To Err Is AI!

Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems

He, Gaole; Bharos, Abri; Gadiraju, Ujwal

DOI
10.1145/3648188.3675130

Publication date
2024

Document Version
Final published version

Published in
HT 2024

To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems

Gaole He
Delft University of Technology
Delft, The Netherlands
g.he@tudelft.nl

Abri Bharos
Delft University of Technology
Delft, The Netherlands
a.r.j.bharos@gmail.com

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju@tudelft.nl

ABSTRACT

Powerful predictive AI systems have demonstrated great potential in augmenting human decision making. Recent empirical work has argued that the vision for optimal human-AI collaboration requires '*appropriate reliance*' on AI systems. However, accurately estimating the trustworthiness of AI advice at the instance level is quite challenging, especially in the absence of performance feedback pertaining to the AI system. In practice, the performance disparity of machine learning models on out-of-distribution data makes the dataset-specific performance feedback unreliable in human-AI collaboration. Inspired by existing literature on critical thinking and mindsets, we propose debugging an AI system as an intervention to foster appropriate reliance. This paper explores whether a critical evaluation of AI performance within a debugging setting can better calibrate users' assessment of an AI system. Through a quantitative empirical study (N = 234), we found that our proposed debugging intervention does not work as expected in facilitating appropriate reliance. Instead, we observe a decrease in reliance on the AI system — potentially resulting from an early exposure to the AI system's weakness. Our findings have important implications for designing effective interventions to facilitate appropriate reliance and better human-AI collaboration.

ACM Reference Format:
Gaole He, Abri Bharos, and Ujwal Gadiraju. 2024. To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems. In *35th ACM Conference on Hypertext and Social Media (HT '24), September 10–13, 2024, Poznan, Poland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3648188.3675130

1 INTRODUCTION

With the rise of deep learning systems over the last decade, there has been widespread adoption of AI systems in supporting human decision makers [33], albeit without always fully understanding the societal impact or downstream consequences of relying on such systems [13, 14]. Due to the opaqueness of some AI systems, users (especially laypeople) have struggled to determine when exactly they are trustworthy. To realize the full potential of complementary team performance [3], human decision makers need to identify when they should rely on AI systems and when they are better off relying on themselves. Such a reliance pattern has been defined as *appropriate reliance* [26, 33, 43, 51].

In practice, it is common that users need to deal with data from unknown distributions and unseen contexts, meaning that AI systems in the real world need to provide advice on out-of-distribution data [9, 42]. Under such circumstances, the estimated performance of an AI system or the so-called 'stated accuracy' [25, 49, 61] of the system (*i.e.,* accuracy on pre-defined test sets) cannot faithfully reflect the trustworthiness of the AI system. To help users assess the trustworthiness of AI systems, a practical solution that has been proposed, is to provide meaningful explanations along with AI advice [50, 56]. Post-hoc explanations have been found to improve user understanding of AI advice in empirical studies exploring human-AI decision making [33, 58]. However, such explanations require relatively high AI literacy [8] and domain expertise [23] to verify. As a result, most existing XAI methods have remained ineffective in helping laypeople assess the trustworthiness of AI advice at the instance level, adversely affecting their degree of appropriate reliance (often inducing over-reliance) on AI systems [7, 58].

To realize the goal of appropriate reliance, human decision makers need to be capable of evaluating AI advice and the trustworthiness of the AI system critically. We argue that such a critical mindset can help users avoid blindly following AI advice (*i.e.,* avoiding *over-reliance*), and also prevent them from distrusting AI advice when it can be productive (*i.e.,* avoiding *under-reliance*). Inspired by recent works on explanation-based human debugging of AI systems [1, 37], we propose explanation-based debugging as a training intervention to increase appropriate reliance on AI systems. We posit that such a debugging intervention has the potential to help laypeople understand the limitations of AI systems — ***that neither explanations of the AI advice nor the advice itself are always reliable***. Recognizing these limitations can help users better understand when an AI system is trustworthy and thereby increase appropriate reliance on the system. In this paper, we aim to empirically evaluate the effectiveness of using a debugging intervention as a means to increase appropriate reliance. We address the following research questions — (**RQ1**) How can a debugging intervention help users to estimate the performance of an AI system, both at the instance and at the global level? and (**RQ2**) How does a debugging intervention affect the reliance of users on an AI system?

To this end, we propose three hypotheses considering the effect of the debugging intervention on AI performance assessment as well as reliance, and the task ordering effect of debugging intervention on appropriate reliance. We tested these hypotheses in an empirical study (N = 234) of human-AI decision making in a deceptive review detection task (*i.e.,* identifying whether a review excerpt is written based on real experience). However, we found

HT '24, September 10–13, 2024, Poznan, Poland

ACM ISBN 979-8-4007-0595-3/24/09
https://doi.org/10.1145/3648188.3675130

that the proposed debugging intervention neither calibrates user estimation of AI performance nor promotes appropriate reliance.

Our results highlight that when presented with the weakness of the AI system in an early stage of the debugging intervention, users underestimate AI performance and rely less on the AI system. Users' overestimation of their own competence may further amplify such an effect. Through an analysis exploring relatively less-competent individuals, we found that the underestimation of AI trustworthiness may also play a role in shaping under-reliance, which is potentially relevant to the metacognitive bias called the Dunning-Kruger effect [31]. Our work has important implications for designing effective interventions to promote appropriate reliance in the context of human-AI decision making.

2 RELATED WORK

We position our work in the context of (a) studies on human-AI decision making, (b) appropriate reliance on AI systems, and (c) explanation-based debugging of machine learning systems.

Human-AI Decision making. In recent years, deep learning methods have been used in a wide range of applications (like medical image analysis [40], autonomous driving [21]). However, due to the intrinsic uncertainty and opaqueness, it would be undesirable to make such AI systems automate decision making, especially in high-stakes scenarios (*e.g.,* legal judgment, medical diagnosis). Under such circumstances, AI systems are expected to play a supporting role for human decision makers. According to GDPR, users have the right to obtain meaningful explanations to work with such AI systems [53]. Motivated by this, a series of work has proposed constructing human-centered explainable AI systems [11, 12, 24, 39] for better human-AI collaboration. Existing work has widely explored how different user factors (*e.g.,* expertise [10, 45], risk perception [20], machine learning literacy [8]) and interaction designs (*e.g.,* performance feedback [2, 49, 60], explanation [58], user tutorial [34, 44]) will affect user trust in and reliance on AI systems.

Appropriate Reliance on AI Systems. One important goal of human-AI decision making is complementary team performance [3, 42], which requires appropriate reliance [36]. In practice, however, humans always misuse (*i.e.,* over-reliance [48]) or disuse (*i.e.,* under-reliance [5, 58, 59]) AI systems. Such inappropriate reliance results in sub-optimal team performance, which is even worse than AI alone [3, 42]. To mitigate such issues, existing work has proposed different interventions including user tutorials [8, 34], cognitive force functions [4], and improving AI literacy [9]. Another stream of work proposed to improve the transparency of AI systems with effective explanations [35, 58], performance feedback [43], and global model properties [6]. In summary, these works presented users with extra information about AI systems (more than advice) or changed users' mindset and knowledge of AI systems.

Explanation-based Debugging. Explanation-based debugging was found to be helpful for improving human understanding of machine learning system [32]. Recent works [1, 37] have explored how to leverage explanations for model debugging across a wide range of tasks. The core idea of explanation-based debugging is to check whether the explanations from AI systems misalign with human knowledge. From human feedback, it would be possible to improve machine learning models' robustness, *e.g.,* by reducing spurious reasoning patterns [41, 54] and bias in dataset [27]. Debugging in programming is the process by which programmers can determine the potential errors in the source code and resolve these errors [57]. Inspired by this, we proposed debugging as an intervention to help participants understand the limitations of both explanations and advice of AI systems. In such an error finding and resolution process, users may learn when the AI system is trustworthy as a result of engaging directly with their pitfalls.

3 TASK, HYPOTHESES, AND INTERVENTION

3.1 Deceptive Review Detection Task

In this paper, we base our experiment within a challenging task – deceptive review detection. In each task, based on a hotel review, participants are asked to identify whether it is genuine (*i.e.,* written by real customers) or deceptive (*i.e.,* written by people who did not stay at the hotel). An example of this task is shown in Figure 1(a). This task has been used in prior work exploring Human-AI decision making [34, 35]. We also used the same public dataset [35]. [1]

Using Text Highlights as Explanations. In our study, we consider a real-world scenario where the performance of an AI system is not provided or available. To help participants assess the trustworthiness of advice from the AI system in each instance of decision making, we provide local explanations for each prediction. Following Lai *et al.* [34], we adopted BERT-LIME to generate text highlights as local explanations for each AI advice. We first finetuned the BERT [29] (bert-base-uncased) on the deceptive review detection dataset, and then generated the top-10 highlighted features from post-hoc XAI method LIME [50] as explanations.

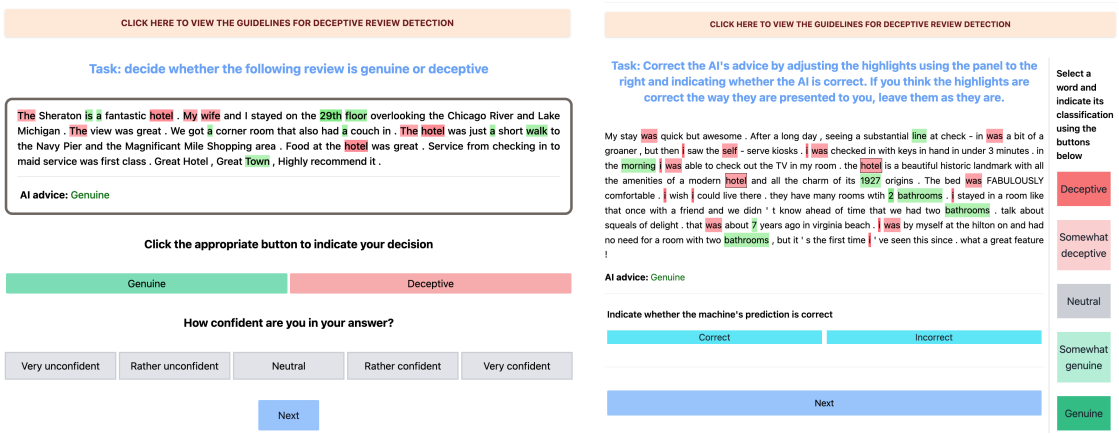
Selection of Tasks. To measure the effect of the debugging intervention in our study, two batches of tasks with compatible difficulty levels are required. For that purpose, we conducted a pilot study on human performance over 20 tasks randomly sampled from evaluation and test set of the deceptive review detection dataset. We divided the trial cases into two sets of 10 tasks with equal human performance in a pilot study (10 participants). In each task batch, the AI system achieved 80% accuracy.

Two-stage Decision Making. Following existing empirical study design of human-AI decision making [18, 19, 26], all participants in our study work on each trial case with two stages of decision making. In the first stage, only task input is provided; participants make an initial decision by themselves. After that, the same task input along with a local explanation and AI advice are provided. Participants make their final decision based on all information.

3.2 Hypotheses

Our experiment was designed to answer questions surrounding the impact of the proposed explanation-based debugging intervention on user estimation of AI performance, and user reliance on AI systems. Putting users into a debugging setting, they will try to challenge the AI advice and explanations. Along with the real-time feedback about the debugging results, they can have a better understanding of how the AI system works and when the explanation

[1] https://github.com/vivlai/deception-machine-in-the-loop



(a) Task interface of the deceptive review detection task.

(b) The debugging interface.

Figure 1: Screenshots of deceptive review detection task in our study.

and advice are reliable. Thus, they can more accurately estimate the performance of the AI system when no performance of the AI system is provided, and rely on the AI system more appropriately. Based on this, we expect to observe:

- **H1**: Encouraging users to critically evaluate the trustworthiness of AI advice at the instance level in a debugging intervention, will improve their assessment of the AI system's performance at the instance and global levels.
- **H2**: Encouraging users to critically evaluate the trustworthiness of AI advice at the instance level in a debugging intervention, will improve appropriate reliance on the system.

Within a debugging intervention, to present a balanced view of AI systems, we considered showing both the strength and weakness of an AI system (by providing accurate or inaccurate advice). Thus, multiple tasks of different characteristics will be presented in the debugging intervention. When these tasks are presented in different orders, users may show different learning effects, which further affects the reliance on AI systems. Thus, we hypothesize that:

- **H3**: The trustworthiness of AI advice at the instance level in a debugging intervention corresponds to an ordering effect with respect to appropriate reliance.

3.3 Debugging Intervention

To help participants accurately assess the trustworthiness of AI advice at the instance level and calibrate their reliance on the AI system, we designed a debugging intervention with explanations generated with post-hoc explanation methods LIME [50]. Our data, code, and further analysis are available at the companion page.[2]

Explanation-Based Human Debugging. Through the debugging phase, all participants are supposed to learn two important facts about the AI system: (1) the AI advice is not always correct, and (2) explanations are not always informative and helpful in identifying the trustworthiness of AI advice. Thus, we considered two main factors for each task: (1) the correctness of AI advice, and (2) whether an explanation is informative (*i.e.,* combined with guidelines, whether or not such explanations can help participants easily identify the correct answer). Participants subjected to training were presented with a hotel review with explanatory elements consisting of a model prediction and color-coded highlights showing 10 predominant features. Each token highlight shows the contribution of the token to the model prediction on a 5-point Likert scale: *deceptive*, *somwhat deceptive*, *neutral*, *somewhat genuine*, *genuine*. This difference in the contribution is distinguished by the color and intensity of the highlight shown in the interface.

An example of the debugging phase is shown in Figure 1(b). They are instructed to read the text and, when deemed necessary, refine the explanations by adjusting the highlights and indicating whether the AI advice is correct. After each task, the correctness of AI advice and missed adjustments will be shown as real-time feedback. Besides realizing the explanations are not always helpful, we hope participants can learn patterns they can rely on to make decisions given the guidelines. With that wish, the authors manually adjusted the highlights generated with BERT-LIME according to the task guidelines (from [34]). The adjusted highlights are taken as ground truth for the debugging phase.

Selection of the Debugging Tasks. The eight tasks presented in our debugging phase are: (1) two tasks with correct AI advice and informative explanations, (2) two tasks with correct AI advice and uninformative explanations, (3) two tasks with incorrect AI advice and informative explanations, (4) two tasks with incorrect AI advice and uninformative explanations. The tasks are balanced in whether explanations are informative and AI advice is correct. While the informative explanations are manually selected, the correctness of AI advice is determined randomly. In practice, it would be impossible to estimate the performance of the AI system on data sampled from an unknown distribution. In our study, the accuracy of the AI system in the debugging phase was intentionally set to 50%, which is lower than the 80% accuracy during the task phase. On the one hand, our debugging phase is proposed to develop a critical mindset about the AI system, instead of informing users

[2]https://osf.io/dh34y/?view_only=6a6833eafdbd4d5daa8c036579247159

how well the AI system performs. On the other hand, this setup prevents information leakage about AI performance, which may substantially affect user reliance behaviors.

Ordering Effect. When presenting the debugging phase to participants, the order of tasks may have an impact on their estimation of AI performance and reliance on the AI system. According to existing work [46, 55], first impressions greatly affect user estimation of AI performance and user trust in AI systems. Overall, both correct AI advice and informative explanations tend to leave positive impression on users. As pointed out by a recent study [47], the public would prioritize the accuracy of AI systems over interpretability. Thus, compared with "wrong AI advice, informative explanation" case, we would consider "correct AI advice, uninformative explanation" will leave participants a better impression. With these concerns, we designed three orders of tasks: (1) *Random order*. (2) *Decreasing impression order*: correct AI advice, informative explanation → correct AI advice, uninformative explanation → wrong AI advice, informative explanation → wrong AI advice, uninformative explanation. (3) *Increasing impression order*: wrong AI advice, uninformative explanation → wrong AI advice, informative explanation → correct AI advice, uninformative explanation → correct AI advice, informative explanation.

4 STUDY DESIGN

Experimental Conditions. In all conditions, the top-10 most important features obtained from BERT-LIME are highlighted as an explanation for AI advice to help participants identify the trustworthiness of AI advice. The differences between conditions are whether debugging intervention is adopted and the order of debugging tasks. To comprehensively study the effect of debugging intervention, we considered four experimental conditions in our study: (1) no debugging intervention (**`Control`**), (2) with debugging intervention, debugging tasks in random order (**`Debugging-R`**), (3) with debugging intervention, debugging tasks in decreasing impression order (**`Debugging-D`**), (4) with debugging intervention, debugging tasks in increasing impression order (**`Debugging-I`**).

Measures And Variables. To verify **H1**, we assessed participants' global estimation of AI system's performance with two questions: "From the previous 10 tasks, on how many tasks do you estimate the AI advice to be correct?" and "From the previous 10 tasks, how many questions do you estimate to have been answered correctly? (after receiving AI advice)". The answers to the two questions correspond to participants' estimation of AI performance and team performance respectively. We can refer to the estimated trustworthiness as estimated AI performance (**EAP**) and estimated team performance (**ETP**). Comparing that performance estimation with actual performance in abstract difference, we can calculate the degree of miscalibration of AI performance (**MAP**) and team performance (**MTP**). For the AI performance estimation at instance level, we calculated the number of tasks they made the correct final decision with an indication of "Very Confident" (**CCD**).

To verify **H2** and **H3**, we measured both reliance and appropriate reliance of participants on the AI system. Details of calculating these measures can be found in [26]. The reliance is measured with two widely adopted metrics: the **Agreement Fraction** and the **Switch Fraction**. These look at the degree to which participants are in agreement with AI advice, and how often they adopt AI advice in cases of initial disagreement. As for the appropriate reliance, we followed Max *et al.* [51] to calculate the appropriate reliance with relative positive AI reliance (**RAIR**) and relative positive self-reliance (**RSR**). The two measures are calculated when an initial disagreement between the human initial decision and AI advice exists where the correct answer occurs in one of them. In addition, we consider the accuracy in batches to measure participants' performance with AI assistance.

For a deeper analysis of our results, a number of additional measures were considered based on observations from existing literature [38, 52, 55]: Trust in Automation (TiA) questionnaire [30], a validated instrument to measure (subjective) trust [55] consisting of 6 subscales. Affinity for Technology Interaction Scale (ATI) [16], administered in the pre-task questionnaire. Thus, we account for the effect of participants' affinity with technology on their reliance on systems [55]. NASA-TLX questionnaire [22] for the working load assessment of the debugging intervention.

Participants. Following previous work [25, 26], we computed the required sample size in a power analysis for a Between-Subjects ANOVA using G*Power [15]. This resulted in a required sample size of 230 participants. We thereby recruited 324 participants from the crowdsourcing platform Prolific, in order to accommodate potential exclusion. All participants were rewarded with £3.8, amounting to an hourly wage of £7.6 (estimated completion time was 30 minutes). We rewarded participants with extra bonuses of £0.05 for every correct decision in the 20 trial cases. Such an extra bonus for correct decisions provides an additional monetary incentive for crowd workers to try their best on each task, a widely adopted design in existing work to increase participant reliability [9, 34].

Filter Criteria. All participants were proficient English speakers above the age of 18. For a high-quality study, we require participants to have an approval rate of at least 90% and more than 80 successful submissions on the Prolific platform. After reading the basic introduction and guidelines about the deceptive review detection task, participants who failed any qualification test (about understanding the task) were removed from our study. After data collection, we excluded participants from our analysis if they failed any attention check (90 participants). The resulting sample of 234 participants had an average age of 39 (SD = 13) and a gender distribution (48.7% female, 49.6% male, 1.7% other).

Procedure. The complete procedure of our study is illustrated in Figure 2. All participants are first presented with an introduction of the deceptive review detection task. According to Lai *et al.* [34], guidelines about identifying deceptive reviews are highly useful in improving user performance on this task. Thus, we also follow them to provide the guidelines in the introduction. Then, participants will be checked with two qualification questions to ensure they carefully read the instruction and understand this task. Any failure at the qualification test will result in removal from our study. All reserved participants will then be asked to answer a pre-task questionnaire.

For all conditions, participants will work on the first batch of tasks and go through a post-task questionnaire to assess AI performance and subjective trust in AI system (*i.e.,* with TiA subscales).

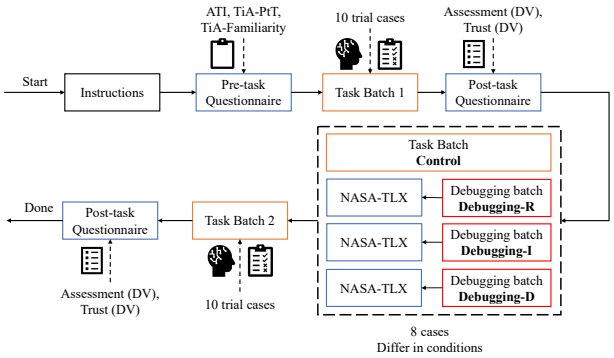


Figure 2: Illustration of the procedure that participants followed within our study.

The main difference between conditions (shown in the dashed box of Figure 2) is the 8 tasks presented after the post-task questionnaire. In condition **`Control`**, participants will work on the 8 tasks as normal trial cases. No debugging intervention and result feedback will be provided. In conditions with debugging intervention, the participants will go through the debugging tasks with different task orders and be asked about the task working load resulting from the debugging intervention, using the NASA-TLX [22] questionnaire. Then, participants in all conditions will continue to work on another batch of tasks and answer the same post-task questionnaire as the one after the first task batch. The AI accuracy on each task batch is 80%. To eliminate the potential ordering effect of trial cases, we randomly assigned one batch of selected tasks (see section 3.1) as the first batch and shuffled the task order within each batch.

5 RESULTS AND ANALYSIS

In our analysis, we only consider participants who passed all attention checks as a control of participant reliability [17]. Participants were distributed in a balanced fashion over experimental conditions: 57 (**`Control`**), 59 (**`Debugging-R`**), 60 (**`Debugging-D`**), 58 (**`Debugging-I`**). On average, participants spend around 51 minutes. Due to the page limit, more results can be found in the Appendix.

Performance Overview. On average across all conditions, participants achieved an accuracy of 0.64 over the two batches of tasks, still lower than the aforementioned AI accuracy of 0.8. The agreement fraction is 0.66 while the switching fraction is 0.31. With these measures, we confirm that participants in our study did not blindly rely on the AI system when disagreement appeared. In the two batches of tasks, the average estimated AI performance is 5.81 and 5.79 respectively; the average estimated team performance is 6.64 and 6.44 respectively. Overall, participants underestimated the performance of the AI system and believed they could outperform the AI system on this task after receiving AI advice.

5.1 Hypothesis Tests

5.1.1 H1: the effect of critical evaluation setting on AI performance estimation. To verify **H1**, we used Wilcoxon signed rank tests to compare all assessment-based dependent variables of participants before and after the debugging intervention. The results are shown in Table 1. Although no significant results were found to support **H1**, participants in **`Debugging-D`** condition showed a worse **MTP** after the debugging intervention, in contrast to our expectations. Thus, **H1** is not supported.

Table 1: Wilcoxon signed ranks test results for H1 on AI performance estimation. "†" indicates the effect of variable is significant at the level of 0.017 (adjusted alpha).

Condition	Debugging		Debugging-R		Debugging-D		Debugging-I	
DV	T	p	T	p	T	p	T	p
MAP	3833	.662	363	.742	463	.238	457	.826
MTP	4006	.957	512	.892	324	**.992**†	528	.160
CCD	3761	.753	474	.717	379	.660	429	.603

5.1.2 H2: the effect of critical evaluation setting on appropriate reliance. Similarly, to verify **H2**, we used Wilcoxon signed rank tests to compare all reliance-based dependent variables of participants before and after the debugging intervention. Overall in all conditions with the debugging intervention, the improvement in reliance caused by debugging intervention was not statistically significant. With a post-hoc Mann-Whitney test on **Accuracy**, we found that: after the debugging intervention, the accuracy drops significantly. For a fine-grained analysis, we further conducted Wilcoxon signed rank tests on each condition with the debugging intervention. We found that participants in the **`Debugging-I`** condition show a significant difference in **RAIR**, while no significant difference is found with post-hoc Mann-Whitney test. The observed results do not support the **H2**. Although no significant improvement was found in the performance and reliance measures due to debugging intervention, we did witness a drop in reliance measures generally: **Accuracy** (0.67 → 0.63), **Agreement Fraction** (0.68 → 0.66), **Switch Fraction** (0.34 → 0.28), **RAIR** (0.38 → 0.30), **RSR** (0.64 → 0.61). This is evident in the condition **`Debugging-I`**: **Accuracy** (0.68 → 0.63), **Agreement Fraction** (0.71 → 0.66), **Switch Fraction** (0.39 → 0.29), **RAIR** (0.43 → 0.29), **RSR** (0.59 → 0.61). When AI advice is in disagreement with users' initial decisions, users tend to rely on themselves more than they should. This results in decreased (appropriate) reliance and accuracy. In the deceptive review detection tasks, the AI system achieved a better performance than the participants. The reduced reliance (mainly under-reliance) may help explain why we found a decreased accuracy on average.

5.1.3 H3: ordering effect of debugging tasks. To analyze the ordering effect in the debugging phase, we compared the difference of reliance-based dependent variables (*i.e.,* the difference between the second batch and the first batch) and user reliance on the second batch with participants of all conditions with Kruskal Wallis test. No significant difference is found with such comparisons. To compare the task working load brought by debugging intervention of different ordering, we conducted Kruskal-Wallis H-test on the six measures in the NASA-TLX questionnaire. No significant difference was found, rejecting **H3**. To further look at how the ordering effect of debugging tasks affects the final performance of participants. We counted the participants who achieved an accuracy level above 80% (*i.e.,* compatible with or better than provided AI system) in the second task batch. After filtering out the participants who blindly rely on the AI system (*i.e.,* **Agreement Fraction** is 1.0), we found

the number of participants in condition **`Debugging-D`** (14) is clearly more than in condition **`Debugging-R`** (9) and **`Debugging-I`** (9). In comparison, the number of participants who achieved an accuracy level above 80% in condition **`Control`** is 11. Although the ordering effect does not show a significant statistical difference, such an observation lends partial support to **H3**.

5.2 Exploratory Analyses

Users' Estimation of AI Trustworthiness. To further understand how users' estimation of AI trustworthiness affects their reliance and performance, we split the participants in all conditions into performance-based quartiles: top quartile (top 25%), bottom quartile (bottom 25%), and middle quartile (other). To avoid the impact of debugging intervention, we only considered user performance in the first batch of tasks. To understand how these participants differ in their appropriate reliance and estimation of AI trustworthiness, we adopted the Kruskal-Wallis H test and Post-hoc Mann-Whitney tests to compare the estimated performance and their assessment of the AI system's performance at the instance and global levels. Generally, participants in the top quartile showed more appropriate reliance (*i.e.,* **RAIR** and **RSR**) than the bottom and middle quartile (with statistical significance). The results of user estimation of performance, AI trustworthiness, and miscalibration of performance are shown in Table 2. Overall, participants in the top quartile showed significantly higher **EAP** and **ETP** in comparison with the bottom and middle quartile. Meanwhile, the top quartile also has a more precise estimation of AI performance and team performance (*i.e.,* significantly lower **MAP** and **MTP**) and makes more correct decisions confidently (significantly higher **CCD**). It also indicates that the underestimation of AI trustworthiness can be the main cause of under-reliance, which results in lower accuracy.

Table 2: Kruskal-Wallis H-test results for user estimated trustworthiness and miscalibration of estimated performance based on performance quartiles. "†" indicates the effect of variable is significant at the level of 0.017 (adjusted alpha).

Variable	H	p	Post-hoc results
EAP	41.54	**<.001**†	Top > Middle > Bottom
ETP	15.85	**<.001**†	Top > Middle, Bottom
MAP	40.89	**<.001**†	Top < Middle < Bottom
MTP	22.67	**<.001**†	Top, Middle < Bottom
CCD	23.17	**<.001**†	Top, Middle > Bottom

6 DISCUSSION

Key Findings. To promote appropriate reliance on AI systems by calibrating user estimation of AI performance, we proposed a debugging intervention to educate participants that AI systems are not always reliable and that the explanations may not always be informative. As opposed to our hypotheses backed by existing work, we found that such a debugging intervention fails to calibrate participants' estimation of AI performance at both the global and local levels. Participants tended to rely less on the AI system after receiving the debugging intervention. Through an exploratory analysis based on performance quartiles, we found that participants who performed worse in our study tended to underestimate AI performance. Thus, their suboptimal team performance can be largely explained by the under-reliance on the AI system. The *plausibility* of the XAI intervention [28] can alternatively explain our findings. The debugging intervention may have potentially made the XAI (*i.e.,* text highlights in our study) seem less plausible to users, resulting in a tendency to underestimate AI performance.

In our study, no significant difference was found between the different ordering of debugging tasks across experimental conditions. However, participants who were exposed to the weakness of the AI system at the beginning of the debugging intervention showed a more obvious tendency to disuse the AI system. Such under-reliance was found to result in sub-optimal team performance. This finding is in line with recent work that has uncovered similar ordering effects and cognitive biases influencing outcomes in human interaction with intelligent systems [46, 55]: a bad first impression of an AI system can lead to an underestimation of AI competence and reduced reliance on the system.

Implications. Our findings suggest that the debugging intervention and similar interventions with training purposes (*e.g.,* user tutorials) may suffer from the cognitive bias brought by the ordering effect within such interventions. While using such interventions to demonstrate the strength and weakness of AI systems, we should be careful not to leave users with a bad first impression, highlighting the weakness of the AI system. In our study, participants tended to be optimistic about the team performance while underestimating the AI performance. This is possibly caused by the meta-cognitive bias of the Dunning-Kruger effect [26, 31]. In our study, we found that less-competent individuals showed a greater tendency to underestimate the AI performance and make fewer correct decisions with confidence (see Table 2). This indicates that underestimating AI systems can also contribute to under-reliance in the context of human-AI decision making. According to He *et al.* [26], an overestimation of self-competence can result in under-reliance on the AI system. Both the overestimation of self-competence and the underestimation of AI competence can contribute to an illusion of superior competence over the AI system. As a result, users with such an illusion tend to disuse the AI system.

7 CONCLUSION

In an age where laypeople have democratized access to various AI-based decision support systems on the web, promoting appropriate reliance on AI systems in decision-making contexts is an important goal. In this paper, we present an empirical study to understand the impact of a debugging intervention on the estimation of AI performance and user reliance on the AI system. Our results suggest that we should be careful in presenting the weakness of the AI system to users to avoid anchoring effects which may result in under-reliance. While our experimental results do not provide support to our original hypotheses informed by prior work, more work is required to understand the potential of debugging interventions in facilitating appropriate reliance on AI systems. Future work may explore how to mitigate potential bias brought by the users' overestimation of themselves and the underestimation of AI performance.

REFERENCES

[1] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2022. How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[2] 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.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] 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.

[5] 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.

[6] 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.

[7] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? *arXiv preprint arXiv:1810.12366* (2018).

[8] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, Giulio Jacucci, Samuel Kaski, Cristina Conati, Simone Stumpf, Tuukka Ruotsalo, and Krzysztof Gajos (Eds.). ACM, 148–161.

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

[10] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies* 162 (2022), 102792.

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

[12] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[13] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For What It's Worth: Humans Overwrite Their Economic Self-interest to Avoid Bargaining With AI Systems. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[14] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 43–52.

[15] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[16] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[17] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1631–1640.

[18] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[19] 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.

[20] Ben Green and Yiling Chen. 2020. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *arXiv preprint arXiv:2012.05370* (2020).

[21] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* 37, 3 (2020), 362–386.

[22] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[23] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2022. It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 89–101.

[24] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2024. Opening the Analogical Portal to Explainability: Can Analogies Help Laypeople in AI-assisted Decision Making? *Journal of Artificial Intelligence Research* (2024).

[25] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[26] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[27] Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K Thiruvathukal, and Ming Yin. 2020. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*. 2955–2961.

[28] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2023. Rethinking AI Explainability and Plausibility. *arXiv preprint arXiv:2303.17707* (2023).

[29] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[30] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.

[31] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[32] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[33] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[34] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–13.

[35] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

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

[37] Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics* 9 (2021), 1508–1528.

[38] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. 2019. No risk no trust: Investigating perceived risk in highly automated driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 177–185.

[39] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[40] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[41] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. 2020. Towards hybrid human-AI workflows for unknown unknown detection. In *Proceedings of The Web Conference 2020*. 2432–2442.

[42] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum. Comput. Interact.* 5, CSCW2 (2021), 1–45.

[43] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 78:1–78:16.

[44] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5323–5331.

[45] Mahsan Nourani, Joanie T. King, and Eric D. Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. (2020).

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

[47] Anne-Marie Nussberger, Lan Luo, L Elisa Celis, and Molly J Crockett. 2022. Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. *Nature Communications* 13, 1 (2022), 1–13.

[48] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI: Literature review. (2022).

[49] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, 535:1–535:14.

[50] Marco Túlio 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, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144.

[51] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. In *ACM Conference on Human Factors in Computing Systems (CHI'22), Workshop on Trust and Reliance in AI-Human Teams (trAIt)*.

[52] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* 2, 8 (2020), 476–486.

[53] Andrew Selbst and Julia Powles. 2018. "Meaningful Information" and the Right to Explanation. In *Conference on Fairness, Accountability and Transparency*. PMLR, 48–48.

[54] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2022. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In *Proceedings of the ACM Web Conference 2022*. 882–892.

[55] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 77–87.

[56] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283.

[57] Gamze Türkmen and Sonay Caner. 2020. THE INVESTIGATION OF NOVICE PROGRAMMERS'DEBUGGING BEHAVIORS TO INFORM INTELLIGENT E-LEARNING ENVIRONMENTS: A CASE STUDY. *Turkish Online Journal of Distance Education* 21, 3 (2020), 142–155.

[58] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[59] Ilan Yaniv and Eli Kleinberger. 2000. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes* 83, 2 (2000), 260–281.

[60] 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.

[61] 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 *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 295–305.