You Can Only Verify When You Know the Answer: Feature-Based Explanations Reduce Overreliance on AI for Easy Decisions, but Not for Hard Ones

Zelun Tony Zhang
zhang@fortiss.org
fortiss GmbH, Research Institute of the Free State of Bavaria
Munich, Germany
LMU Munich
Munich, Germany

Felicitas Buchner
fe.buchner@campus.lmu.de
LMU Munich
Munich, Germany

Yuanting Liu
liu@fortiss.org
fortiss GmbH, Research Institute of the Free State of Bavaria
Munich, Germany

Andreas Butz
butz@ifi.lmu.de
LMU Munich
Munich, Germany

ABSTRACT

Explaining the mechanisms behind model predictions is a common strategy in AI-assisted decision-making to help users rely appropriately on AI. However, recent research shows that the effectiveness of explanations depends on numerous factors, leading to mixed results, with many studies finding no effect or even an increase in overreliance, while explanations do improve appropriate reliance in other studies. We consider the factor of decision difficulty to better understand when feature-based explanations can mitigate overreliance. To this end, we conducted an online experiment ($N = 200$) with carefully selected task instances that cover a wide range of difficulties. We found that explanations reduce overreliance for easy decisions, but that this effect vanishes with increasing decision difficulty. For the most difficult decisions, explanations might even increase overreliance. Our results imply that explanations of the model's inner workings are only helpful for a limited set of decision tasks where users easily know the answer themselves.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

explainable AI, overreliance, human-AI decision-making, AI-assisted decision-making, decision difficulty, online experiment

ACM Reference Format:
Zelun Tony Zhang, Felicitas Buchner, Yuanting Liu, and Andreas Butz. 2024. You Can Only Verify When You Know the Answer: Feature-Based Explanations Reduce Overreliance on AI for Easy Decisions, but Not for Hard Ones. In *Proceedings of Mensch und Computer 2024 (MuC '24), September 01–04, 2024, Karlsruhe, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3670653.3670660

MuC '24, September 01–04, 2024, Karlsruhe, Germany

ACM ISBN 979-8-4007-0998-2/24/09
https://doi.org/10.1145/3670653.3670660

1 INTRODUCTION

As AI is increasingly used to support human decision-making [20], concerns about the opaqueness of modern machine learning algorithms and calls for making AI explainable have become commonplace [1, 10, 15]. A typical approach for using explanations in AI-assisted decision-making is to explain to end users how the AI produced its recommendation [26]. The hope is that this helps decision makers to rely on AI *appropriately* [33], i.e., to adopt correct or beneficial recommendations and to reject wrong or detrimental ones. However, many studies have found that explanations do not effectively improve appropriate reliance [3, 33, 38, 41] or even increase *overreliance* on AI [2, 4, 17, 21, 34], i.e., people become more likely to adopt detrimental AI recommendations. Recent evidence suggests that the ineffectiveness of explanations is due to users' lack of cognitive engagement with them [6, 14]. This has become an influential perspective on the issue and suggests a fundamental challenge to the approach of explaining AI recommendations to decision makers. However, in other studies, explanations do help users to rely on AI more appropriately [8, 37, 39], showing that there *are* conditions under which explanations have the intended effect, though it is often unclear what these conditions are.

In this paper, we aimed to systematically investigate how the effect of feature-based explanations—one of the most popular explanation styles [20, 36]—on overreliance depends on the difficulty of a decision, as AI can be used to support decision tasks of varying difficulties. Often, the goal is to improve human decision-making in difficult high-stakes decisions, such as medical diagnosis [4, 17], creditworthiness assessment [16], or recidivism prediction [24, 38], where the call for explainability is particularly prominent and at the same time, overreliance is especially undesired. We found that

the effectiveness of feature-based explanations for reducing overreliance can strongly depend on decision difficulty. In our experiment, explanations reduced overreliance for easy decisions, but not for difficult ones. We discuss the implications of these results for the use of explanations in AI-assisted decision-making.

2 RELATED WORK

2.1 Explainable AI (XAI)

Various explanation methods have been proposed, most of them being *post-hoc*, i.e., these methods explain already-trained models that are not inherently interpretable [36]. Post-hoc methods are criticized for their lack of faithfulness to how the model actually works and for being not informative enough to support decision-making [31], but are often used due to the difficulty of building inherently interpretable models for complex tasks. Post-hoc methods are commonly categorized by whether they are *local* (explaining a specific model prediction) or *global* (explaining how the model works as a whole), and by whether they are *specific* to a certain model or can be applied *agnostically* to any model [36]. For empirical studies on AI-assisted decision-making, the more relevant distinction is the result of the explanation, with *feature-based* (showing how individual features contribute to a model output) and *example-based* (showing representative examples from the training set) explanations being the most popular [20]. In this work, we focus on feature-based explanations, which are more common to our chosen study task (see Section 3.1).

2.2 Task Difficulty and Reliance

Several studies have been conducted in recent years to understand how human reliance on AI is influenced by task difficulty or related constructs, often in the context of decision-making. Parkes [28] found that rather than objective task complexity, it is subjective task difficulty that leads to increased reliance on a decision aid. Lu and Yin [25] studied how people's heuristics for reliance on AI depends on their observations of model performance on decisions where they have high confidence, i.e., decisions that they perceived as easier. In a study by Chiang and Yin [9], people relied more on AI on out-of-distribution decision tasks, since they perceived their own performance to be worse on these tasks, i.e., they perceived these tasks as harder. Papenmeier et al. [27] found that perceived AI accuracy is lower when the AI makes mistakes on easy decision tasks and higher when mistakes happen on difficult ones. Cao and Huang [7] observed that participants looked at AI recommendations for a longer time on more difficult tasks, even though it did not translate into higher agreement with the AI in their case. Taken together, these results consistently suggest that people are more likely to rely on AI in more difficult decisions.

However, not many studies investigate the role of explanations in this relationship. Wang and Yin [38] studied the effect of different explanations in two decision-making tasks, one where participants had more domain knowledge in, and one where they had less. They found that feature contribution explanations led to more appropriate reliance in the high-domain-knowledge task, but not in the low-domain-knowledge task. But while the amount of domain knowledge is linked to task difficulty, the study gives no direct evidence about the relationship between explanations, overreliance, and task difficulty, as a task can be easier or harder independently from the amount of domain knowledge.

A work where task difficulty was directly studied was conducted by Vasconcelos et al. [37]. They found that explanations reduce overreliance in more difficult tasks when they enable easy verification of the AI. However, their notion of task difficulty differs from what we aimed to study in this work. To manipulate task difficulty, Vasconcelos et al. varied the complexity of the mazes that participants had to solve. More complex mazes required more cognitive effort, but with enough effort, a single clear solution could always be found. In contrast, we were interested in settings like those studied by Lu and Yin [25] or Papenmeier et al. [27]. In these task settings, the effort remains constant across task instances, and difficulty differs in terms of how hard it is to decide between multiple plausible options. This notion of difficulty more closely reflects difficulty in many real-world tasks, e.g., when assessing medical images, the image complexity does not change from case to case, but in some cases, the diagnosis is less clear than in others. To differentiate this notion of difficulty from complexity-based task difficulty, we refer to it as *decision difficulty*.

We conclude that from related work, it remains unclear how explanations affect overreliance under varying levels of decision difficulty. We therefore pose the following research question:

> **RQ**: How does the effect of explanations on overreliance depend on decision difficulty?

Given that previous work indicates increasing reliance with increasing decision difficulty, we also expected that overreliance increases with decision difficulty. As previous work has found that explanations can induce blind trust in AI [2, 12], especially for difficult tasks [21], we anticipated explanations to further increase overreliance in difficult decisions. For easy decisions, our intuition was that AI should not have a big impact, as users can easily make the decisions themselves. We therefore expected no effect of explanations in easy decisions.

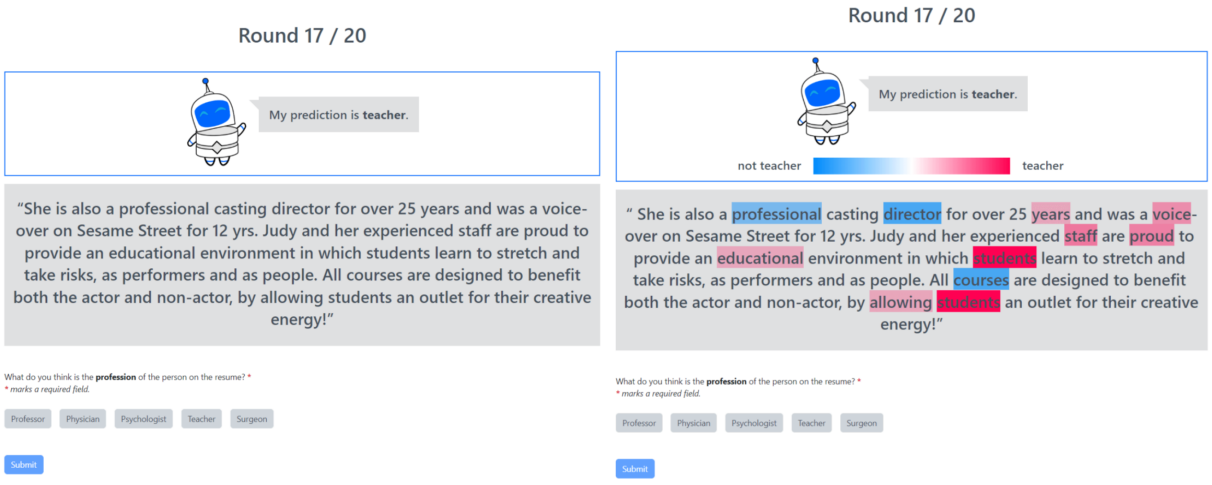
3 METHOD

We first describe the overall task on which we built the experiment (Section 3.1), followed by how we measured decision difficulty (Section 3.2) and how we selected the individual task instances for our experiment (Section 3.3). Lastly, we describe the study procedure (Section 3.4).

3.1 Study Task

As decision task, we chose profession classification based on a dataset by De-Arteaga et al. [11] of short biographies scraped from the internet. Each biography is labeled with one of 28 professions. This task has been used in previous studies on AI-assisted decision-making [8, 24, 29, 35]. We chose this task for two reasons: (1) It is accessible to non-expert participants. (2) Decisions have an objective ground truth, making it easier to assess overreliance.

The participants' task was to read a series of short biographies and decide which profession the described person has. To make the decision more manageable, we restricted the biographies to the same five professions as Liu et al. [24]: teacher, professor, physician, surgeon, psychologist. Participants were supported by an AI model which gave its prediction above the biography (Figure 1). Depending



(a) Without explanation. **(b) With explanation.**

Figure 1: Screenshots of the study interface.

on the study condition, participants in addition saw feature-based explanations in the form of text highlights (Figure 1b), indicating which words were most influential for the AI's prediction, and which words were speaking most against it. This interface mimics the typical interface for text-based studies on AI-assisted decision-making (e.g., [2, 21, 24, 34, 35]). We trained a logistic regression model with bag-of-words features on the five classes to generate the AI predictions and used LIME [30] to generate the feature-based explanations. The model had a test set accuracy of 0.89.

3.2 Measurement of Decision Difficulty

Our initial idea was to obtain decision difficulties by directly asking for participants' subjective difficulty ratings after each decision. However, while testing this setup, we found it to be unreliable, as participants' difficulty ratings could drift over the course of the experiment[1], and asking for the decision difficulty after each decision highly distracted from the actual decision-making task.

Instead, we settled on using the agreement between participants as measure for decision difficulty, similar to Papenmeier et al. [27]. The rationale was that high disagreement between participants indicates more than a single plausible answer, making the decision difficult. For our measure, we recorded how participants classified a biography *without* AI support and noted a_T, which is the share of participants who chose the most frequently chosen answer for task instance T. We linearly transformed a_T such that the resulting decision difficulty score d_T for task instance T lies between 0 for the easiest decisions, and 1 for decisions that are so hard that humans can only guess randomly:

$$d_T = 1 - \frac{a_T - a_{min}}{1 - a_{min}} = 1 - \frac{a_T - 0.2}{0.8}. \quad (1)$$

a_{min} is the theoretical minimum for a_T and is required to map d_T onto a range of $[0, 1]$. Since there are five possible answers, $a_{min} = 0.2$. Note that with this score, we did *not* consider the difficulty of identifying the correct answer, but only the difficulty of choosing between answers. If one answer stood out as the single most obvious answer, we treated it as an easy decision, even if the answer was wrong.

3.3 Selection of Task Instances

To measure overreliance, we sought tasks where the AI makes predictions that are (1) wrong and (2) different from what participants would independently decide without AI support. The first requirement follows from the commonly used definition of overreliance[2], which is that the human accepts a wrong AI recommendation [37, 39]. The latter requirement was important to ensure that when participants agreed with a wrong AI prediction, it was indeed the result of overreliance rather than an instance where participants' own opinion simply matched the AI recommendation. To answer our research question, we further had to ensure that the task instances in our experiment covered a wide range of decision difficulties. We carefully selected task instances according to these requirements by conducting a pre-study with 12 volunteers (average age: 31.1 ± 13.2 years; 5 female, 7 male). Figure 2 shows an overview of the entire procedure.

We first manually chose a set of 30 biographies out of those that were wrongly classified by our logistic regression model. We selected one biography for each combination of profession, gender[3], and decision difficulty, where decision difficulty in this first step was subjectively classified by the authors as either easy, medium, or hard. We then asked each of the 12 volunteers to label the professions for all 30 biographies, without seeing any AI predictions. The biographies were presented in random order to avoid ordering effects. We first excluded eleven biographies where the volunteers'

[1] As an illustrative example, one might have rated the first three tasks with the lowest difficulty of 1/5, but after three more tasks notice there are even easier ones, and would in hindsight rate the previous tasks 2/5.

[2] We acknowledge that—while popular and straightforward to operationalize—this definition has its limitations, as argued by Fok and Weld [13].

[3] The dataset only covers male and female genders.

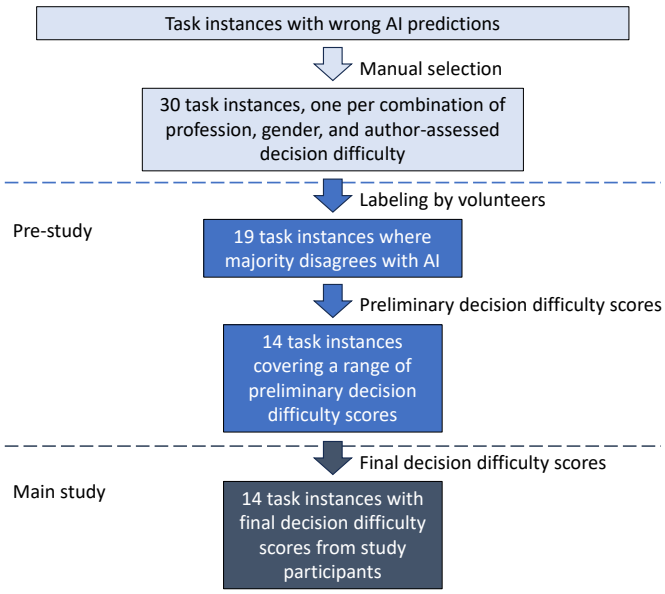


Figure 2: Overview of steps to select a set of tasks with well-distributed decision difficulty scores and wrong AI predictions.

most frequent answer matched the AI prediction. Based on the volunteers' labels, we then calculated preliminary decision difficulty scores for the remaining 19 task instances according to Equation 1. Finally, given these preliminary scores, we selected 14 out of these 19 task instances (see Appendix A) for our experiment, aiming for an even coverage of a wide range of preliminary decision difficulty scores. We also tried to balance professions and gender in this step. We kept only 14 task instances to keep the main study (see Section 3.4) short, as excessively long studies may inadvertently contribute to overreliance [42].

The decision difficulty scores in this step were preliminary since our volunteers did not necessarily match the demographics of our study participants. Note that this does not affect the validity of the main study results: As described in Section 3.4, we recalculated the final decision difficulty scores based on labels from our study participants and only used these final scores for the statistical analysis presented in Section 4. The preliminary scores were merely meant to ensure as much as possible that we went into the actual study with task instances covering a wide range of decision difficulties and were not used for later analyses.

3.4 Study Procedure

We ran a between-subject online study on Prolific[4] with two groups: with and without explanations. We restricted participants to native English speakers residing in the US, UK, or Canada, to ensure that participants would be familiar with the academic system referred to in the biographies. For the same reason, we required participants to have completed at least an undergraduate degree. Lastly, we required participants to have a minimum approval rate of 99% on Prolific.

After giving informed consent and completing a demographic survey, participants got instructions to the study task and interface. The main part of the study consisted of two phases: a labeling phase and a trial phase (Figure 3). In the labeling phase, participants made ten decisions *without* any AI support, including one attention check. In the trial phase, participants made another 20 decisions, again including one attention check, this time *with* AI support according to their study condition. The labeling phase had two purposes: first, helping participants to familiarize with the task, and second, obtaining labels from participants to calculate the final decision difficulty scores according to Equation 1. To get as many labels as possible, the labeling phase contained only task instances randomly sampled from the set of 14 task instances described in Section 3.3 (plus the attention check). The remaining five of these 14 task instances were included into the trial phase. The distribution of the 14 task instances of interest between labeling and trial phase was randomly resampled for each new participant.

The trial phase in addition contained 14 task instances where the AI made a *correct* prediction. Since overreliance can only be measured on task instances with wrong AI recommendations, these task instances with correct AI predictions were *not* of interest for our analyses, but only used to ensure that participants experienced a reasonable AI accuracy of 73.7% (excluding the attention check). We did not disclose the AI accuracy to participants and also gave no feedback whether they answered a task instance correctly or not. The set of task instances with correct AI predictions remained the same for all participants and where selected manually so that each profession and gender was equally represented. We also selected these task instances to cover a range of difficulties, as assessed subjectively by the authors. The order of all task instances in both phases was randomized between participants.

After the main part of the study, participants could choose to answer an optional exit survey where they were presented with the five task instances from the trial phase with wrong AI predictions along with their own answers. For each task instance, participants were asked how much they considered the AI in their decision (*"How much did you think about the AI prediction in this case?"*). For task instances where participants' answer was the same as the AI's, we additionally asked how much they considered to choose another answer (*"How much did you consider to choose a different answer than the AI prediction?"*). For task instances where participants decided differently than the AI, we asked how much they considered switching to the AI prediction (*"How much did you consider choosing the same answer as the AI prediction?"*). All questions were answered on a five-point scale (1 = *not at all*, 5 = *very strongly*). The exit survey was optional to ensure high-quality answers and to keep the study short for participants who were not motivated to answer the exit survey.

Participants were paid a base amount of £2.00 for completing the study. As incentive for accurate answers, participants were rewarded a bonus of £0.01 for each correct answer. Including the bonus, participants received an average payment of £2.21 for an average time of 16.46 minutes, resulting in an hourly rate of £8.06 (≈ US$10.24). The study was approved by the Ethics Committee

[4] https://www.prolific.com/

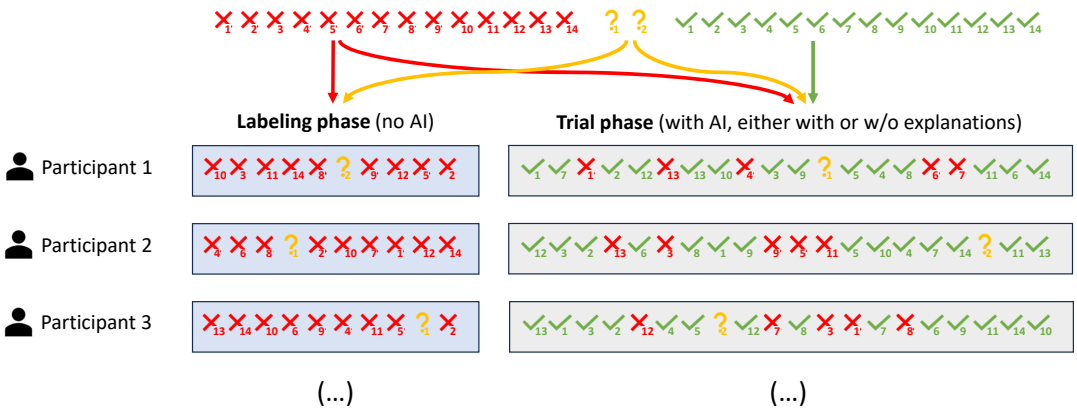


Figure 3: Distribution of task instances over labeling and trial phase. Red crosses symbolize task instances where the AI makes wrong predictions and which were of interest for our analyses. Green check marks symbolize task instances where the AI makes correct predictions and which were *not* of interest for our analyses. Yellow question marks symbolize attention checks.

of the Faculty of Mathematics, Computer Science and Statistics at LMU Munich.

4 RESULTS

A total of 237 participants completed the study in June 2023. We discarded the data of 34 participants for failing at least one attention check and filtered out three more participants who spent on average less than five seconds per task instance. Of the remaining 200 participants, 97 saw AI explanations and 103 saw no explanations. Participants had an average age of 39.3 ± 14.6 years. 100 participants self-identified as female and the other 100 as male. Additional data on participants' demographics is given in Appendix B.

The final decision difficulty scores of our 14 task instances with wrong AI predictions covered a range from 0.1 to 0.66, with a concentration between 0.4 and 0.5 (Figure 4). Each score was based on the answers of 128.6 ± 12.3 participants on average. In the trial phase, each of the 14 task instances was answered by an average of 71.4 ± 12.3 participants in both conditions together.

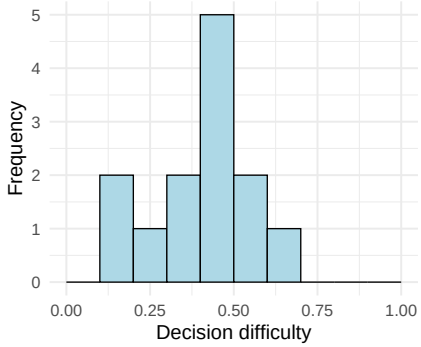


Figure 4: Histogram of final decision difficulty scores.

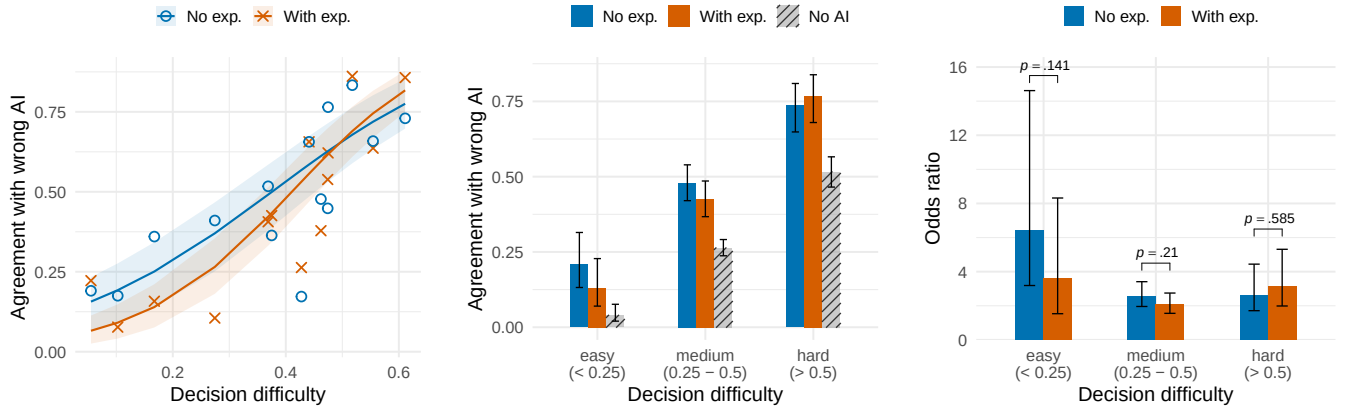
We investigated overreliance by three metrics: participants' agreement with wrong AI predictions (Section 4.1), their accuracy on task instances with wrong AI predictions (Section 4.2), and their subjective ratings of how much they considered the AI (Section 4.3). Details of the models mentioned in the following sections are given in Appendix C.

4.1 Agreement With Wrong AI

We considered the binary outcome for each task instance and participant whether the participant agreed with the wrong AI prediction (Figure 5). We fitted a mixed-effects logistic regression model with random intercepts and slopes for individual participants to account for repeated measures. As fixed effects, we added decision difficulty, explanation condition, and the interaction of both. We further controlled for participants' gender, age, education, machine learning knowledge, and AI attitude by adding those as fixed effects to the model. Likelihood ratio tests revealed significant[5] main effects for both decision difficulty ($OR = 581.11$, 95% CI [204.53, 2181.32], $\chi^2(1) = 136.38$, $p < .001$) and explanation ($OR = 2.99$, 95% CI [1.3, 6.83], $\chi^2(1) = 4.73$, $p = .03$). The interaction between both was also significant ($OR = 0.11$, 95% CI [0.02, 0.64], $\chi^2(1) = 4.25$, $p = .039$). As shown in Figure 5a, for easy decisions, agreement with wrong AI predictions was higher without explanations. Agreement increased for both conditions with decision difficulty, but faster with explanations, such that for the most difficult decisions, agreement was higher with explanations. None of the fixed effects for participant demographics were significant.

To further investigate the interaction effect, we divided the task instances into three clusters of decision difficulty to reduce the noise of individual task instances: easy (difficulty score < 0.25), medium (0.25–0.5), and hard (> 0.5). For each cluster, we fitted a separate mixed-effects logistic regression model with random intercepts for individual participants. As fixed effect, we added the form of AI support, including the two explanation conditions as well decisions without AI during the labeling phase. Figure 5b shows the estimated marginal means for these models, confirming that explanations led to less agreement with wrong AI for easy decisions, but more for hard decisions. However, both explanation conditions increased

[5] If not stated otherwise, we used a significance level of $p < .05$ throughout this paper.



(a) Agreement with wrong AI over decision difficulty. Dots represent the mean agreement for each task instance, lines the probabilities predicted by the model for an average participant.

(b) Estimated marginal mean agreement with wrong AI for clusters of decision difficulty.

(c) Odds ratios for agreement with wrong AI compared to when participants saw no AI in the labeling phase.

Figure 5: Agreement with wrong AI predictions. Error bands and bars represent 95% confidence intervals.

participants' agreement with wrong AI compared to decisions without AI for all decision difficulties. We ran linear hypothesis tests to test whether this increase in agreement differed between conditions within each decision difficulty cluster, but found no significant differences (Figure 5c).

4.2 Accuracy on Task Instances With Wrong AI

Similar to agreement, we considered the binary outcome for each task instance and participant whether the participant chose the correct profession when given a wrong AI recommendation (Figure 6). We fitted a mixed-effects logistic regression model and ran likelihood ratio tests akin to Section 4.1. Again, we found significant main effects for decision difficulty ($OR = 7.1 \times 10^{-4}$, 95% CI $[1.9 \times 10^{-4}, 1.9 \times 10^{-3}]$, $\chi^2(1) = 151.86$, $p < .001$) and explanation ($OR = 0.36$, 95% CI [0.16, 0.75], $\chi^2(1) = 4.02$, $p = .045$). An interaction between both is visible in Figure 6a, as for easy decisions, accuracy is higher with explanations, but the gap gradually closes with increasing decision difficulty and disappears entirely for the most difficult ones, although this interaction effect was not statistically significant ($OR = 6.32$, 95% CI [1.21, 33.1], $\chi^2(1) = 2.44$, $p = .118$).

As with agreement, we also analyzed accuracy with separate mixed-effects logistic regression models per decision difficulty cluster. The estimated marginal means in Figure 6b confirm that accuracy was higher with explanations for easy decisions, but almost the same for hard decisions. Similar to agreement, accuracy was lower than without AI for all decision difficulties, regardless of whether participants saw explanations. Linear hypothesis tests again revealed no significant differences in the accuracy reductions between the conditions (Figure 6c).

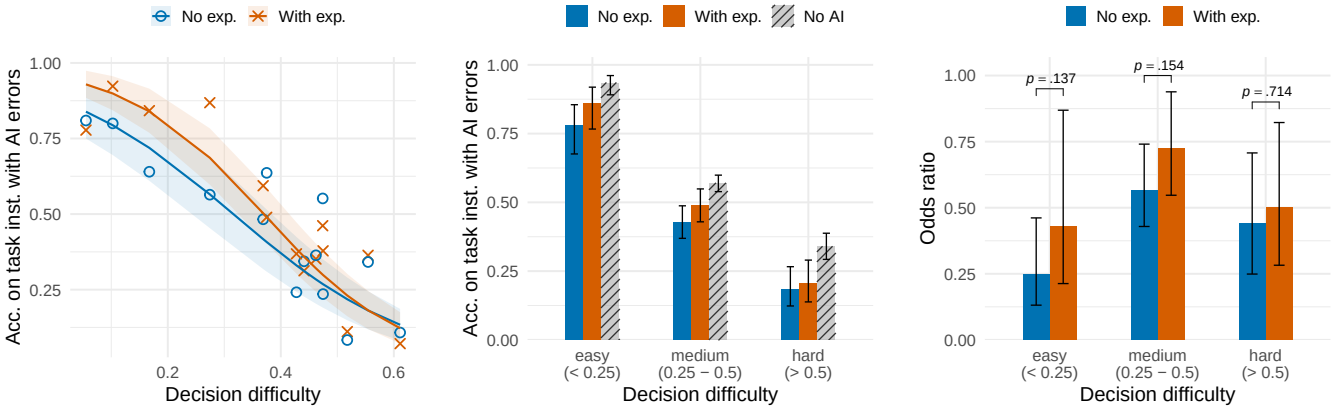
4.3 Self-Rated Consideration of AI

The optional exit survey was answered by 91 of the 200 participants. We fitted mixed-effects linear regression models to analyze participants' five-point scale answers, again with random intercepts and slopes for individual participants and controlling for participant demographics as fixed effects (Figure 7). As with the objective measures, we ran likelihood ratio tests to test for the main effects of decision difficulty and explanation and their interaction.

When asked how much they thought about the AI recommendation in task instances with wrong AI predictions (Figure 7a), participants' answers were not significantly affected by decision difficulty (Coef. = 0.22, 95% CI [-0.26, 0.7], $\chi^2(1) = 0.8$, $p = .371$) nor explanation condition (Coef. = -0.22, 95% CI [-0.69, 0.23], $\chi^2(1) = 0.9$, $p = .342$); the interaction of both was also not significant (Coef. = 0.77, 95% CI [-0.26, 1.84], $\chi^2(1) = 2.43$, $p = .119$).

We asked participants who correctly rejected a wrong AI recommendation how much they considered agreeing with the AI (Figure 7b). The main effect of decision difficulty on participants' answers was not significant (Coef. = 0.58, 95% CI [-0.22, 1.38], $\chi^2(1) = 2.07$, $p = .15$), while participants who saw explanations considered agreeing with the AI slightly more. This main effect of explanations was significant at the $p < 0.1$ level (Coef. = -0.64, 95% CI [-1.36, 0.04], $\chi^2(1) = 3.55$, $p = .06$). The interaction between decision difficulty and explanation was not significant (Coef. = 1.2, 95% CI [-0.29, 2.92], $\chi^2(1) = 2.2$, $p = .138$), even though the trend suggests that participants without explanations considered agreeing with the AI less on easier decisions, while the answers of participants with explanations remained constant for all decision difficulties.

We also asked participants who agreed with a wrong AI recommendation the complementary question of how much they considered disagreeing with the AI (Figure 7c). We found neither significant main effects for decision difficulty (Coef. = 0.54, 95% CI [-0.64,

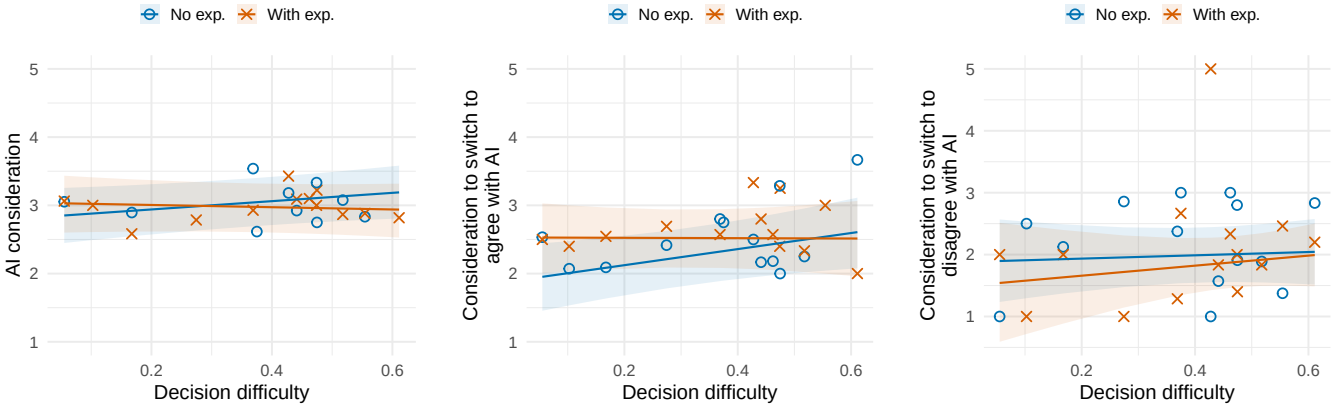


(a) Accuracy on task instances with wrong AI over decision difficulty. Dots represent the mean accuracy for each task instance, lines the probabilities predicted by the model for an average participant.

(b) Estimated marginal mean accuracy on task instances with wrong AI for clusters of decision difficulty.

(c) Odds ratios for accuracy on task instances with wrong AI compared to when participants saw no AI in the labeling phase.

Figure 6: Accuracy on task instances with wrong AI predictions. Error bands and bars represent 95% confidence intervals.



(a) *"How much did you think about the AI prediction in this case?"*

(b) When participants chose a different answer than the AI: *"How much did you consider choosing the same answer as the AI prediction?"*

(c) When participants chose the same answer as the AI: *"How much did you consider to choose a different answer than the AI prediction?"*

Figure 7: Participants' self-reported consideration of AI predictions in their decisions. Dots represent the means of the individual answers for each task instance. Lines represent model predictions for an average participant, error bands represent 95% confidence intervals.

1.87], $\chi^2(1) = 0.71$, $p = .399$) and explanation condition (Coef. = 0.38, 95% CI [-0.87, 1.64], $\chi^2(1) = 0.41$, $p = .52$), nor a significant interaction effect (Coef. = -0.54, 95% CI [-2.99, 1.85], $\chi^2(1) = 0.19$, $p = .662$).

4.4 Summary

We found that in line with previous work, overreliance increased strongly with decision difficulty, with higher agreement with wrong AI predictions and lower accuracy on more difficult decisions. To our surprise, AI recommendations induced overreliance even for the easiest decisions. Explanations reduced overreliance on these easy decisions, but did not completely prevent overreliance. With increasing decision difficulty, this positive effect of explanations became smaller and disappeared entirely for the most difficult decisions in our experiment.

All of this appears to happen unconsciously, as participants' subjective responses were mostly unaffected by both decision difficulty and explanation condition. Only when participants chose the correct profession, explanations led them to consider the wrong AI prediction slightly more, even though it did not lead to a stronger

adoption. Apparently, the explanations were plausible, but at least for easy decisions, still helped participants to reject wrong AI predictions.

5 DISCUSSION

Given the ineffectiveness of explanations in many studies, some authors have lately called the entire premise of explaining AI recommendations into question [5, 18, 26]. Other works have shown that explanations can effectively reduce overreliance under certain circumstances. Vasconcelos et al. [37] employed a cost-benefit framework to show that explanations can be effective when they reduce the cost of engaging with the task. On a theoretical level, Fok and Weld [13] argue that explanations only improve appropriate reliance when they enable verification of the AI's correctness. Our results also suggest that explanations can reduce overreliance in some scenarios. But similar to previous work, these scenarios are quite limited. Namely, explanations only seem to help for decisions that people could easily make themselves. This is consistent with Fok and Weld's theory, as people are likely able to verify the AI on such easy decisions. An application where this applies in practice could be the automation of easy, but high-volume decisions, such as unambiguous cases in content moderation [19], where explanations could help human supervisors to spot false machine classifications.

However, what researchers usually envision are more ambitious human-AI collaborations that augment human performance on tough decisions. On this front, our results are more in line with the voices that are skeptical about current approaches in AI explainability, as explanations did not reduce overreliance for more difficult decisions. Again, this is consistent with the theory of verifiability. We assume that for difficult decisions where humans are uncertain about the answer, they lack a reference against which they can verify the AI. The trend in our results suggests that explanations might even increase overreliance on even harder decisions than the most difficult ones in our experiment, which would be consistent with previous work on a similar setup, but where the task was very hard for humans [21].

We caution against overgeneralizing our results, as we only explored a single task with a single explanation style, with a single specific setup (e.g., showing AI recommendations before participants made an initial decision, not disclosing AI accuracy to participants, not giving them immediate feedback for each task instance, etc.). All of these factors can have an influence on people's reliance behavior, as shown by numerous empirical studies [6, 25, 32, 38]. Nevertheless, our results highlight that decision difficulty can be an important factor for how effectively explanations can help calibrating decision makers' reliance on AI. Our results further indicate that to improve human decision-making in more challenging tasks, different approaches than explaining how the AI came to its prediction may be necessary. Instead of introducing the secondary task of verifying the AI, which in difficult decisions may be hardly possible, AI systems could focus more on supporting the primary decision task. For instance, explanations could provide domain-specifc information that aligns with users' decision-making, instead of making transparent the inner working of the AI model. In a clinical setting, Yang et al. [40] explored supplementing AI recommendations with references to medical literature, similar to how clinicians validate suggestions from colleagues. Lim et al. [23] explained cardiac diagnosis predictions with murmur diagrams, which are a common representation for changes in heart sound loudness. Alternatively, Miller [26] proposed to circumvent the AI verification task by forgoing AI recommendations entirely. Instead, AI could help to explore and evaluate different hypotheses. In light of our results as well as previous work, we see both approaches, domain-specific explanations and alternatives to recommendation-centric AI, as promising ways forward to augment human decisions with AI in challenging domains.

6 CONCLUSION

We have conducted an online experiment in which we carefully selected task instances with a wide range of difficulties. We found that the effectiveness of feature-based explanations for reducing overreliance on AI can strongly depend on decision difficulty. Our results suggest that for easy decisions, explanations can reduce overreliance, likely because users can easily make the decision themselves and hence verify the AI using the explanations. For difficult decisions, explanations did not reduce overreliance, probably since users were uncertain of the answer themselves and were thus also not able to verify the AI. We therefore argue that explanations can be useful when automating easy, but high-volume decisions under human supervision. But to augment human decision-making in more challenging domains like medical diagnoses, other approaches than explaining the mechanisms behind AI recommendations may be necessary.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Montreal, Canada, 582:1–582:18. https://doi.org/10.1145/3173574.3174156

[2] 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 (CHI '21)*. ACM, Yokohama, Japan, 81:1–81:16. https://doi.org/10.1145/3411764.3445717

[3] Astrid Bertrand, James R. Eagan, and Winston Maxwell. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, Chicago, IL, USA, 943–958. https://doi.org/10.1145/3593013.3594053

[4] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics (ICHI 2015)*. IEEE, Dallas, TX, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[5] Zana Buçinca, Alexandra Chouldechova, Jennifer Wortman Vaughan, and Krzysztof Z. Gajos. 2022. Beyond end predictions: stop putting machine learning first and design human-centered AI for decision support. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–4.

[6] 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 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287

[7] Shiye Cao and Chien-Ming Huang. 2022. Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 471:1–471:23. https://doi.org/10.1145/3555572

[8] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Oct. 2023), 370:1–370:32. https://doi.org/10.1145/3610219

[9] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *Proceedings of the*

13th ACM Web Science Conference 2021 (WebSci '21). ACM, Virtual Event, United Kingdom, 120–129. https://doi.org/10.1145/3447535.3462487

[10] Giovanni Cinà, Tabea Röber, Rob Goedhart, and Ilker Birbil. 2022. Why we do need explainable AI for healthcare. https://doi.org/10.48550/arXiv.2206.15363 arXiv:2206.15363 [cs].

[11] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: a case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, Atlanta, GA, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[12] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, Glasgow, Scotland, UK, LBW0243:1–LBW0243:6. https://doi.org/10.1145/3290607.3312787

[13] Raymond Fok and Daniel S. Weld. 2023. In search of verifiability: explanations rarely enable complementary performance in AI-advised decision making. https://doi.org/10.48550/arXiv.2305.07722 arXiv:2305.07722 [cs].

[14] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. ACM, Helsinki, Finland, 794–806. https://doi.org/10.1145/3490099.3511138

[15] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. 2021. Reviewing the need for explainable artificial intelligence (xAI). In *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS-54)*. Computer Society Press, Virtual Event, 1284–1293. https://doi.org/10.24251/HICSS.2021.156

[16] 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 (Nov. 2019), 50:1–50:24. https://doi.org/10.1145/3359152

[17] 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 (June 2021), 108:1–108:9. https://doi.org/10.1038/s41398-021-01224-x

[18] Sean Koon. 2022. A human-capabilities orientation for human-AI interaction design. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–5.

[19] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI collaboration via conditional delegation: A case study of content moderation. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New Orleans, LA, USA, 54:1–54:18. https://doi.org/10.1145/3491102.3501999

[20] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a science of human-AI decision making: an overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, Chicago, IL, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[21] 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 (FAT* '19)*. ACM, Atlanta, GA, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[22] Philip Leifeld. 2013. texreg: conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software* 55, 8 (2013), 1–24. https://doi.org/10.18637/jss.v055.i08

[23] Brian Y. Lim, Joseph P. Cahaly, Chester Y. F. Sng, and Adam Chew. 2023. Diagrammatization: rationalizing with diagrammatic AI explanations for abductive-deductive reasoning on hypotheses. https://doi.org/10.48550/arXiv.2302.01241 arXiv:2302.01241 [cs].

[24] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 408:1–408:45. https://doi.org/10.1145/3479552

[25] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Yokohama, Japan, 78:1–78:16. https://doi.org/10.1145/3411764.3445562

[26] Tim Miller. 2023. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, Chicago, IL, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[27] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 2022. How accurate does it feel? – human perception of different types of classification mistakes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New Orleans, LA, USA, 180:1–180:13. https://doi.org/10.1145/3491102.3501915

[28] Alison Parkes. 2017. The effect of individual and task characteristics on decision aid reliance. *Behaviour & Information Technology* 36, 2 (Feb. 2017), 165–177. https://doi.org/10.1080/0144929X.2016.1209242 Publisher: Taylor & Francis.

[29] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. 2022. Investigations of performance and bias in human-AI teamwork in hiring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. Vancouver, BC, Canada, 12089–12097. https://doi.org/10.1609/aaai.v36i11.21468

[30] 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 (KDD '16)*. ACM, San Francisco, CA, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[31] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x Publisher: Nature Publishing Group.

[32] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. ACM, Oxford, United Kingdom, 617–626. https://doi.org/10.1145/3514094.3534128

[33] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, Sydney, NSW, Australia, 410–422. https://doi.org/10.1145/3581641.3584066

[34] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-AI collaboration: impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction (CD-MAKE 2020)*. Springer International Publishing, Dublin, Ireland, 431–449. https://doi.org/10.1007/978-3-030-57321-8_24

[35] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2023. On explanations, fairness, and appropriate reliance in human-AI decision-making. https://doi.org/10.48550/arXiv.2209.11812 arXiv:2209.11812 [cs].

[36] Timo Speith. 2022. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, Seoul, Republic of Korea, 2239–2250. https://doi.org/10.1145/3531146.3534639

[37] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 129:1–129:38. https://doi.org/10.1145/3579605

[38] Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, College Station, TX, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[39] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. ACM, Cagliari, Italy, 189–201. https://doi.org/10.1145/3377325.3377480

[40] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 14:1–14:14. https://doi.org/10.1145/3544548.3581393

[41] 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 (FAT* '20)*. ACM, Barcelona, Spain, 295–305. https://doi.org/10.1145/3351095.3372852

[42] Zelun Tony Zhang, Sven Tong, Yuanting Liu, and Andreas Butz. 2023. Is over-reliance on AI provoked by study design?. In *Lecture Notes in Computer Science (INTERACT 2023, Vol. 14144)*, José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, and Marco Winckler (Eds.). Springer, York, UK, 49–58. https://doi.org/10.1007/978-3-031-42286-7_3

A DECISION TASK INSTANCES

Table 1: Decision task instances with wrong AI predictions, ordered by decision difficulty, from easy to difficult.

Biography	Ground Truth	AI Prediction	Decision Difficulty
She received her Elementary education degree from Clemson University in 2002 and began her teaching career at West End Elementary School. She earned her master's degree from Southern Wesleyan University in 2006 and has been National Board Certified in literacy: Reading/Language Arts since 2009. ___ enjoys volunteering her time at Crosswell on the PTA Leadership Team and being the fundraising coordinator for March of Dimes and The American Cancer Society.	Teacher	Professor	0.055
He has taken that practical experience into his quality improvement work for population health and systems re-design across BC. He was the co-chair of the first chronic disease management collaborative in BC, the Congestive Heart Failure Collaborative. He is the physician lead for chronic disease management with Vancouver Coastal Health.	Physician	Professor	0.103
He is rated 5.0 stars out of 5 by his patients. He has indicated that his clinical interests include peripheral neuropathy, reconstructive surgery, and functional neurosurgery. Dr. ___ is affiliated with Massachusetts General Hospital. He is an in-network provider for Blue Cross/Blue Shield, Coventry, Humana ChoiceCare Network, and more. Dr. ___ has an open panel. He attended medical school at Stanford University School of Medicine. For his professional training, Dr. ___ completed a residency program at Massachusetts General Hospital. Dr. ___ is conversant in Hebrew.	Surgeon	Physician	0.167
His current research interests include the development of metacognitive capacity through individual psychotherapy for persons with schizophrenia. Related SubjectsPersonality Disorders in AdultsSchizophrenia & Other Psychotic Disorders in AdultsMood Disorders in Adults - Depression, Mania, Bi-polar	Psychologist	Professor	0.274
She enjoys developing relationships with her patients and following them throughout their lives from adolescence to adulthood. She strives to inform patients of their medical and surgical options and helps them to make decisions that best suit their needs.	Physician	Surgeon	0.369
She specializes in eating and food related issues, burnout, and stress reduction. Her comfortable office is located close to the Syracuse University campus. She also authors her monthly newsletter BreatheTasteSavor, as well as contributes to other popular publications.	Psychologist	Professor	0.375
Her teaching reflects this. She was trained at Harvard as a cognitive scientist & studied insight and discovery phenomena. She is currently Professor Emerita of Psychology and Computer Sciences at Rutgers University. She left Rutgers to teach dharma full-time.	Teacher	Psychologist	0.428

Biography	Ground Truth	AI Prediction	Decision Difficulty
He developed an interprofessional education anatomy course that brings together a variety of students in different Health Sciences professional programs, as well as students from other faculties. Of particular note is his effort in helping to design the Anatomy Learning Commons — an appealing learning environment that enhances the student experience by allowing them to engage with and learn from each other. ___ is applauded for his overall excellence in teaching, his national recognition for innovative interprofessional teaching methods and his multiple innovations to serve the needs of students.	Professor	Teacher	0.441
During her fellowship, she participated in the care of collegiate and professional athletes, and has experience in transplantation techniques for cartilage disorders and sports-related foot and ankle injuries. Dr. ___ has also published several research articles and book chapters on foot and ankle issues.	Surgeon	Professor	0.462
She attended Harvard University and the Feinberg School of Medicine at Northwestern University, and completed her surgical training at Yale University, the National Cancer Institute, and UCLA. She also served as a faculty member at UCLA, and was named the UCLA Outstanding Physician of the Year in 1999.	Surgeon	Physician	0.474
Currently, he is a teaching researcher at the Teaching Research Section of the Shanghai Municipal Education Commission. His responsibility is guiding mathematics teaching and mathematics teachers' professional development in the city. He is also a coauthor of a set of school mathematics textbooks for middle schools in Shanghai. His main research interests are school mathematics teaching and curriculum.	Teacher	Professor	0.475
His current research interests include relationship formation and maintenance through computer mediated communication channels (e.g., smartphone apps, social media, VoIP programs, and virtual reality) and artificial intelligence and the self. He received a MA in Sociology from the University of New Orleans and is currently a PhD candidate at Louisiana State University.	Psychologist	Professor	0.518
Her passion is to understand why people hold the views they do, the relationship between their views and their behavior, and how both can change over time. She earned a B.A. in psychology summa cum laude from Harvard College, and a Ph.D. in organizational behavior from Harvard Business School. ___ served as a professor of public policy at Harvard for 10 years, where she was awarded the Manuel Carballo Prize by the students. Her research on teams, networks, and reward systems has been published in such journals as Harvard Business Review, Organization Science, Academy of Management Executive, Small Group Research, and Social Justice Research.	Psychologist	Professor	0.554

Biography	Ground Truth	AI Prediction	Decision Difficulty
Through the centre that she founded, ___ has helped children – and their parents – identify and overcome their learning difficulties and challenges, using movement as one of the approaches. Movement therapist ___ met up with ___ for yet another insightful conversation on this.	Teacher	Psychologist	0.611

B PARTICIPANT DETAILS

Table 2: Participants' education.

Highest completed degree	With exp. (% in group)	No exp. (% in group)	Total (%)
Undergraduate	59 (60.82%)	70 (67.96%)	129 (64.5%)
Graduate	31 (31.96%)	30 (29.13%)	61 (30.5%)
PhD	5 (5.15%)	1 (0.97%)	6 (3%)
Other	2 (2.05%)	2 (1.94%)	4 (2%)

Table 3: Participants' machine learning experience.

Self-rated machine learning experience	With exp. (% in group)	No exp. (% in group)	Total (%)
None	71 (73.20%)	76 (73.79%)	147 (73.5%)
Basic	18 (18.56%)	17 (16.50%)	35 (17.5%)
Intermediate	7 (7.22%)	6 (5.83%)	13 (6.5%)
Advanced	1 (1.03%)	4 (3.88%)	5 (2.5%)
Expert	0 (0%)	0 (0%)	0 (0%)

Table 4: Participants' attitude toward AI.

Self-rated AI attitude	With exp. (% in group)	No exp. (% in group)	Total (%)
Negative	1 (1.03%)	1 (0.97%)	2 (1%)
Rather negative	9 (9.28%)	10 (9.71%)	19 (9.5%)
Neutral	43 (44.33%)	58 (56.31%)	101 (50.5%)
Rather positive	36 (37.11%)	25 (24.27%)	61 (30.5%)
Positive	8 (8.25%)	9 (8.74%)	17 (8.5%)

C MODELS

To fit the models, we combined some sparsely populated categories in participants' demographic data:

- Education: We combined *undergraduate* with *other*, and *graduate* with *PhD*.
- Machine learning knowledge: We combined *intermediate* with *advanced*.
- AI attitude: We combined *positive* with *rather positive* and *negative* with *rather negative*.

In another step, we further combined *(rather) positive* with *neutral* AI attitude since those two categories were highly correlated. The explanation condition was contrast-coded, with -0.5 for *with explanations* and 0.5 for *no explantions*. All of the following tables were created with texreg [22].

Table 5: Mixed-effects logistic regression models in Figure 5a and Figure 6a.

	Agreement	Accuracy
(Intercept)	−2.61 (0.55)***	2.83 (0.55)***
study_condition_contrast	1.10 (0.51)*	−1.03 (0.51)*
decision_difficulty	6.36 (0.66)***	−7.25 (0.68)***
user_gendermale	0.17 (0.19)	−0.25 (0.19)
user_age	−0.01 (0.01)	0.01 (0.01)
user_educationundergraduate/other	0.23 (0.19)	−0.24 (0.19)
user_ml_knowledgeintermediate/advanced	−0.29 (0.37)	−0.02 (0.37)
user_ml_knowledgenone	−0.11 (0.26)	0.10 (0.25)
user_ai_attitudenon-negative	0.08 (0.30)	−0.29 (0.30)
study_condition_contrast:task_difficulty	−2.22 (1.11)*	1.84 (1.14)
AIC	1221.98	1176.65
BIC	1285.78	1240.45
Log Likelihood	−597.99	−575.33
Num. obs.	1000	1000
Num. groups: user_ID	200	200
Var: user_ID (Intercept)	2.27	2.38
Var: user_ID task_difficulty	3.66	4.80
Cov: user_ID (Intercept) task_difficulty	−2.88	−3.38

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6: Mixed-effects logistic regression models in Figure 5b/Figure 5c and Figure 6b/Figure 6c.

	Agreem. easy	Agreem. medium	Agreem. hard	Acc. easy	Acc. medium	Acc. hard
(Intercept)	−3.19***	−1.03***	0.06	2.65***	0.28***	−0.67***
	(0.35)	(0.07)	(0.10)	(0.28)	(0.06)	(0.11)
study_conditionno_explanations	1.86***	0.95***	0.97***	−1.40***	−0.57***	−0.82**
	(0.35)	(0.14)	(0.24)	(0.31)	(0.14)	(0.26)
study_conditionwith_explanations	1.29***	0.73***	1.14***	−0.85*	−0.32*	−0.69**
	(0.39)	(0.14)	(0.25)	(0.35)	(0.14)	(0.26)
AIC	410.74	1971.83	779.91	464.49	2199.84	710.04
BIC	428.32	1993.34	797.50	482.08	2221.35	727.62
Log Likelihood	−201.37	−981.91	−385.96	−228.25	−1095.92	−351.02
Num. obs.	600	1600	600	600	1600	600
Num. groups: user_ID	200	200	200	200	200	200
Var: user_ID (Intercept)	0.96	0.00	0.00	0.73	0.01	0.00

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7: Mixed-effects linear regression models in Figure 7.

	AI consideration	Consideration to switch to agree with AI	Consideration to switch to disagree with AI
(Intercept)	2.38 (0.46)***	1.79 (0.66)**	2.49 (0.64)***
study_condition_contrast	−0.22 (0.23)	−0.64 (0.34)	0.38 (0.60)
decision_difficulty	0.22 (0.25)	0.58 (0.40)	0.54 (0.63)
user_gendermale	0.15 (0.18)	0.60 (0.24)*	−0.06 (0.22)
user_age	−0.00 (0.01)	−0.01 (0.01)	−0.00 (0.01)
user_educationundergraduate/other	−0.29 (0.17)	0.02 (0.23)	−0.39 (0.22)
user_ml_knowledgeintermediate/advanced	0.87 (0.37)*	0.68 (0.55)	−0.30 (0.48)
user_ml_knowledgenone	0.12 (0.22)	0.33 (0.30)	−0.38 (0.28)
user_ai_attitudenon-negative	0.66 (0.26)**	0.56 (0.35)	0.06 (0.36)
study_condition_contrast:task_difficulty	0.77 (0.49)	1.20 (0.80)	−0.54 (1.24)
AIC	970.51	629.34	505.00
BIC	1025.30	676.27	548.23
Log Likelihood	−471.26	−300.67	−238.50
Num. obs.	370	211	162
Num. groups: user_ID	91	76	78
Var: user_ID (Intercept)	0.34	0.76	0.27
Var: user_ID task_difficulty	0.04	0.33	0.01
Cov: user_ID (Intercept) task_difficulty	0.12	−0.34	0.05
Var: Residual	0.53	0.65	0.86

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$