Background Explanations Reduce Users' Over-reliance on AI: A Case Study on Multi-Hop Question Answering

NAVITA GOYAL, University of Maryland, USA
ELEFTHERIA BRIAKOU, University of Maryland, USA
AMANDA LIU, University of Maryland, USA
CONNOR BAUMLER, University of Maryland, USA
CLAIRE BONIAL, U.S. Army Research Lab, USA
JEFFREY MICHER, U.S. Army Research Lab, USA
CLARE R. VOSS, U.S. Army Research Lab, USA
MARINE CARPUAT, University of Maryland, USA
HAL DAUMÉ III, University of Maryland & Microsoft Research, USA

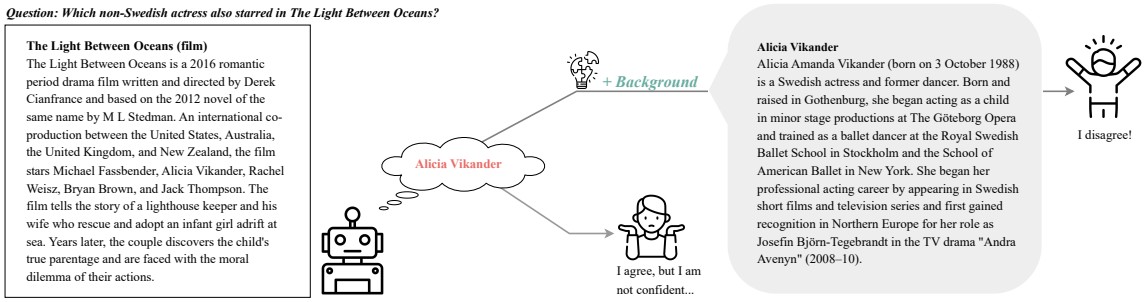

Fig. 1. User's over-reliance on AI predictions is reduced when provided with **background** explanations.

AI systems are becoming increasingly large, infused with a massive amount of prior "knowledge", enabling them to reason about information that is external to their inputs. In such cases, users may be unable to assess the correctness of AI predictions without additional background. In this work, we study how background explanations that provide external information affect users' reliance on AI predictions in a question-answering setting. Our study reveals that users rely on AI predictions even in the absence of sufficient information needed to assess its correctness. Background explanations help users do a better job of spotting AI errors, reducing over-reliance on incorrect AI predictions. However, background explanations also increase users' confidence in their correct as well as incorrect judgments. Highlight-based explanations, on the other hand, do not affect reliance or confidence.

Authors' addresses: Navita Goyal, navita@umd.edu, University of Maryland, College Park, MD, USA; Eleftheria Briakou, ebriakou@umd.edu, University of Maryland, USA; Amanda Liu, amandastephanieliu@gmail.com, University of Maryland, USA; Connor Baumler, baumler@umd.edu, University of Maryland, USA; Claire Bonial, claire.n.bonial.civ@army.mil, U.S. Army Research Lab, Adelphi, MD, USA; Jeffrey Micher, jeffrey.c.micher.civ@army.mil, U.S. Army Research Lab, USA; Clare R. Voss, clare.r.voss.civ@army.mil, U.S. Army Research Lab, USA; Marine Carpuat, marine@umd.edu, University of Maryland, USA; Hal Daumé III, me@hal3.name, University of Maryland & Microsoft Research, USA.

Additional Key Words and Phrases: explanations, artificial intelligence, over-reliance, human-AI collaboration

1 INTRODUCTION

Despite the promise of AI systems to assist humans in decision-making tasks [15, 16], recent studies show that human-AI collaboration exhibits a persistent failure mode: instead of using their own insights and reasoning to perform a task [8], humans often rather over-rely on AI assistance by accepting incorrect suggestions [2, 6, 13, 17]. With the continued use of AI systems in high-stake domains [6, 13, 17], over-reliance on AI poses an important concern that recent work attempts to address through explainable methods. Therein, the hypothesis is that providing humans with explanations of an AI prediction will help them better assess its correctness and thus improve their appropriate reliance upon it. Yet, most human-centered evaluations of current explainability methods show that explanations can adversely increase "blind trust" in the AI model, exacerbating over-reliance [2, 4, 5, 13, 21, 27, 31], especially amongst non-experts [25, 26].

With the advent of large language models pre-trained on massive data, AI systems are able to reason about information that is external to their input, based on their implicit factual or commonsense knowledge [14, 22] or by employing shortcuts to get to the correct answer [7, 20, 28]. However, in such incomplete information settings, users may lack important background knowledge required to assess the correctness of AI predictions: consider the example of Figure 1. Here, a user interacts with an AI model to seek information about which non-Swedish actress starred in the movie "Light Between Oceans," with the system producing *Alicia Vikander* as an answer. However, because the document that the AI system draws the answer from does not mention Vikander's nationality, in order to assess the accuracy of this prediction, the user must already know that Vikander is Swedish and so the answer provided is not possible. Despite such knowledge gaps in scenarios where humans and AI systems operate on insufficient information, much work on explainable AI seeks to justify an AI prediction based only on signals explicitly present in its input [1, 19, 24].

In this work, we study how users interact with AI systems when some of the information required to understand the correctness of the AI prediction is missing. We introduce the use of *background explanations*—explanations that provide information that is external to an AI input—to help humans understand the correctness of the AI prediction. We hypothesize that this alternative type of explanation addresses user's reliance on AI more directly, as its goal is not to *justify* the AI prediction but rather to provide shared context by establishing a common ground between humans and the AI model, a prerequisite for effective communication [9].

We design an explainable AI experiment to study how background explanations impact several measures of a user's perception of the AI model, such as their reliance, confidence, and accuracy. We study this in question-answering setup, as users often engage in question-answering with AI-infused search engines or conversational systems for their information needs. We build an AI system that predicts the answer to a complex question based on an input that is insufficient alone to answer the question without additional background knowledge. Using this system, we design a user-study where we test how the user's agreement with the AI system changes across different conditions, including both traditional highlight-based explanations that focus exclusively on the AI system's input document, as well as background explanations that provide external information. Our findings are summarized as follows:

- Confirming previous findings, users over-rely on AI predictions in the absence of background explanations.
- Background explanations help users better identify incorrect AI predictions, thus significantly reducing over-reliance on AI.
- Background explanations significantly increase users' confidence in their own judgments, both when they are correct and when they are incorrect.

- Highlight-based explanations do not have a significant effect on users' reliance on the AI system or their confidence in their judgments.

2 METHODS

We conduct an IRB-approved[1] online user-study based on a quasi-simulated explainable question-answering system to examine how users' reliance and confidence on AI predictions change based on explanations. We start with outlining the user-study task description (§2.1), provide details on the explainable AI experiment (§2.2), the user-study conditions (§2.3), and procedures (§2.4), and conclude with the measures (§2.5) used to analyze users' perception of the AI across conditions.

2.1 Task Description

Participants are presented with a question, along with a Wikipedia context paragraph and an AI-predicted answer corresponding to a span of text in the context paragraph. Depending upon the condition that they are assigned to, participants may additionally be shown relevant background information, with or without highlights (see §2.3 for the complete set of conditions). For each question, participants have to indicate: (1) whether they agree or disagree with the AI-predicted answer and (2) their level of confidence in their judgment, based on a 5-point Likert scale. Additionally, participants are asked to complete an aggregate survey on their perception of the AI model (see §2.4).

2.2 The Quasi-Simulated Explainable AI Model

We design an experiment based on a quasi-simulated explainable AI question-answering system where the (answer) predictions that the participants see are genuinely from an AI system, but where the explanations provided are synthetic, based on ground-truth annotations from the dataset we use. Using ground-truth explanations lets us explore the full potential of correct, human-provided explanations in this paper, in advance of future research studying confounding factors introduced by AI-predicted and, thus, noisy explanations. We simulate a setting where an AI system predicts an answer given an input that is not sufficient to answer the question in the absence of additional background knowledge. To that end, we use question-answering dataset requiring multi-hop reasoning, i.e., performing multiple inference steps, to identify the correct answer. For instance, answering the question "Which country got independence when World War II ended?" is based on the assumed background knowledge that World War II ended in 1945. We repurpose HotPotQA dataset [30], a popular machine reading comprehension dataset consisting of questions that require gathering information from different parts of Wikipedia articles to be answered.

Real AI System Predictions. We choose a question-answering system with the following desiderata: we want a system that is (1) a realistic representative of state-of-the-art question-answering systems and (2) can make correct predictions based on its implicit background knowledge (as acquired from pre-training data), even when given questions and accompanying context paragraphs without sufficient information needed to answer the questions. Based on this criteria, we select a 336M parameter BERT model [10] fine-tuned on the Stanford Question Answering Dataset (SQuAD) [23], which consists of extractive question-answering pairs over Wikipedia context paragraphs. We employ a single-hop question-answering model, as opposed to a model fine-tuned on multi-hop QA task, as the underlying assumption is that the model is able to answer the complex question, even without the background information. The BERT model fine-tuned on SQuAD achieves an accuracy of $\sim$ 76% on the development set of HotPotQA both in the presence

[1]IRB number 1941629 − 1.

and absence of complete input context that is, in principle, needed to answer the question. For comparable model performance across conditions in our study, we only select questions from HotPotQA where the AI model's answer is the same across the two aforementioned conditions. We conjecture that the model performs well in the absence of complete input context using heuristic shortcuts [7, 20, 28] or implicit knowledge acquired throughout pre-training [14, 22].

Simulated Explanations. Each question in HotPotQA is associated with two or more relevant context paragraphs: the context paragraph containing the gold-standard answer to the complex question and other intermediate paragraphs providing background context corresponding to the required reasoning steps implicit in the question. Moreover, each paragraph has annotations indicating which of its sentences are essential *supporting facts* for answering the question. We construct two types of *simulated explanations* for each question and its AI predicted answer: (1) a *background* explanation corresponding to an intermediate paragraph as defined by HotPotQA and (2) *highlight-based* explanations corresponding to the provided supporting facts, again, defined by HotPotQA.

2.3 Conditions

We designed four conditions that varied the simulated explanations provided to the participants:

(1) No background and no highlights;
(2) No background and supporting facts highlighted;
(3) With background and no highlights;
(4) With background and supporting facts highlighted.

We conduct a between-subjects study with participants randomly assigned to one of these four conditions. Across conditions, the participants are provided with the context paragraph containing the gold-standard answer, as well as the AI-predicted answer. Whenever applicable, we highlight the supporting facts in both the context paragraph containing the gold-standard answer and the additional paragraph providing the background explanation.

2.4 Procedures

We conduct our user-study online on Prolific.[2] We first present participants with details about the study and obtain their consent to participate. We then show a tutorial introducing relevant terminology (e.g., background, highlights), along with instructions on how to perform the task. We assign each participant to one of the four conditions randomly and present them with 10 questions in the same condition, as detailed in §2.3. We show each participant 7 correct and 3 incorrect predictions to ensure that model accuracy on the observed examples is close to the true model accuracy ($\sim$ 76%). We hold this distribution fixed so as to avoid any effect of the observed accuracy on participants' trust in the system. We consider the same pool of questions across conditions and only sample each question once per condition.

We include two attention-check questions asking participants to indicate their answer to the previous question (one attention-check concerning their agreement and one their confidence level) after they have clicked away. We discard responses from participants who fail both attention checks. After completing the ten questions, participants are asked to complete a questionnaire aimed at assessing their perception of the utility of various explanation types, if applicable, their confidence in the AI model, and their self-confidence in their own responses. Participants are also asked to provide optional free-text feedback or comment on the study. Finally, we collect optional demographic information, such as age and gender. The study has been approved by University of Maryland Institutional Review Board.

[2] http://prolific.co

Participants. We recruited a total of 100 participants, 25 per condition. The study took a median time of about 9 minutes to complete. Each participant was allowed to complete the study only once. Out of 96 participants who completed the survey, we discard the responses from participants that fail both attention checks, leaving us with 95 responses. All participants were compensated at a rate of US$15 per hour. 41% of the participants self-identified as women, 58% as men, and 0% in any other gender identity. 25% of participants were between the ages of 18-25, 44% between 25-40, 25% between 40-60 and 6% over the age of 60.

2.5 Measures

We measure several aspects of users' behavior during the study, including their agreement with both correct and incorrect AI predictions, and their confidence in their own judgments. In what follows, we refer to the predictions made by the AI system as "predictions" and the decisions by the users to agree or disagree with that prediction as "judgments". Users' judgments are subsequently deemed "correct" when they agree with correct AI predictions or disagree with incorrect AI predictions and "incorrect" otherwise.

- *Users' accuracy:* Percentage of agreement with correct AI predictions and disagreement with incorrect AI predictions.
- *Appropriate Reliance:* Percentage of agreement with correct AI predictions.
- *Over-reliance:* Percentage of agreement with incorrect AI predictions.
- *Users' confidence:* Average confidence (on a scale of 1-5) in correct or incorrect judgments.

Additionally, in the post-task survey, we collect the following self-reported subjective measures to study users' overall perception of the AI system [11]. For each measure, users are shown the corresponding statement and asked to indicate their level of agreement on a five-point Likert scale from strongly disagree to strongly agree:

- *Usefulness of Highlights:* "The highlights were useful. I feel that highlights helped in determining whether the AI predictions were correct".
- *Usefulness of Background:* "The background information was useful. I feel that background helped in determining whether the AI predictions were correct".
- *Confidence in AI:* "I am confident in the AI system, including the predictions, highlights and background".
- *Self-confidence:* "I am confident in my decisions".
- *Satisfaction with AI:* "I would like to use the AI system for decision making".

To avoid multiple testing effect, we perform Benjamini-Hochberg correction [3] and report "significance" with a false discovery rate (FDR) of 0.05 [12], resulting in a significance threshold of $p < 0.01$.

3 RESULTS

RQ1: Do background explanations affect users' reliance on AI?. As a measure of reliance, we study users' rate of agreement with AI predicted answers. As seen in Figure 2, we find that background explanations help combat over-reliance on incorrect predictions; users exhibit a significantly lower rate of agreement on incorrect predictions in the *with background explanation* condition (0.61 ± 0.04) than the *without background explanation* condition (0.47 ± 0.04). Background explanations do not affect appropriate reliance; that is, the rate of agreement on correct AI predictions is the same with (0.88 ± 0.02) and without background (0.88 ± 0.02) explanations. Even when predictions are correct, high reliance without background explanation is concerning as it indicates that users blindly trust AI despite insufficient information. We note an example of this behavior in a user's feedback at the end of the study: *"Some questions asked for*

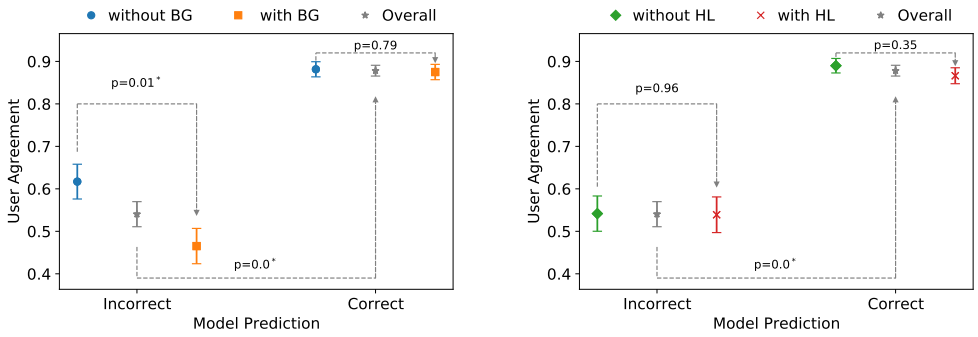

Fig. 2. User agreement with correct vs incorrect model predictions. The graphs show aggregate agreement in *with/without background* conditions (left) and *with/without highlight* conditions (right). The overall agreement is higher for correct predictions than incorrect predictions across conditions. Background explanations (left) reduce over-reliance on incorrect model predictions, with no effect on reliance on correct predictions. Highlight-based explanations (right), on the other hand, do not affect agreement. ∗ indicates significance after Benjamini-Hochberg multiple-testing correction with a false discovery rate of 0.05.

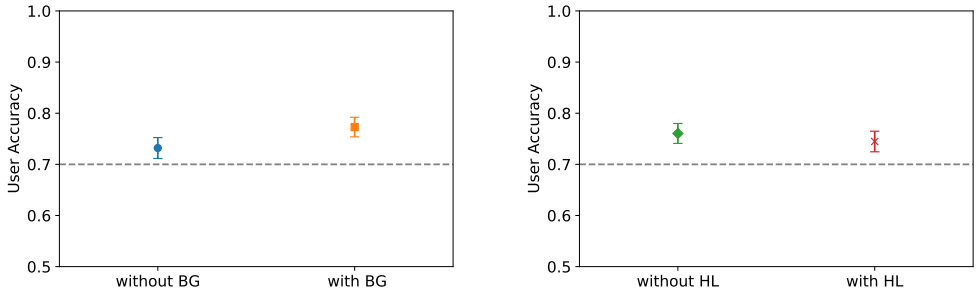

Fig. 3. User accuracy measured as agreement with correct AI prediction and disagreement with incorrect AI prediction. The black line indicates the baseline sample accuracy—if users agree with all predictions, then they will have an accuracy of 70%. User accuracy is marginally higher with background explanations (77%) than without (73%) (left)and marginally lower with highlights (76%) than without (74%). These differences are non-significant (perhaps partly because only 30% of the examples a user sees are incorrect).

2 things, like the type of game for 2 games, but the article only has one game info in there. However, I made my decision based on what information I can get from the article, ***I think most of the time, AI made the right prediction so I chose "certain" about the AI's decision****."*

On the whole, the users are able to adjust their reliance based on the correctness of the AI prediction regardless of whether they are shown any background explanations: users rely on incorrect AI predictions significantly less (0.54 ± 0.03) than correct AI predictions (0.88 ± 0.01). However, the agreement with incorrect AI predictions is substantial nonetheless (over 50%), indicating over-reliance on AI.

RQ2: Do background explanations affect users' ability to detect correct vs. incorrect AI predictions? Comparing the user accuracy in *with and without background explanation* conditions in Figure 3 (left), we find that the user accuracy in detecting correct vs incorrect AI predictions is marginally higher with background explanations (0.77 ± 0.02) than without (0.73 ± 0.02). This is a natural extension of RQ1 as we observe a close agreement rate for correct predictions but a much lower agreement rate for incorrect predictions, in the *with background explanation* condition, which results

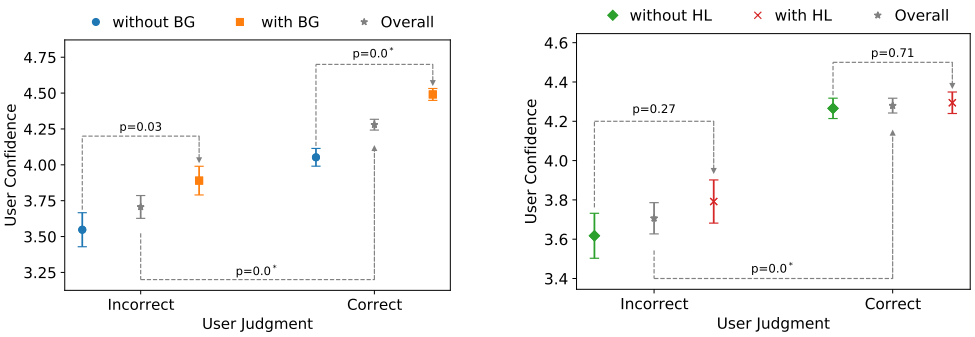

Fig. 4. User confidence in their incorrect vs correct judgments (recorded on a scale of 1 (Very Uncertain)-5 (Very Certain). Overall, users are more confident in their correct judgments than their incorrect judgments, regardless of presence of background or highlight explanations. Background explanations (left) help calibrate trust in correct judgments (more confidence in correct judgments with background than without), but background explanations also lead to overconfidence in incorrect judgments (more confidence in incorrect judgments with background than without). On the other hand, highlight-based explanations do not increase confidence in correct or incorrect judgments. $*$ indicates significance after Benjamini-Hochberg multiple-testing correction with a false discovery rate of 0.05.

in an overall higher accuracy. However, this effect is not significant, perhaps partly because only 30% of the examples a user sees are incorrect.

RQ3: Do background explanations affect users' confidence in decision making? We study whether background explanations help users calibrate their confidence in their judgments: a lower confidence when users' judgments are incorrect and a higher confidence when users' judgments are correct. As seen in Figure 4 (left), we observe that although background explanations improve users' confidence in correct judgments, these explanations also increase over-confidence in incorrect judgments. That is, users exhibit a higher confidence in their incorrect judgments with the background explanations (3.89 ± 0.10) than without (3.55 ± 0.12) with a p-value of 0.03. This reflects that although background explanations help calibrate trust in correct judgments, they lead to overconfidence in incorrect judgments.

On the whole, users' confidence in their judgments is fairly calibrated: users are significantly more confident in their correct judgments (4.28 ± 0.04) than their incorrect judgments (3.71 ± 0.08) (Figure 4 (left)). This is true regardless of background explanations. However, the jump in confidence between incorrect and correct judgments is higher with background explanations (Cohen's d: 0.70) than without (Cohen's d: 0.42).

RQ4: Do highlight-based explanations affect users' reliance on AI predictions and confidence in their judgments? Contrary to background explanations (RQ1), highlight-based explanations do not reduce over-reliance on incorrect AI predictions (Figure 2 (right)). Further, in line with user accuracy of detecting correct vs incorrect AI predictions with and without background explanations (RQ2), the difference in user accuracy with and without highlights is non-significant (Figure 3 (right)). In contrast to background explanations, users have marginally higher accuracy without highlights (0.76 ± 0.02) than with highlights (0.74 ± 0.02).

Lastly, we find that although users are more confident in their incorrect judgments in the presence of highlights (3.79 ± 0.11) than without (3.62 ± 0.11) (Figure 4 (right)), this difference is not significant. This indicates that, in contrast to background explanations, highlight-based explanations do not significantly affect users' overconfidence in their

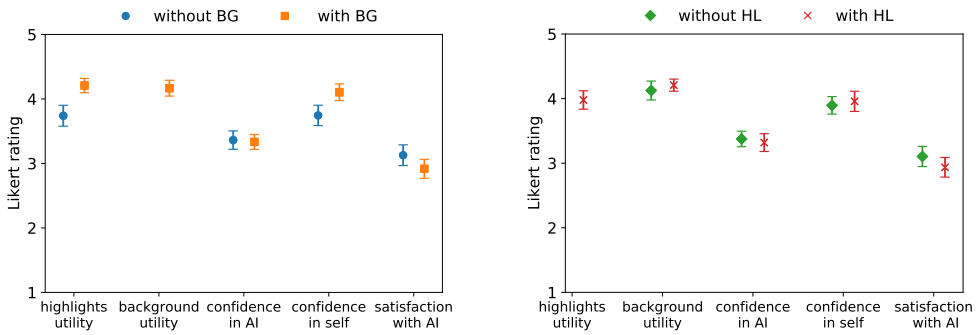

Fig. 5. Users' subjective rating of the system for usefulness of highlights, background, their confidence in the AI system, self-confidence, and satisfaction with AI. Users rate highlights marginally more useful in *with background conditions* than *without*, whenever applicable. Users also rate self-confidence marginally higher with the background explanations than without (left). However, users rate the satisfaction with AI with background explanations slightly lower than without. The satisfaction rating is also slightly lower with highlight explanations than without. However, there are no other discernible differences in ratings for *with vs without highlight conditions*.

incorrect judgments. This also holds for correct judgments: the difference in users' confidence in their correct judgments with and without highlights is negligible.

In summary, we find that highlights do not affect human-AI decision making process. This is possibly because the passages in the HotPotQA dataset are not too long. Thus, highlight-based explanations might not be as useful in understanding the AI predictions.

How does users' perception of AI change with/without background or highlight-based explanations? Figure 5 shows the user rating for the subjective measures across different conditions. We find that the differences in user ratings with vs without background explanations and with vs without highlight-based explanations are non-significant. However, we observe a marginally higher rating for highlights utility in the *with background* condition (4.21 ± 0.16) than the *without background* condition (3.74 ± 0.23). Adding background also increases users' confidence in their own judgment (from 3.74 ± 0.16 to 4.10 ± 0.13), consistent with users' aggregate confidence over individual examples (Figure 4 (left)). Despite the decrease in the rate of over-reliance on incorrect model predictions with background explanations (RQ1), users rate the satisfaction with AI system marginally lower in the *with background explanation* conditions (2.91 ± 0.15) than *without* (3.13 ± 0.16). This might be due to additional cognitive load required to consume background explanations.

These differences in users' ratings are much less prominent for highlight-based explanations where users assign similar ratings for confidence in AI predictions and confidence in self judgments with or without highlights. This parallels our findings on task performance where we see negligible improvements on adding highlight-based explanations, both in *with and without background* explanation conditions. Surprisingly, users rate the satisfaction with AI in the *with highlights* conditions marginally lower (2.94 ± 0.15) than the *without highlights* conditions (3.10 ± 0.16).

4 DISCUSSION AND CONCLUSION

In contrast to existing explainability methods focused on explaining salient parts of inputs [1, 19, 24], we study users' trust and reliance on AI systems with background explanations that provide additional information external to the model's input. We conduct our study in a multi-hop question answering setup. We design our study such that some

part of the information required to perform the reasoning is missing in the *without background explanation* conditions. This missing information is included as background explanation in the *with background explanation* conditions. These background explanations, although not necessarily faithful to the model's reasoning process [29], may still be helpful to users in assessing the correctness of AI predictions [18].

Our study reveals that users' reliance on correct AI predictions is fairly high (88%) even without sufficient information to assess the correctness of the predictions (without relevant background). This indicates blind trust in the AI system to some extent. We find that although background explanations do not affect reliance on correct model predictions, they significantly reduce over-reliance on incorrect model predictions (RQ1). This indicates that even though users trust the AI system without sufficient information, they are able to catch AI errors better when provided with the relevant information. On the flip side, our study reveals that the addition of more information also increases their confidence in their own judgments of the AI predictions, both those that are correct and, unfortunately, those that are incorrect. This is especially concerning as even if users are better at the task with background explanations, uncalibrated confidence might be detrimental in critical decision-making tasks.

Our work highlights the utility and pitfalls of including background information in explanations. Such background explanations can take many shapes—implicit information encoded in the pre-trained models used for reasoning or external information required to fill in knowledge gaps. Our work studies human interaction with such explanations in a quasi-simulated setting. More work is needed to study how such explanations can be gathered or generated automatically. Further, our work highlights the issue of overconfidence stemming from more information, even when that information points users with evidence to make the correct judgments. This indicates that explanations alone are not sufficient to garner appropriate trust and reliance. More efforts are needed to educate lay audience to inspect explanations critically and diligently.

ACKNOWLEDGMENTS

We would like to thank the reviewers and members of the CLIP lab at UMD for their constructive feedback. This work was funded in part by U.S. Army Grant No. W911NF2120076.

REFERENCES

[1] David Alvarez Melis, Harmanpreet Ka ur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (Oct. 2021), 35–47. https://ojs.aaai.org/index.php/HCOMP/article/view/18938

[2] Gagan Bansal, Tongshuang Sherry Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2020).

[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[4] Zana Buccinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces* (2020).

[5] Zana Buccinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To Trust or to Think. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 21.

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics* (2015), 160–169.

[7] Jifan Chen and Greg Durrett. 2019. Understanding Dataset Design Choices for Multi-hop Reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4026–4032. https://doi.org/10.18653/v1/N19-1405

[8] Valerie Chen, Qingzi Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *ArXiv* abs/2301.07255 (2023).

[9] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

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

[12] James X Hu, Hongyu Zhao, and Harrison H Zhou. 2010. False discovery rate control with groups. *J. Amer. Statist. Assoc.* 105, 491 (2010), 1215–1227.

[13] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy Jr., Roy 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 (2021)* (2021).

[14] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics* 8 (07 2020), 423–438. https://doi.org/10.1162/tacl_a_00324 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00324/1923867/tacl_a_00324.pdf

[15] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 4070–4073.

[16] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-Scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Valencia, Spain) *(AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 467–474.

[17] Vivian Lai and Chenhao Tan. 2018. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2018).

[18] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[19] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[20] Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

[21] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *ArXiv* abs/1907.12652 (2019).

[22] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[23] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[25] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, and Ghassemi M. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med* (2021).

[26] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[27] Maximilian Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the Influence of Explainable AI on Automation Bias. *ArXiv* abs/2204.08859 (2022).

[28] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is Multihop QA in DiRe Condition? Measuring and Reducing Disconnected Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8846–8863. https://doi.org/10.18653/v1/2020.emnlp-main.712

[29] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[30] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[31] Yunfeng Zhang, Qingzi Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).