task is not suitable for detecting task difficulty differences and posit that we previously had a conceptual error in our hypotheses. The former may be due to the fact that modulating the paragraph length does not incur any significant extra cognitive cost, as it may be easy to skim to find the relevant key words. The conceptual error 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.

A.4 Previous Pre-registrations

We previously pre-registered two studies[3].

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 [6, 18, 37, 38].

Hypotheses. Our specific hypotheses can be found in the preregistration links.

Results. None of our hypotheses in this section were approaching statistical significance, so we ended the study early at a number of participants lower than pre-registered; therefore, we do not report these values.

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