Knowing About Knowing

An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems

He, Gaole; Kuiper, Lucie; Gadiraju, Ujwal

DOI
10.1145/3544548.3581025

Publication date
2023

Document Version
Final published version

Published in
CHI 2023 - Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems

Citation (APA)
He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *CHI 2023 - Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* Article 113 (Conference on Human Factors in Computing Systems - Proceedings). Association for Computing Machinery (ACM). https://doi.org/10.1145/3544548.3581025

Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems

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

Lucie Kuiper
Delft University of Technology
Delft, The Netherlands
lucieakuiper@gmail.com

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

ABSTRACT

The dazzling promises of AI systems to augment humans in various tasks hinge on whether humans can appropriately rely on them. Recent research has shown that appropriate reliance is the key to achieving complementary team performance in AI-assisted decision making. This paper addresses an under-explored problem of whether the Dunning-Kruger Effect (DKE) among people can hinder their appropriate reliance on AI systems. DKE is a metacognitive bias due to which less-competent individuals overestimate their own skill and performance. Through an empirical study (N = 249), we explored the impact of DKE on human reliance on an AI system, and whether such effects can be mitigated using a tutorial intervention that reveals the fallibility of AI advice, and exploiting logic units-based explanations to improve user understanding of AI advice. We found that participants who overestimate their performance tend to exhibit under-reliance on AI systems, which hinders optimal team performance. Logic units-based explanations did not help users in either improving the calibration of their competence or facilitating appropriate reliance. While the tutorial intervention was highly effective in helping users calibrate their self-assessment and facilitating appropriate reliance among participants with overestimated self-assessment, we found that it can potentially hurt the appropriate reliance of participants with underestimated self-assessment. Our work has broad implications on the design of methods to tackle user cognitive biases while facilitating appropriate reliance on AI systems. Our findings advance the current understanding of the role of self-assessment in shaping trust and reliance in human-AI decision making. This lays out promising future directions for relevant HCI research in this community.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Human-AI Decision Making, Appropriate Reliance, XAI, Dunning-Kruger Effect

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581025

ACM Reference Format:
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 (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581025

1 INTRODUCTION

In the last decade, powerful AI systems (especially deep learning systems) have shown better performance than human experts on many tasks, sometimes outperforming humans by a large margin [58, 87]. Attracted by the predictive capability of such AI systems, researchers and practitioners have started to adopt such systems to support human decision makers in critical domains (*e.g.,* financial [33], medical domains [48]). With the wish of complementary team performance, one goal of such human-AI collaboration is *appropriate reliance*: human decision makers rely on an AI system when it is accurate (or perhaps more precisely, when it is more accurate than humans) and do not rely on it when the system is inaccurate (or, ideally, whenever it is wrong). In such a collaborative decision process, human factors (*e.g.,* knowledge, mindset, cognitive bias) and the explanations for AI advice are important for trust in the AI system and for human reliance on the system. Several prior works have carried out empirical studies within this context of human-AI decision making, to explore the effectiveness of different kinds of explanations and the role of human factors in shaping such collaboration [3, 8, 24, 33, 52, 62, 79, 87].

In recent literature exploring human-AI interaction, researchers have shown a great interest in understanding what shapes user trust and reliance on AI systems. They found that factors like first impression [76], AI literacy [8], risk perception [33, 34], and performance feedback [55, 61] among others, play important roles in shaping human trust and reliance on AI systems. Explanations (*e.g.,* feature attribution of input) have been found to be useful in promoting human understanding and adoption of AI advice [3, 52, 79, 87] and He *et al.* [35] recently proposed analogies as an instrument to increase the intelligibility of explanations. However, prior studies observed improvements in performance in the presence of explanations only when the AI system outperformed both the human and the best team [3]. One reason for such phenomenon is under-reliance, which indicates humans do not rely on accurate AI predictions as often as it is ideal to [23, 79, 82]. In this work, we explore whether Dunning-Kruger effect (DKE) [43] – a metacognitive bias due to which individuals overestimate their competence and performance – affects user reliance on AI systems. This a particularly important metacognitive bias to understand in the context of human-AI decision making, since one can intuitively understand how inflated self-assessments and illusory superiority over an AI system can

result in overly relying on oneself or exhibiting under-reliance on AI advice. This can cloud human behavior in their interaction with AI systems. However, to the best of our knowledge no prior work has addressed this. In addition, DKE is closely related to user confidence in decision making, which has been identified as an important user factor and has been recently explored in the context of human-AI decision making [10, 33]. To achieve the goal of appropriate reliance, users are expected to adequately calibrate their self-confidence and their confidence in the AI system. Our work can lead to fundamental HCI insights that can help facilitate appropriate reliance of humans on AI systems.

To explore the impact of DKE on user reliance, we need to first identify participants who demonstrate the DKE (*i.e.,* participants who perform relatively poorly but overestimate their performance). According to existing research on the DKE [16, 18], the participants representing the bottom performance quartile tend to overestimate their skill and depict an illusory superiority, while those in the top performance quartile do not exhibit such a trend. Researchers have also operationalized self-assessments to serve as indicators of competence in different online tasks [29]. Informed by such prior work, we consider overestimated self-assessments in the context of human-AI decision making as an indicator of the DKE and explore it further. Through an explicit analysis of participants' performance in the bottom quartile, we verified that the overestimation in their performance is highly indicative of DKE in our study. In this scope, we explore whether we can design interventions to help users improve their own calibration of their skills in the task at hand.

Inspired by existing work in mitigating cognitive biases such as the DKE [43] and promoting appropriate reliance [8, 45, 79], we propose to leverage tutorials to calibrate their self-assessment through revealing the actual performance level of participants with performance feedback. In such a tutorial, after the initial decision making, participants are provided with correct answers and explanations to contrast with their final choice (if they make a wrong choice). As pointed out by existing research [17], one cause of DKE can be that people place too much confidence in the insightfulness of their judgments. When the correct answer differs from their own choice, they may refrain from trusting such ground truth in the absence of additional rationale. To ensure the effectiveness of revealing users' shortcomings, we provide them with contrastive explanations which point out not only the reason for correct answers, but also why their choice was incorrect. Based on prior work, we expect such a training session to help users realize their errors and calibrate their self-assessment. Furthermore, they become more skillful at the task, which is also highlighted by Kruger *et al.* [43] in mitigating DKE.

When AI advice disagrees with human decisions, the lack of rationales may be a reason not to adopt AI advice. To help participants interpret the AI advice, we leverage logic units-based explanations which reveal the AI system's internal states. When users recognize that an explanation provides reasonable evidence for supporting AI advice, it is much easier for them to resolve disagreement in their decision making. As a result, participants have a better opportunity to know and understand when they "should" in fact rely on AI systems. From this standpoint, effective explanations alongside the tutorial may help mitigate the impact of the Dunning-Kruger Effect on user reliance. To analyze the impact of DKE on user reliance on AI systems in this paper, we aim to find answers for the following two research questions:

> **RQ1:** How does the Dunning-Kruger Effect shape reliance on AI systems?
> **RQ2:** How can the Dunning-Kruger Effect be mitigated in human-AI decision making tasks?

To answer these questions, and based on existing literature, we proposed four hypotheses considering the effect of the overestimation of performance on (appropriate) reliance, the effect of the tutorial intervention on self-assessment calibration and reliance for participants with miscalibrated self-assessment, the effect of logic units-based explanations and tutorial intervention on reliance and team performance. We tested these hypotheses in an empirical study (N = 249) of human-AI collaborative decision making in a logical reasoning task (*i.e.,* multi-choice logical question answering based on a context paragraph). We found a negative impact of the DKE on human reliance behavior, where participants with DKE relied significantly less on the AI system than their counterparts without DKE. To mitigate such effects, we designed a tutorial intervention for making users aware of their miscalibrated self-assessment and provided logic units-based explanations to help explain AI advice. Although we found that the intervention tutorial was highly effective in improving participants' self-assessments, their improvement in appropriate reliance and performance is limited (statistically non-significant). Moreover, no obvious benefits were found with introducing logic units-based explanations in the logical reasoning task.

Our results highlight that the overestimation of performance will result in under-reliance, and such miscalibrated self-assessment can be improved with our proposed tutorial intervention. We also found that participants who overestimated their performance demonstrated an increased appropriate reliance, which the calibration of self-assessment can partially explain. However, this was in contrast to participants who initially underestimated their performance – while they calibrated their self-assessment, they achieved significantly worse appropriate reliance and performance. One potential cause is that such tutorials help them recognize their actual performance but also cause the illusion of superiority to AI systems. Such finding is also in line with algorithm aversion [12], where users are less tolerant of the mistakes made by AI systems. In addition, we found that the users' general propensity to trust goes a long way in shaping trust in AI systems, despite our tutorial not having an effect in reshaping subjective trust. Based on the results from our empirical study, we provide guidelines for designing more comprehensive user tutorials and point out promising future directions for further research around self-assessments in the context of human-AI decision making. Although we found that miscalibrated self-assessments may hinder appropriate reliance (*i.e.,* participants with DKE relied less on AI systems), the participants with accurate self-assessment did not necessarily show optimal appropriate reliance (*e.g.,* we found that participants with underestimation showed better appropriate reliance and performance). This interplay between self-assessment and reliance on AI

systems is potentially more complex than what can be explained by a linear relationship and, therefore, deserves further research.

In summary, we explored the effectiveness of a tutorial intervention to mitigate the DKE and, in turn, facilitate appropriate reliance. We found evidence suggesting its effectiveness through an empirical study in a logical reasoning task. Our work has important implications for HCI research in the realm of human-AI interaction. Our findings indicate that incorrect self-assessments and a prevalent meta-cognitive bias can affect user objective reliance on the AI system. Thus, while designing for optimal human-AI interaction, it is important to consider the extent to which users are aware of their own abilities and that of the AI system. Our work is an important first step towards furthering our understanding of how cognitive biases shape human reliance on AI systems, an understudied aspect in this quickly evolving realm of research. Considering the unique and evolving landscape of AI systems, the associated metaphors, and end-user expectations that are mediated through abstractions and their own experiences, we believe that studying the role of the DKE in the human-AI decision making context is a timely and unique contribution. We hope that our work can inform future research on designing human-AI interactions that can facilitate appropriate reliance on AI systems.

2 BACKGROUND AND RELATED WORK

This paper contributes to the growing literature on human-AI interaction, collaboration, and teaming, by exploring **how the Dunning-Kruger Effect shapes user reliance on AI systems** and **whether such effect can be mitigated with a user tutorial that highlights the fallibility of AI advice and logic units-based explanation**. Thus, we position our work in different strands of related literature: the general literature on AI-assisted decision making and what roles explanations play in such collaboration (2.1), more specific literature on promoting appropriate reliance (2.2), the contradicting literature on algorithm aversion and algorithm appreciation (2.3), and finally the literature on self-assessments, which has been explored in both psychology and other HCI studies (2.4).

2.1 Human-AI Collaborative Decision Making

In recent years, AI-assisted decision making has received more and more attention. In such collaboration, user factors and interaction with AI systems are observed to be of much impact on final user behaviors. Among these work, most researchers are interested in how users shape their trust in AI systems and how user behaviors will be affected by AI systems. Topics like performance feedback [2, 55], risk perception [27, 34], uncertainty [77] and confidence [10, 79, 87] of machine learning models, impact of explanations [3, 46] have been extensively studied in human-AI decision making. Meanwhile, fairness, accountability, and transparency of incorporating AI systems for collaborative decision making received more and more attention from a wide range of stakeholders [19, 21]. For a more comprehensive survey of existing work on Human-AI decision making, readers can refer to [44].

According to GDPR, the users of AI systems should have the right to access meaningful explanations of model predictions [68]. Under this perspective, more and more researchers have started to provide human-centered explainable AI (XAI) solutions to promote human-AI collaboration [20–22, 50, 78]. Up to now, the benefits of incorporating XAI methods in human-AI collaboration are still limited [3, 44]. As reported by most existing work, though XAI methods can aid understanding of AI advice, such effect does not necessarily lead to clear performance improvement [3, 52]. For instance, Liu *et al.* [52] observed that interactive explanations may "reinforce human biases and lead to limited performance improvement". Based on a comprehensive literature review, Wang *et al.* [79] proposed three desiderata of AI explanations to promote appropriate reliance: (1) critical for people to understand the AI, (2) recognize the uncertainty underlying the AI, and (3) calibrate their trust in the AI in AI-assisted decision making. With such ideal properties, effective explanations may also potentially help participant realize their weakness and mistake when they disagree with AI advice. Under this perspective, we also explored whether logic units-based explanations can help participants calibrate their self-assessment and promote appropriate reliance.

2.2 Empirical Studies on Appropriate Reliance

AI systems and human decision makers are supposed to achieve complementary team performance through taking advantage of both powerful predictive capability of AI systems and flexibility of human users to handle complex decision tasks. However, existing literature still struggles to find such complementary team performance — in most empirical studies, AI alone performs much better than human-AI team [44, 52]. With further analysis, researchers point out two main causes: (1) under-reliance, users fail to fully take advantage of powerful AI systems, and (2) over-reliance, users fail to rely on themselves when they actually outperform AI systems.

To promote appropriate reliance, existing research mainly focused on mitigating under-reliance and over-reliance. Different interventions like cognitive forcing functions [5], user tutorial [8, 9] and explanations [79] are proved to be highly effective in mitigating such unexpected reliance patterns. Buçinca *et al.* [5] introduced three types of cognitive functions to mitigate over-reliance: show AI advice on demand, update decision with AI advice after the initial decision, and keep participants waiting for a while before providing advice. Their experimental results indicate that such cognitive forcing functions are even more effective than simple XAI methods in mitigating over-reliance. With a comparative study of four types of different explanations, Wang *et al.* [79] reported that feature importance and feature contribution explanations can promote appropriate reliance with mitigating under-reliance.

"User tutorials, when presented in appropriate forms, can help some people rely on ML models more appropriately" [8]. Another important branch is educating users with user tutorials, which stands out in recent years. On one hand, such user tutorials make users aware of the weakness of AI systems, which further calibrate user trust and reliance on AI systems. For example, Chiang *et al.* [9] found that a brief education session (to increase people's awareness of the machine learning model's possible performance disparity on different data) can effectively reduce over-reliance on out-of-distribution data. On the other hand, such a system can educate participants with domain-specific knowledge extracted from an AI system, which further improves users' capability. As a typical example, Lai *et al.* [45] proposed model-driven tutorials to help

humans understand patterns learned by models in a training phase. Inspired by this series of research, we also explored whether DKE can be mitigated with user tutorial. For the purpose of calibrating self-assessment, we include performance feedback and explanations to contrast wrong user choice with correct answers.

2.3 Algorithm Aversion and Algorithm Appreciation

In the face of intelligent predictive agents, which may outperform human experts, people show two contradicting altitudes: *Algorithm Aversion* and *Algorithm Appreciation*. Compared to human forecasters, people more quickly lose confidence in AI systems after seeing them make the same mistakes [12]. Thus, some users are reluctant to use superior but imperfect algorithms [6]. Such a phenomenon is called "Algorithm Aversion," which has been observed across multiple domains, like moral decision making [31], economic bargains [23], medical diagnosis [54], and autonomous driving [14]. Burton *et al.* [6] summarized the cause and solution of algorithm aversion with five aspects: expectations and expertise, decision autonomy, incentivization, cognitive compatibility, and divergent rationalities. Meanwhile, Dietvorst *et al.* [13] found that such algorithm aversion can be overcome with the chance to modify algorithm advice. Readers can refer to two recent survey papers [6, 40] for a comprehensive literature review. In contrast, Logg *et al.* [53] found that users were influenced more by the algorithmic decision instead of human decision, and they first coined the notion of "Algorithm Appreciation" to describe such a phenomenon. Others revealed similar findings in contexts where tasks are perceived as being more objective [7], machines share rationale with humans [75] or with prior exposure to similar systems [42].

Besides contradicting attitudes towards the use of AI systems, prior work has shown how different human factors such as algorithmic literacy [72], expertise [53], and cognitive load [84] can affect users' final adoption of algorithmic advice. For example, users' algorithmic literacy [71–73] about fairness, accountability, transparency, and explainability is found to greatly affect their trust and privacy concern in adopting the advice from AI systems [70, 74]. Logg *et al.* [53] found that experts may even show more tendency to discount algorithmic advice when compared to laypeople. Furthermore, these factors can also affect the extent to which users show algorithm aversion or algorithm appreciation. For instance, You *et al.* [84] argue that algorithm appreciation declines when the transparency of the advice source's prediction performance further increases. In their study, they used a series of numbers instead of aggregated average performance, which increases the transparency of prediction performance. But they observed a decrease in algorithm appreciation, which was explained by the greater cognitive load imposed by the elaborated format. A recent work [37] found that the choice of framings of human agents and algorithmic agents may affect user perception of agent competence (*i.e.,* expert power), which further affects user behavior and cause inconsistent observations of algorithm aversion and algorithm appreciation. In this work, since we explore means to facilitate appropriate reliance of humans on AI systems, we position our findings in the context of the research breaching algorithm aversion and appreciation. Future work can further explore the role of algorithmic aversion and appreciation in the context of interventions to facilitate appropriate reliance on AI systems.

2.4 Self-assessment in HCI Studies

Evaluating one's own performance on a task, typically known as "self-assessment", is perceived as a fundamental skill, but people appear to calibrate their abilities [39] poorly. In general, most people tend to overestimate their own abilities. The cause of such an effect is multi-fold, like people tend to think they are above average and people place too much confidence in the insightfulness of their judgments [17]. With self-assessment, existing HCI research has explored using it as a measure for different purposes: Gadiraju *et al.* [29] used self-assessment for competence-based pre-selection in crowdsourcing marketplaces, Green *et al.* [33] measured users' risk assessment with comparing self-reported confidence with their actual performance, and Chromik *et al.* [11] compared perceived understanding of XAI methods with their actual understanding to reveal users' illusion of explanatory depth.

Dunning-Kruger effect (DKE) [43] described the dual burden the unskilled suffer from, besides the low performance, the unskilled will also lack the skill to estimate their own ability. Kruger *et al.* also found that a training session to increase the skills of participants is highly successful in mitigating such effect [43]. It had some positive effects and showed that by increasing knowledge, the overestimation could also be reduced. Further work also proved the effectiveness of such training in different domains like medicine [4] and economics [64].

Besides the popularity in psychology research, Dunning-Kruger effect was also studied in human-computer interaction field. In a recent study, Schaffer *et al.* [65] conducted a user study based on Diner's Dilemma game. They found that participants who considered themselves very familiar with the task domain showed more trust in an intelligent assistant but relied less on it. Presenting explanations was not as effective as expected, and sometimes even resulted in automation bias. Using logical reasoning tasks with varying difficulty levels, Gadiraju *et al.* [29] showed that online crowd workers also fall prey to the DKE. The authors proposed the use of self-assessments in a pre-selection strategy to improve quality-related outcomes. Informed by prior literature, we selected logical reasoning tasks as the exploratory lens to address our research questions since the tasks themselves are straightforward to understand for laypeople, but with increasing difficulty, they also create room for inviting AI advice. This serves suitably to study the DKE in the context of human-AI decision making.

3 METHOD AND HYPOTHESIS

In this section, we describe the logical reasoning task (*i.e.,* multi-choice logical question answering based on a context paragraph) and present our hypotheses.

3.1 Logical Reasoning Task

Prior work in the human-AI decision making context has explored how one can reliably study human behavior in proxy tasks. These work has established the importance of designing tasks, where users can find that there is a need to rely on AI (*e.g.,* owing to the

Figure 1: An example of a logical reasoning task used to obtain an initial human decision in the two-stage decision making process.

task difficulty or a perceivable benefit) and where there is a risk associated with such reliance (*e.g.,* dealing with an imperfect AI system) [5, 76]. This follows from the work of Lee and See [47] who defined trust in the Human-AI interaction context as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.*" The basis for our experimental setup is a task where participants are asked to choose an option in a multi-choice setting based on a paragraph of context presented to them (an example of the interface page is shown in Figure 1). We use the publicly available Reclor[1] [85] dataset to this end. The dataset corresponds to characteristically high difficulty of logical reasoning tasks and has been used in prior work exploring Human-AI team performance [3]. This task was chosen as a realistic scenario for human-AI collaboration, where humans incentivized to complete the task accurately, may have the capability to reason accurately and find the right answer, but may also evidently perceive a benefit in adopting AI advice. In addition, the Dunning-Kruger Effect which has been widely replicated in a variety of contexts has been shown to be prevalent in the domain of logical reasoning as well [16, 43].

In the basic setting of the task, participants are presented with three snippets of information: (1) a context paragraph, (2) a question related to this context, and (3) four different options corresponding to the question. Among the four options, a single option is deemed to be the best match to the question (*i.e.,* ground truth). Participants are asked to first go through the context paragraph, and then make a choice based on the question. This simulates a realistic scenario where participants make decisions in a reading comprehension setting. While humans are capable of handling such tasks, AI systems may outperform them by extracting useful information and dealing with complex reasoning structures which require a larger working memory capacity. The task interface is shown in Figure 2(a).

[1]https://whyu.me/reclor/

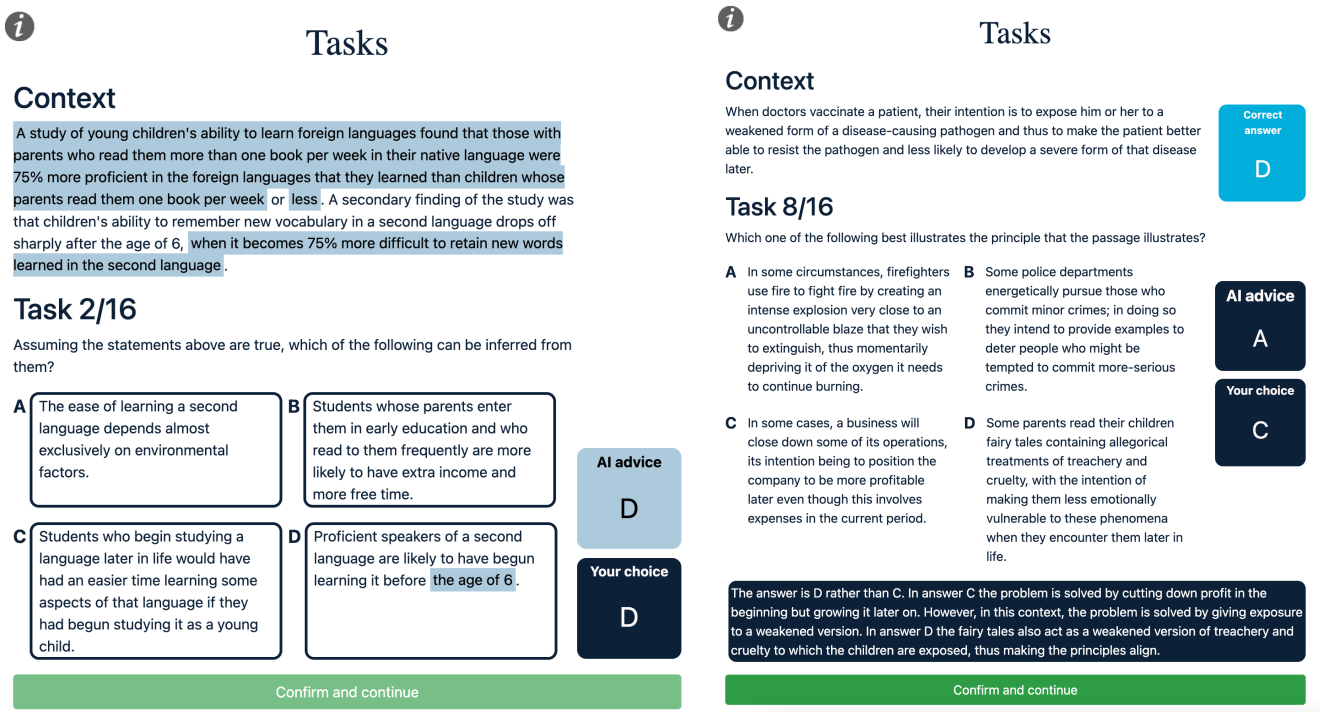
Two-stage Decision Making. To analyze human reliance on AI systems, all participants in our study worked on tasks with a two-stage decision making process. In the first stage, only task information was provided, and participants were asked to make decisions themselves (example shown in Figure 1). After that, we showed the same task with AI advice (and *explanations* depending on the experimental condition) and provided an opportunity for the participants to alter their initial choice. An example of second stage is shown in Figure 2(a), where "Your choice" shows the initial decision participants made in the first stage. This setup of an initial unaided decision and the presentation of advice from an AI system in order to make a second and final choice is similar to the update condition in [33], and in line with findings that people first make a decision on their own and only then decide whether to incorporate system advice [32]. It also fits with the research of Dietvorst *et al.* [13] on trust in two-stage decision making.

Quality Control. To ensure participant reliability and that participants worked on the logical reasoning tasks genuinely (*i.e.,* read the context paragraph and question carefully), we employed three attention check questions during the study process [56]. For this purpose, we embedded explicit instructions asking participants to select a specific option either in the context paragraph (once) or the question (twice). For example, we embedded the instruction, "*Confirm that you have read the context by selecting answer B.*" into a context paragraph on the task interface (which looks nearly identical to other tasks). A conservative estimate through trial runs reflected that participants would take at least 1 minute to complete each task. As a further quality control measure, we deactivated the submit button corresponding to each task page (including tasks in tutorial phase) for 30 seconds. Since attention check pages do not require deliberation, we reduced that time to 5 seconds.

3.2 Logic Units-based Explanations

In natural language processing tasks, feature attribution methods (*e.g.,* text highlights on input) are the most popular in existing literature. However, multiple pieces of research work point out that such token-level highlights are still hard to interpret [69, 81, 86]. Meanwhile, since logical reasoning tasks highlight the potential for logical reasoning congruent to human understanding, explanations based on logic units (*i.e.,* text spans) may be a better choice to reveal how AI systems reach their final decision. With this perspective, we drew inspiration from LogiFormer, proposed by Xu et al. [80], who conducted logical reasoning with logic units based on pre-trained language models to generate such explanations. LogiFormer adopted a graph transformer network for logical reasoning of logic units, where the logic units are text spans connected with causal relations. Following this interpretability design, we also relied on the self-attention matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ (n indicates the number of logic units) from the last layer of the graph transformer network and identified the important logic units with the following formula:

$$E = Argmax_k(\sum_{j=1}^{j=n} \mathbf{A}_{ij}), \tag{1}$$



(a) Logical question answering page with AI advice.

(b) Tutorial page with manual explanation.

Figure 2: Screenshots of the task interface. In panel (a), the logic units-based explanations are highlighted with a light blue background color in the context paragraph and the option suggested by the AI system. In panel (b), we show the rationale of correct answers in contrast with users' final choice at the bottom (when users do not select the correct answer in the decision making stage) at the bottom.

where E is the top-k logic units which receive most attention from other logic units (*i.e.,* we calculated it with the sum along each column of the self-attention matrix). One example of such explanation is shown in figure 2(a).

Our implementation and extracted logic units-based explanations can be found in Github repo.[2] To generate the explanations described above, we first trained the LogiFormer model on the Reclor dataset. With the trained model, we generated logic units-based explanations according to Equation 1. In this study, we specify $k = 5$ to highlight the most important logic units for each task. Notice that, such explanations are generated for each option, and the spans are only extracted from the context paragraph and each option. For more details about the LogiFormer model, we refer readers to the original paper [80] and the corresponding implementation.[3]

3.3 Proposing a Tutorial Intervention to Help Users Calibrate Their Skills

To answer RQ2, we need to verify whether our proposed intervention can help mitigate the DKE among the same participants who demonstrated it in the absence of the intervention. This requires two batches of tasks that can facilitate comparative performance assessment and on which participants can be asked to self-assess their performance. Based on the effectiveness of tutorials as interventions in previous HCI literature [8, 9, 45], we designed a tutorial as a means to shed light on the fallibility of AI advice. In our paper, we, therefore, considered the tutorial as an intervention and analyzed its effectiveness by comparing participants' reliance and self-assessment before and after the tutorial was delivered. Inspired by existing work to mitigate different kinds of cognitive biases through revealing such biases to users [1, 38], we decided to adopt a tutorial to help users calibrate their skills through self-assessment on logical reasoning tasks. To this end, we designed a tutorial with the aim of revealing to users that they may not be as capable in such tasks as they may believe. Furthermore, to ensure the effectiveness of revealing their mistakes, we designed persuasive explanations for users. To achieve that goal, we chose to provide contrastive explanations which point out not only the reason for correct answers but also the reason to reject users' wrong choices. As none of the existing off-the-shelf toolkits can be used to obtain such strongly persuasive explanations, we manually created explanations for each option in the four tasks considered in the tutorial phase. These explanations corresponding to each task have also been made available on the Open Science Framework companion page. An example of such performance feedback and contrastive explanation can be found in Figure 2(b). On this page, we showed

[2]https://github.com/RichardHGL/CHI2023_DKE
[3]https://github.com/xufangzhi/Logiformer

the correct answer in a box with light blue background color. The final decision of the participant after receiving AI advice, and the AI advice itself are shown in boxes with a dark blue background color. The contrastive explanation is shown at the bottom of this page. Through such a performance feedback intervention, we hope that users with inflated self-assessments can realize their true capability with respect to the tasks and recalibrate their self-assessment. Such an intervention can potentially help users improve their reliance on AI systems [36].

3.4 Pilot Study for Task Selection

To answer our research questions, we need to analyze the impact of the Dunning-Kruger effect on reliance measures and the effectiveness of the proposed intervention to mitigate such an effect. Note that the Dunning-Kruger effect corresponds to one's skills in a given task [16]. To operationalize this, we need two batches of tasks with similar difficulty levels, through which we can verify the effectiveness of the intervention by comparing performance before and after the intervention. Meanwhile, for the tutorial tasks, we need tasks that may trigger the Dunning-Kruger effect. In other words, tasks that participants may make mistakes on with high confidence. For these purposes, we conducted a pilot study with 10 participants from the Prolific crowdsourcing platform.[4] In the pilot study, each participant worked on 30 questions randomly sampled from the validation set of the Reclor dataset. We collected their choice and confidence level for each task. With six participants who passed all the attention checks, we assessed the difficulty of each task based on the number of participants who answered the task correctly. Considering that most participants spend around 1 minute to fully understand a task and make a decision, we considered the batch size to be six. We collected two batches of tasks which are of similar difficulty (informed by the average accuracy on the tasks in the pilot study). To make the tutorial effective but not cumbersome, we selected four tasks for the tutorial. The tasks for the tutorial were selected in a similar fashion as the other batches, as the tutorial only has four questions instead of six, the tasks with the lowest and highest accuracy were removed. Such selection strategy creates a batch similar in difficulty to the other batches. Among the four tasks, we configured the AI advice to be correct on two of them and misleading on the other two. All participants were rewarded with hourly wage of £7.5 (estimated completion time was 33 minutes), and extra bonus of £0.05 for each correct decision.

3.5 Hypotheses

Our experiment was designed to answer questions surrounding the impact of Dunning-Kruger effect on user reliance on AI systems, and how to mitigate such potentially undesirable impact. People who are less competent in a task struggle more with estimating their own performance in the task, compared to the more competent counterparts [43]. Impacted by DKE, users with the option to rely on AI advice may overestimate their own performance in a task and tend to rely on themselves when they are actually less capable than the AI systems. Apart from them, some users can exhibit accurate self-assessment. Such accurate self-assessments can be indicative of a good understanding of the task difficulty and personal skills, which may help these users rely on AI systems more appropriately. Meanwhile, effective explanations may amplify such an effect. Thus, we hypothesize that:

> **(H1)** Users overestimating their own performance will demonstrate relatively less reliance on AI systems than users demonstrating accurate self-assessment.

According to previous work [26, 57], interventions that provide users with feedback on their performance may help improve their self-assessment. By providing users with an opportunity to reflect on their skills and recalibrate their skills on the given task, we argue that the impact of the DKE can be mitigated. As a result of an improved calibration of oneself, such users are better suited to rely on AI systems appropriately when making decisions. Therefore, we hypothesize that:

> **(H2)** Making users aware of their miscalibrated self-assessment, will help them improve their self-assessment.
>
> **(H3)** Making users aware of their miscalibrated self-assessment will result in relatively more appropriate reliance on AI systems.

Performance feedback can potentially help participants improve their self-assessment, which may facilitate appropriate reliance. At the same time, explanations have been shown to improve the human understanding and interpretation of AI advice [3, 52, 79], which can also potentially contribute to appropriate reliance. Thus, we hypothesize to observe the following in a human-AI decision making context:

> **(H4)** Providing performance feedback and meaningful explanations can facilitate appropriate reliance on the AI system.

4 STUDY DESIGN

This section describes our experimental conditions, variables, statistical analysis, procedure, and participants in our main study.

4.1 Experimental Conditions

In our study, all participants worked on logical reasoning tasks with two-stage decision making process (described in Sec. 3.1). The only difference is whether tutorial is presented and whether explanations are provided along with AI advice. To comprehensively study the effect of each factor and their interaction effect, we considered a 2×2 factorial design with four experimental conditions: (1) no tutorial, no XAI (represented as × **Tutorial,** × **XAI**), (2) with tutorial, no XAI (represented as ✓ **Tutorial,** × **XAI**), (3) no tutorial, with XAI (represented as × **Tutorial,** ✓ **XAI**), (4) with tutorial, with XAI (represented as ✓ **Tutorial,** ✓ **XAI**). In conditions with tutorial, participants were presented with four selected tasks with performance feedback and contrastive explanation for correct answers against wrong choice (when participants missed the wrong answer). While in conditions without tutorial, the four

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

Table 1: The different appropriate reliance patterns considered in [66]. d_i is initial human decision, while d_f is the final decision after AI advice.

d_i	AI advice	d_f	Reliance
Incorrect	Correct	Correct	Positive AI reliance
Incorrect	Correct	Incorrect	Negative self-reliance
Correct	Incorrect	Correct	Positive self-reliance
Correct	Incorrect	Incorrect	Negative AI reliance

tasks selected are presented as normal tasks without any performance feedback or explanation for correct answers, to prevent any learning effect. In conditions with XAI, the top-5 most important logic units are highlighted as an explanation for AI advice.

For each batch of six tasks, the AI system was configured to provide correct advice on four of them and misleading advice on two tasks. So the accuracy of AI systems is around 66.7%. To avoid any ordering effect, we randomly assign one batch of tasks as first batch of tasks for each participant and further shuffled the order of tasks within each batch.

4.2 Measures and Variables

We measure the reliance of participants on the AI system via two 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. They are commonly used in the literature, for example in [83, 87]. In addition, we consider the accuracy in batches to measure participants' performance with AI assistance. Since cases without initial disagreement do not clearly signal reliance on the system we restrict the scope of the appropriate reliance measure to accurately understand how participants handle divergent system advice. Max *et al.* [66] presented four conditions of appropriate reliance patterns (see Table 1) when the disagreement exists and correct answer exists in human initial decision or AI advice. We followed them to adopt *Relative positive AI reliance* (**RAIR**) and *Relative positive self-reliance* (**RSR**) as appropriate reliance measures. The two measures assessed users' appropriate reliance from two dimensions, which can help analyze the dynamics of reliance. To provide an overview of participants' appropriate reliance under initial disagreement, we considered **Accuracy-wid** (*i.e.,* accuracy with initial disagreement). These measures are computed as follows:

$$\textbf{Agreement Fraction} = \frac{\text{Number of decisions same as the system}}{\text{Total number of decisions}},$$

$$\textbf{Switch Fraction} = \frac{\text{Number of decisions user switched to agree with the system}}{\text{Total number of decisions with initial disagreement}},$$

$$\textbf{Accuracy} = \frac{\text{Number of correct final decisions}}{\text{Total number of decisions}},$$

$$\textbf{Accuracy-wid} = \frac{\text{Number of correct final decisions with initial disagreement}}{\text{Total number of decisions with initial disagreement}},$$

$$\textbf{RAIR} = \frac{\text{Positive AI reliance}}{\text{Positive AI reliance + Negative self-reliance}},$$

$$\textbf{RSR} = \frac{\text{Positive self-reliance}}{\text{Positive self-reliance + Negative AI reliance}}.$$

To measure the self-assessment of users, we gathered responses on the following question after each batch of tasks – "From the previous 6 questions, how many questions do you estimate to have been answered correctly? (after receiving AI advice)". Comparing that estimation with the actual correct number, we can calculate the degree of miscalibration and self-assessment as: **Degree of Miscalibration** = |Estimated correct number - Actual correct number|, **Self-assessment** = Estimated correct number - Actual correct number. Meanwhile, for conditions with explanations, we also assessed the helpfulness of explanations with the question, "To what extent was the explanation (*i.e.,* the highlighted words/phrases) helpful in making your final decision?" Responses were gathered on a 5-point Likert scale from *1* to *5* corresponding to the labels *not helpful*, *very slightly helpful*, *slightly helpful*, *helpful*, *very helpful*.

For a deeper analysis of our results, a number of additional measures were considered based on observations from existing literature [49, 67, 76]:

- Trust in Automation (TiA) questionnaire [41], a validated instrument to measure (subjective) trust [76]. In this study we adopted two subscales: *Propensity to Trust* (TiA-PtT), *Trust in Automation* (TiA-Trust). Thus, we consider possible effects of trust on reliance, in accordance with Lee *et al.* [47].
- Affinity for Technology Interaction Scale (ATI) [28], administered in the pre-task questionnaire. Thus, we account for the effect of participants' affinity with technology on their reliance on systems [76].

Table 2 presents an overview of all the variables considered in our study.

4.3 Participants

Sample Size Estimation. Before recruiting participants, we computed the required sample size in a power analysis for the 2×2 factorial design using G*Power [25]. To correct for error-inflation as a result of testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{4} = 0.0125$. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.0125$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and the consideration of 4 different experimental conditions. This resulted in a required sample size of 244 participants. We thereby recruited 314 participants from the crowdsourcing platform Prolific[5], in order to accommodate potential exclusion.

Compensation. All participants were rewarded with £2.5, amounting to an hourly wage of £7.5 (estimated completion time was 20 minutes). We rewarded participants with extra bonuses of £0.1 for every correct decision in the 16 trial cases. By incentivizing participants to reach a correct decision, we operationalize the concomitant "vulnerability" discussed by Lee and See [47] as a contextual requirement to encourage appropriate system reliance.

Filter Criteria. All participants were proficient English speakers above the age of 18 and they had an approval rate of at least 90% on the Prolific platform. We excluded participants from our analysis if they failed at least one attention check (65 participants). The resulting sample of 249 participants had an average age of 38 (SD = 12.8) and a gender distribution (48.6% female, 51.4% male).

[5]https://www.prolific.co

Table 2: The different variables considered in our experimental study. "DV" refers to the dependent variable. RAIR, RSR, and Accuracy-wid are indicators of appropriate reliance.

Variable Type	Variable Name	Value Type	Value Scale
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
	Accuracy-wid	Continuous	[0.0, 1.0]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous	[0.0, 1.0]
	RAIR	Continuous	[0.0, 1.0]
	RSR	Continuous	[0.0, 1.0]
Assessment (DV)	Degree of Miscalibration	Continuous, Interval	[0,6]
	Self-assessment	Continuous, Interval	[-6,6]
Trust (DV)	TiA-Trust	Likert	5-point, 1:strong distrust, 5: strong trust
Covariates	ATI	Likert	6-point, 1: low, 6: high
	TiA-PtT	Likert	5-point, 1: tend to distrust, 5: tend to trust
Other	Helpfulness of Explanation	Likert	5-point, 1: not helpful, 5: very helpful

4.4 Procedure

The full procedure that participants followed in our study is illustrated in Figure 3. All participants first read the same basic instructions on the logical reasoning task. Next, participants were asked to complete a pre-task questionnaire to measure their propensity to trust and affinity for technology interaction.

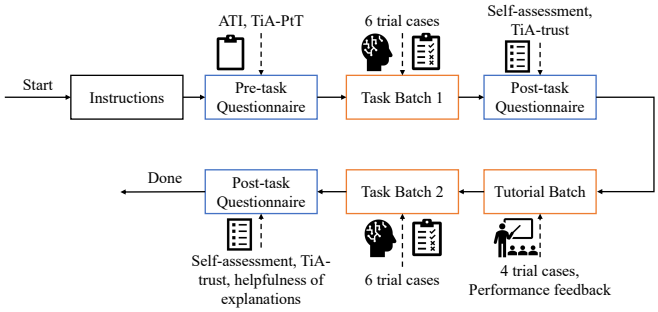


Figure 3: Illustration of the procedure participants followed within our study. This flow chart describes the experimental condition ✓ `Tutorial`, ✓ `XAI`. Blue boxes represent the questionnaire phase, orange boxes represent the task phase.

Participants were then assigned to one experimental condition, which differed in whether or not tutorial feedback is provided and the system's prediction is supplemented with explanation. In × `Tutorial`, × `XAI` and × `Tutorial`, ✓ `XAI` conditions, participants worked on the four trial cases without any difference with the task batch, no extra information was provided. After that, participants will work on 16 tasks (two task phases with six tasks, and one tutorial phase with four tasks). Selection of these cases is described in section 3.4. After each task phase, post-task questionnaires were adopted to assess their self-assessment and trust in AI systems (TiA-trust). Participants in the × `Tutorial`, ✓ `XAI` and ✓ `Tutorial`, ✓ `XAI` conditions were additionally asked for their perceived helpfulness of the explanations they were presented with. To further ensure the reliability of responses gathered in the questionnaire and the task phases, we added four attention check questions spread out at random through the different stages of the procedure [30].

5 RESULTS

In this section, we present the results of our study. We discuss descriptive statistics, the outcomes of the hypothesis tests we conducted, and our exploratory findings. Our code and data can be found on Github.[6]

5.1 Descriptive Statistics

In our analysis, we only kept participants who passed all attention checks, which deemed to be more reliable. Participants were distributed in a balanced fashion over the four experimental conditions as follows: 63 (× `Tutorial`, × `XAI`), 62 (✓ `Tutorial`, × `XAI`), 62 (× `Tutorial`, ✓ `XAI`), 62 (✓ `Tutorial`, ✓ `XAI`). On average, participants spend around 32 minutes (SD = 11 minutes) in our study. We found no significant difference in the time spent across the four experimental conditions.

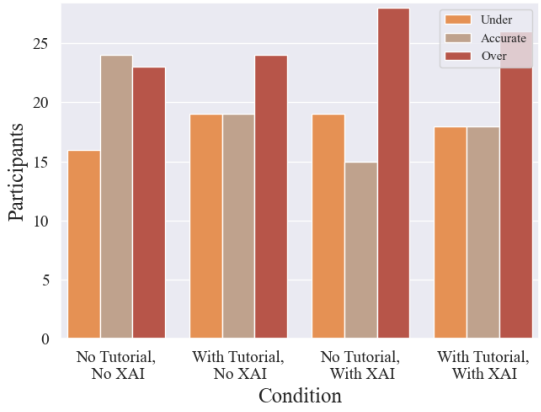


Figure 4: Distribution of participants with underestimated, accurate, and overestimated self-assessment across all experimental conditions in the first batch of tasks.

Distribution of Covariates. The covariates' distribution is as follows: *ATI* (M = 3.73, SD = 0.99, 6-point Likert scale, and *1: low, 6:*

[6] https://github.com/RichardHGL/CHI2023_DKE

Table 3: Kruskal-Wallis H-test results for inflated self-assessments (H1) on reliance-based dependent variables. "††" indicates the effect of variable is significant at the level of 0.0125. "Under", "Accurate", abd "Over" refers to participants who underestimated , accurately estimated, and overestimated their performance on the first batch of tasks, respectively.

Dependent Variables	H	p	$M \pm SD$(Under)	$M \pm SD$(Accurate)	$M \pm SD$(Over)	Post-hoc results
Accuracy	74.06	**<.001**††	0.72 ± 0.16	0.61 ± 0.15	0.45 ± 0.19	Under > Accurate > Over
Agreement Fraction	10.87	**.004**††	0.70 ± 0.18	0.69 ± 0.21	0.59 ± 0.24	Under, Accurate > Over
Switch Fraction	23.31	**<.001**††	0.50 ± 0.28	0.53 ± 0.31	0.32 ± 0.32	Under, Accurate > Over
Accuracy-wid	87.94	**<.001**††	0.65 ± 0.21	0.53 ± 0.27	0.28 ± 0.22	Under > Accurate > Over
RAIR	46.91	**<.001**††	0.65 ± 0.36	0.58 ± 0.37	0.27 ± 0.33	Under, Accurate > Over
RSR	30.23	**<.001**††	0.67 ± 0.44	0.41 ± 0.47	0.27 ± 0.43	Under > Accurate, Over

high), *TiA-Propensity to Trust* ($M = 2.95$, $SD = 0.60$, 5-point Likert scale, *1: tend to distrust, 5: tend to trust*).

Distribution of Participants. Among 249 participants, we identified the participants who underestimated their performance (*i.e.,* Self-assessment < 0), those with an accurate self-assessment (*i.e.,* Self-assessment = 0), and those with overestimation of their performance (*i.e.,* Self-assessment > 0) according to their performance in the first batch of tasks (shown in Figure 4). In general, participants showed relatively balanced distribution into the three types of self-assessment across conditions: (1) the number of participants with underestimated self-assessment lies in the range of 15 ∼ 20, (2) the number of participants with accurate self-assessment lies in the range of 15 ∼ 25, (3) the number of participants with overestimated self-assessment was in the range of 20 ∼ 30. We also compared the time spent by participants with different self-assessment and participants with different experimental conditions, and found no statistically significant difference with Kruskal-Wallis H-tests.

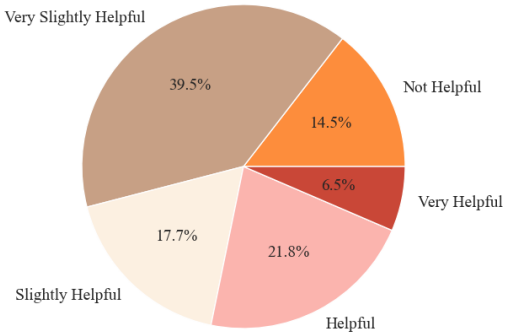


Figure 5: Distribution of participants with perceived helpfulness of logic units-based explanations.

For participants in conditions with explanation (*i.e.,* [× `Tutorial`, ✓ `XAI`] and [✓ `Tutorial`, ✓ `XAI`]), we assessed the helpfulness of logic units-based explanations. The ratios of perceived helpfulness are illustrated with Figure 5. As we can see, most people (57.2%) think it slightly or very slightly helpful, while only 28.3% participants show positive feedback to the logic units-based explanations.

Performance Overview. On average across all conditions, participants achieved an accuracy of 56.9% ($SD = 0.16$) over the two batches of tasks, still lower than the aforementioned AI accuracy of 66.7%. The agreement fraction is 0.665 ($SD = 0.17$) while the switching fraction is 0.453 ($SD = 0.27$). With these measures, we confirm that when disagreement appears participants in our study did not always switch to AI advice or blindly rely on the AI system. As all dependent variables are not normally distributed, we used non-parametric statistical tests to verify our hypotheses.

5.2 Hypothesis Tests

5.2.1 ***H1****: effect of inflated self-assessments on AI system reliance.* To analyze the main effect of participants' inflated self-assessment (*i.e.,* overestimation of performance) on their reliance on the AI system, we conducted Kruskal-Wallis H-tests by considering how participants varied in their self-assessment. We categorize all participants into three groups according to the self-assessment: (1) participants who underestimated their performance (*i.e.,* Self-assessment < 0), (2) participants with accurate performance self-assessment (*i.e.,* Self-assessment = 0), and (3) participants who overestimated their performance (*i.e.,* Self-assessment > 0). For this analysis, we considered all participants across the four experimental conditions, and the performance metrics are calculated based on the first batch of tasks (*i.e.,* 6 tasks). The results are shown in Table 3.

Effect of Overestimated Self-Assessments on Objective Reliance. For all reliance-based measures, we found a statistically significant difference between the performance of the participants who overestimated their performance and those with accurate self-assessment. Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) were used to make pairwise comparisons of performance, revealing that participants who did not overestimate their performance in fact performed significantly better than those who did (The only exception is on metric **RSR**). Overall, participants with accurate self-assessment and underestimation of their own performance performed much better than participants who overestimated their own performance. The main reason is that they showed more reliance on the AI system and achieved better appropriate reliance when their initial decision disagreed with AI advice. The results indicate that participants who overestimate their own performance rely significantly less on AI systems compared to those who do not. Due to such under-reliance and inappropriate reliance when initial disagreement exists, they achieved a significantly lower accuracy on average. Thus, we find support for hypothesis **H1**.

We also found that participants who underestimated their performance achieved significantly higher **Accuracy**, **Accuracy-wid**,

Table 4: Wilcoxon signed ranks test results for H3 on reliance-based dependent variables. For participants with initial underestimation, we report results with one-sided hypothesis that the performance / reliance decrease after tutorial. For participants with initial overestimation, we report results with one-sided hypothesis that the performance / reliance increase after tutorial. "†" and "††" indicates the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Participants	Underestimation					Overestimation				
Dependent Variables	T	p	$M \pm SD$(first)	$M \pm SD$(second)	Trend	T	p	$M \pm SD$(first)	$M \pm SD$(second)	Trend
Accuracy	407.5	**.000**††	0.73 ± 0.17	0.55 ± 0.21	↓	303.0	.075	0.46 ± 0.18	0.51 ± 0.22	-
Agreement Fraction	212.5	.543	0.68 ± 0.20	0.70 ± 0.23	-	451.0	.605	0.60 ± 0.23	0.57 ± 0.23	-
Switch Fraction	267.5	.592	0.47 ± 0.29	0.48 ± 0.36	-	367.5	.147	0.31 ± 0.33	0.36 ± 0.31	-
Accuracy-wid	418.0	**.000**††	0.68 ± 0.22	0.44 ± 0.29	↓	338.0	**.013**†	0.27 ± 0.20	0.41 ± 0.28	↑
RAIR	313.0	**.006**††	0.68 ± 0.37	0.45 ± 0.38	↓	194.0	**.038**†	0.24 ± 0.32	0.36 ± 0.36	↑
RSR	204.0	**.000**††	0.72 ± 0.43	0.30 ± 0.44	↓	151.0	**.020**†	0.29 ± 0.45	0.52 ± 0.48	↑

and **RSR** than participants demonstrating accurate self-assessment. Since they showed similar degrees of reliance (**Agreement Fraction** and **Switch Fraction**) on the AI system, the improvement of overall accuracy is mainly due to appropriate reliance. In general, they showed significantly better **RSR**, which indicates that they have a better chance to rely on themselves to make correct decisions when they initially disagree with misleading AI advice.

In the first batch of tasks, we found no difference (with Kruskal-Wallis H-tests) in reliance and accuracy metrics when comparing participants in XAI conditions (*i.e.,* × `Tutorial`,✓ `XAI` and ✓ `Tutorial`,✓ `XAI`) with participants in non-XAI (*i.e.,*× `Tutorial`,× `XAI` and ✓ `Tutorial`,× `XAI`). To verify how the provided logic units-based explanations affect participants with different self-assessments, we compared the performance and reliance measures of participants with XAI and without XAI in underestimation, accurate self-assessment, and overestimation. No significant effects were found from the logic units-based explanation on performance and reliance for participants with overestimated self-assessment.

5.2.2 *H2: effect of the tutorial on self-assessment.*

To verify **H2**, we used Wilcoxon signed rank tests to compare the performance of participants before and after the tutorial. We considered participants who are provided with the tutorial for self-assessment calibration (*i.e.,* ✓ `Tutorial`,× `XAI` and ✓ `Tutorial`,✓ `XAI`). Meanwhile, we exclude participants who have accurate assessment on the first batch of tasks from this analysis. Finally, we have 87 participants reserved for analysis of **H2**. On average, the participants' self-assessment get improved after receiving the tutorial (*i.e.,* decreased **Degree of Miscalibration**, $M \pm SD$(first) = 1.67 ± 0.91, $M \pm SD$(second) = 1.14 ± 1.04; a smaller value indicates more accurate self-assessment). A Wilcoxon signed rank test indicated that the difference was statistically significant, T=1175.0, p<0.001, which supports **H2**. To further check how the tutorial intervention has an impact on participants with different types of miscalibration, we separately conducted Wilcoxon signed rank tests on participants underestimating their own performance and overestimating their own performance separately. The results indicate that: (1) participants underestimating their own performance calibrated their self-assessment, the difference is significant (T=229.0, p=0.002); (2) participants overestimating their own performance calibrated their self-assessment, the difference is significant (T=381.5, p=0.012). The detailed analysis of participants with different types of miscalibration also supports **H2**.

To further explore the effect of logic units-based explanation on calibrating self-assessment, we conducted a Kruskal-Wallis H-test (among these participants) by considering whether the explanation is provided. We found no significant results, which indicates that logic units-based explanations cannot amplify the effect of the tutorial intervention (*i.e.,* calibrating self-assessment).

5.2.3 *H3: effect of the tutorial on appropriate reliance.*

Similar to the analysis for **H2**, we only considered the participants who showed miscalibration in the first batch of tasks. Overall, there is no significant difference in reliance and performance measures when we compare the participants' performance before and after receiving the tutorial. To further check how our tutorial intervention will affect participants with different miscalibration of self-assessment, we conducted analysis for participants with underestimation and overestimation separately. The results of Wilcoxon signed rank tests corresponding to each of the reliance measures are shown in Table 4. Both participants with underestimation and overestimation did not show any significant difference in reliance measures (*i.e.,* **Agreement Fraction** and **Switch Fraction**). For participants who underestimated their performance in the first batch of tasks, they showed significantly worse performance and appropriate reliance after receiving the tutorial. In contrast, we found some improvement of **Accuracy** and appropriate reliance measures (*i.e.,* **Accuracy-wid**, **RAIR**, **RSR**) for participants who overestimated their performance in the first batch of tasks. However, the improvement is non-significant at the level of 0.0125. Thus, on the whole, we find partial support for **H3**.

Meanwhile, to check how the tutorial intervention affects the participants with initial accurate self-assessment, we also conducted Wilcoxon signed rank tests for their performance before and after the tutorial intervention. No significant difference is found. Combined with the findings from participants with initial miscalibration, we found that: (1) the designed tutorial intervention does not show much impact on participants with accurate self-assessment, (2) the designed tutorial intervention has positive impact on appropriate reliance for participants who initially overestimate themselves, while negative impact on participants with initial underestimation of their performance.

Relation Between Self-assessment Calibration and the Change in Reliance. To further explore the relationship between the change in self-assessment and change with (appropriate) reliance, we conducted the Spearman rank-order test separately for participants

with overestimation and underestimation in the first batch of tasks. As the impact of tutorial intervention on **Agreement Fraction** and **Switch Fraction** is insignificant, we ignore the two metrics in calculating the correlation. The results are shown in Table 5. We found a strong negative monotonic relationship between the two variables in participants with overestimation. Thus, in logical reasoning tasks, the calibration effect in self-assessment accounted for 59.3% of the improved **Accuracy** ($\rho^2 = 0.593, p < 0.001$), 55.5% of the improved **Accuracy-wid** ($\rho^2 = 0.555, p < 0.001$), 32.0% of the improved **RAIR** ($\rho^2 = 0.320, p < 0.001$), and 12.9% of the improved **RSR** ($\rho^2 = 0.129, p = 0.005$). Similarly, the calibration of self-assessment also accounted for 26.2% of the decreased **Accuracy** ($\rho^2 = 0.262, p = 0.001$), 14.8% of the decreased **Accuracy-wid** ($\rho^2 = 0.148, p = 0.009$) for participants with underestimation.

Table 5: Correlation of self-assessment change and reliance change. "††" indicates the effect of variable is significant at the level of 0.0125. "†" indicates the effect of variable is significant at the level of 0.05.

Participants	Underestimation		Overestimation	
Dependent Variables	ρ	p	ρ	p
Accuracy	-0.512	**.001**††	-0.770	**.000**††
Accuracy-wid	-0.385	**.009**††	-0.745	**.000**††
RAIR	-0.293	**.039**†	-0.566	**.000**††
RSR	-0.349	.068	-0.359	**.005**††

In general, for all participants with miscalibrated self-assessment, the difference in self-assessment shows strong negative correlation with the difference in performance and appropriate reliance. In other words, the increase in self-assessment (trend to overestimation) will lead to decrease in performance and appropriate reliance, which is consistent with our findings in **H1**. While the significant negative correlation exists for performance measures in all participants with miscalibrated self-assessment, only participants with overestimation showed significant correlation (in the level of 0.0125) with **RAIR** and **RSR**. The difference indicates that the change of self-assessment can hardly explain why participants with underestimation showed worse appropriate reliance.

To further explore the impact of logic units-based explanations on performance improvement (the difference between performance metrics from the second batch of tasks and those from the first batch of tasks), we conducted a Kruskal-Wallis H-test (among these participants) by considering whether explanations are provided. Overall, no significant difference is found for all behavior-based dependent variables considering all 87 participants who showed miscalibration in the first batch and then received the tutorial intervention. We further check the logic units-based explanation impact according to participants with underestimation (37 participants) and overestimation (50 participants) respectively (cf. Table 6). No significant difference is found for all behavior-based dependent variables. Although participants with explanations show better performance improvement in **RSR**, such difference is not significant.

5.2.4 ***H4****: Two-factor analysis for final performance.*
To verify H4, we conducted a two-way ANOVA to compare the performance and (appropriate) reliance measures of participants under the effect of providing tutorial intervention and logic units-based explanations. In this analysis, only the second batch of tasks are taken into consideration, as the performance of the first batch of tasks is not affected by the tutorial intervention. According to the test results shown in Table 7, no significant impact (in the significance level of 0.0125) is found for tutorial intervention, logic units-based explanations and their interaction effect. Thus, **H4** is not supported.

According to the results of **H3**, the tutorial intervention shows positive impact on participants with initial overestimation, no significant effect on participants with accurate self-assessment, and negative impact on participants with initial underestimation. As indicated by Figure 4, the participants show compatible distribution in the three groups with different initial self-assessment. The contradicting effects on the participants with miscalibrated self-assessment get canceled. That may explain why the tutorial intervention does not show significant impact across experimental conditions. On the other hand, we did not find any support for effectiveness of logic units-based explanations in reliving DKE or facilitating appropriate reliance in analysis of **H1** - **H3**.

5.3 Further Analysis On the DKE

According to Dunning and Kruger [43], participants demonstrating the DKE are less competent and overestimate their performance. For further analysis of DKE in our study, we follow the method in the original study as well as consequent replications [29, 43], to split the participants in all conditions into performance-based quartiles. The top-quartile corresponds to those demonstrating high performance (top 25%), the bottom quartile corresponds to those with low performance (bottom 25%), and we combine the two quartiles in the middle comprising of participants with a medium level of performance in the first batch of tasks. As our tutorial is demonstrated to be effective in calibrating self-assessment, we do not take the second batch of tasks into consideration. In total, 101 participants among 249 participants showed an overestimation of performance in the first batch of tasks. In high accuracy group (63 participants), 35 participants showed underestimation of their own performance, and 21 participants demonstrated accurate self-assessment, while only 7 participants (11.1%) show overestimation of performance in the first batch of tasks. In comparison, 46 participants (73.0%) in low accuracy group (63 participants) show an overestimation of performance in the first batch of tasks, while only 6 participants and 11 participants showed underestimation of their performance and demonstrated accurate self-assessment, respectively. This aligns with the observation of Dunning and Kruger [16, 18]: top-performance group shows the tendency to underestimate their performance, while low-performance group shows tendency to overestimate their performance. With this observation, we can take low accuracy group as a representative group of participants with DKE, and take high accuracy group as a representative group of participants without DKE. This aligns with and validates our motivation to design a tutorial intervention to mitigate DKE, and improve self-assessment and appropriate reliance on AI systems.

The impact of DKE on Reliance. To further analyze how the DKE affects user reliance on AI systems, we compared the reliance-based measures of high accuracy group and low accuracy group

Table 6: Kruskal-Wallis H-test results for logic units-based explanations on performance improvement of reliance-based dependent variables.

Participants	Underestimation				Overestimation			
Dependent Variables	H	p	$M \pm SD$(Exp)	$M \pm SD$(No Exp)	H	p	$M \pm SD$(Exp)	$M \pm SD$(No Exp)
Accuracy	0.00	.963	-0.19 ± 0.15	-0.18 ± 0.24	1.38	.241	0.10 ± 0.27	0.00 ± 0.30
Agreement Fraction	0.00	.963	0.01 ± 0.25	0.04 ± 0.32	0.88	.349	0.01 ± 0.38	-0.06 ± 0.28
Switch Fraction	0.04	.843	-0.03 ± 0.39	0.04 ± 0.41	0.02	.884	0.06 ± 0.47	0.05 ± 0.33
Accuracy-wid	0.00	.951	-0.25 ± 0.30	-0.22 ± 0.31	0.50	.478	0.16 ± 0.36	0.11 ± 0.39
RAIR	0.02	.878	-0.23 ± 0.48	-0.24 ± 0.57	0.00	.968	0.11 ± 0.46	0.14 ± 0.46
RSR	0.96	.327	-0.33 ± 0.50	-0.50 ± 0.51	1.84	.175	0.35 ± 0.72	0.10 ± 0.66

Table 7: ANOVA test results for H4 on behavior-based dependent variables in the second batch of tasks.

Dependent Variables	Accuracy		Agreement Fraction		Switch Fraction		Accuracy-wid		RAIR		RSR	
Variables	F	p	F	p	F	p	F	p	F	p	F	p
Tutorial	2.41	.122	3.74	.054	3.87	.050	1.63	.203	4.70	.031	0.20	.652
XAI	2.10	.148	0.30	.587	1.00	.319	3.35	.068	2.05	.153	0.23	.632
Tutorial × XAI	0.05	.824	0.00	.990	0.00	.956	0.10	.746	0.00	.923	0.05	.832

Table 8: Kruskal-Wallis H-test results for reliance-based measures on high accuracy group and low accuracy group. "††" indicates the effect of variable is significant at the level of 0.0125.

Dependent Variables	H	p	$M \pm SD$(High)	$M \pm SD$(Low)
Agreement Fraction	54.68	**<.001**††	0.75 ± 0.15	0.46 ± 0.18
Switch Fraction	13.09	**<.001**††	0.46 ± 0.32	0.27 ± 0.21
Accuracy-wid	81.00	**<.001**††	0.74 ± 0.24	0.21 ± 0.15
RAIR	25.71	**<.001**††	0.64 ± 0.45	0.21 ± 0.21
RSR	46.41	**<.001**††	0.76 ± 0.39	0.18 ± 0.37

using a Kruskal-Wallis H-test. The results are shown in Table 8. Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) also confirmed the significant difference. As we can see, participants in the low accuracy group (representative for participants with DKE) achieve a relatively poorer appropriate reliance than participants in the high accuracy group. Participants in the low accuracy group demonstrate significantly less reliance and appropriate reliance on AI systems, which also reflects that under-reliance is to blame for their low performance. We also compared the time spent by participants in the high accuracy group with participants in low accuracy group through a Kruskal-Wallis H-test. The difference of time spent on tasks between the two groups is non-significant ($p = 0.018$, borderline significance in Kruskal-Wallis H-test). On average, the high accuracy group spent around 30 minutes (SD=12 minutes), while the low accuracy group spent around 34 minutes (SD=13 minutes). Interestingly, despite the fact that participants in the low accuracy group spent longer time on the task they still relied poorly on the AI system. This is consistent with what has been widely understood as an impact of the DKE metacognitive bias.

5.4 Further Analysis of Trust

In addition to the behavior-based reliance measures, we also assessed the subjective trust of participants in AI systems. In this subsection, we explore the impact of our tutorial intervention and logic units-based explanation on user trust in the AI system.

The effect of tutorial intervention on trust. To explore whether our tutorial intervention had any effect on user trust in AI system, we conducted Wilcoxon signed ranks test comparing the trust before and after the tutorial. On average, participants' trust in the AI system does not show significant difference after the tutorial intervention (increased from 2.996 to 3.016; $T = 1063.5$, $p = 0.952$). This suggests that the main impact of the tutorial was on helping users calibrate their competence (*i.e.,* their self-assessment) without directly shaping their trust in the AI system.

Table 9: ANCOVA test results on trust-related dependent variables. With different self-assessmnet patterns, we divide all participants into three groups. "††" indicates the effect of variable is significant at the level of 0.0125.

Variables	F	p	η^2
Group	1.15	.318	.009
ATI	1.22	.271	.004
TiA-PtT	10.21	**.002**††	.040

To further analyze how other covariates shape user trust in AI system, we decided to conduct AN(C)OVAs despite the anticipation that our data may not be normally distributed because these analyses have been shown to be robust to Likert-type ordinal data [60]. As no significant difference is found between the trust before and after the tutorial, we aggregated the trust across the two batches of tasks as users' trust in the AI system. Considering our main hypothesis, we aimed to explore whether overestimation of performance and accurate self-assessment shape user trust in the AI system. For that purpose, we consider the three groups of participants (based on self-assessment, the same criteria in **H1**) with different self-assessment patterns. The results are shown in Table 9. As we can see, propensity to trust was the only user factor which corresponded to a significant impact on **TiA-Trust**. In a further Spearman rank-order test, we

observed that there is a significant positive correlation between **TiA-PtT** and **TiA-Trust**, $\rho(249) = 0.22$, $p < .001$; suggesting a weak linear relationship between users' propensity to trust an AI system and the subjective trust measured with respect to the AI system in our study. We also conducted the Spearman rank-order tests with **TiA-PtT** and other reliance-based variables. No significant correlation was found between **TiA-PtT** and reliance measures.

6 DISCUSSION

6.1 Key Findings

Our analysis of the impact of miscalibrated self-assessment on reliance suggests that participants with DKE tend to overestimate their own competence and rely less on AI systems, which results in under-reliance and much worse performance. To mitigate such cognitive bias, we introduced a tutorial intervention including performance feedback on tasks, alongside manually crafted explanations to contrast the correct answer with the users' mistakes. Experimental results indicate that such an intervention is highly effective in calibrating self-assessment (significant improvement), and has some positive effect on mitigating under-reliance and promoting appropriate reliance (non-significant results). We also note that after making participants who overestimated their performance aware of their miscalibrated self-assessment, participants tend to rely more (appropriately) on the AI system (*i.e.,* increased **Switch Fraction** and appropriate reliance measures, non-significant results, from Table 4) and achieve a higher performance improvement when logic units-based explanations are provided (insignificant results from Table 6). However, we did not find any significant evidence to support that the logic units-based explanations can amplify the effect of the tutorial intervention in calibrating self-assessment, or relieving the impact of DKE.

The tutorial and calibrated self-assessment demonstrate a positive impact in facilitating appropriate reliance for participants who overestimated themselves, but an opposite trend was observed on participants who underestimated themselves. We found such difference can be explained partially by the change of self-assessment. The calibration of overestimation can bring positive impact, while the calibration of underestimation may also turn into overestimation or algorithm aversion, which may explain the decrease in performance and appropriate reliance. The tutorial was initially designed to reveal the shortcomings of participants with DKE. While for participants without DKE, there is a risk that some participants did not get exposed to their shortcomings in this tutorial and only found the AI system also made mistakes, which in turn even caused overestimation of themselves. An alternative explanation is that the performance feedback in tutorial intervention showed one mistake from the AI system, which led to algorithm aversion. As pointed out by [12]: "people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake." These findings advance our current understanding of human-AI decision making, and provide useful insights that can drive guidelines for designing interventions to promote appropriate reliance.

Positioning in Existing Literature. In our study, we found that DKE can have a negative impact on user reliance on the AI system and our proposed tutorial intervention can mitigate such an impact. In the context of human-AI decision making, DKE is closely relevant to a popular stream of research around user confidence[10, 33]. For the participants who overestimated their performance, the designed tutorial intervention calibrated their self-confidence (as reflected in their self-assessment) and facilitated appropriate reliance. In contrast, the negative impact on participants who underestimated their performance can be explained by: (1) the calibrated self-assessment which can also bring over-confidence, or (2) their confidence/trust in the AI system being eroded by the observed mistake(s) of the AI system [3, 76]. The latter is consistent with findings in the literature on algorithm aversion [12]. More empirical studies are required to confirm and explain these observations, breeding promising grounds for future research.

The participants with DKE show under-reliance on AI systems, which also aligns with the finding from Schaffer *et al.* [65]. Authors found that participants who reported higher familiarity with the task domain relied less on the intelligent assistant. The effectiveness of our tutorial intervention to calibrate self-assessment and mitigate under-reliance is also consistent with existing work using user tutorial / education interventions to mitigate unexpected and undesirable reliance patterns. All these tutorial interventions share a common objective of changing the mindset of users. For example, Chiang *et al.* [8] reported that user tutorials such as machine learning literacy interventions can effectively help high-performance individuals to reduce over-reliance without affecting the reliance of low-performance individuals. Similarly, Chiang *et al.* [9] showed that a brief education session about the possible performance disparity of an ML model (on data with different distribution) can effectively reduce over-reliance on such cases. While their work focused more on changing human understanding of AI systems (performance, uncertainty, etc.), our work aims to help users calibrate their competence (*i.e.,* their self-assessment) on specific tasks. As a result, their main objective was to realize when AI systems are not reliable to reduce over-reliance, while we attempt to mitigate under-reliance for participants who overestimate themselves.

Logic Units-based Explanations Do Not Have the Expected Effect. In our study, the logic units-based explanations did not aid in further amplifying the calibration effect of the tutorial intervention. This is in line with the findings of Wang *et al.* [79] and Schaffer *et al.* [65]. With a comparative study about four types of different explanations, authors found that "on decision making tasks that people are more knowledgeable, explanation that is considered to resemble how humans explain decisions (*i.e.,* counterfactual explanation) does not seem to improve calibrated trust." One potential explanation is that such explanations do not fulfill the three desiderata of AI explanations [79] (refer to section 2.1): the logic units-based explanations may help participants understand the AI, but fail to help them recognize the uncertainty underlying the AI or calibrate their trust in the AI in AI-assisted decision making. Another potential cause is such explanations may introduce automation bias [65], which will cause over-reliance. Our results suggest that logic units-based explanations may still be hard to follow, because participants still need to connect and interpret the logic units by themselves. A limitation of our current work is that we did not gather explicit input from participants on their perceived understanding of the explanations. One further step to ground such logic units into readable logical claims may work better for users. However, we do not

deny the prospect that some XAI methods may have the potential to help mitigate DKE and calibrate user confidence in human-AI decision making. For example, contrastive explanations may work in the context of human-AI decision making [51, 59].

6.2 Implications

As our findings suggest that participants with DKE tend to rely less on AI systems, it implies that future work should look more closely at the effects of self-assessment in human-AI collaboration. Although our tutorial intervention shows significant improvement in calibrating self-assessment, the improvement in appropriate reliance is still limited (with borderline significance). Meanwhile, such calibration of self-assessment may even hurt the team performance for participants with initial underestimation of their performance. For these participants, the tutorial calibrated their underestimation, which may also lead to illusion of superior performance (overestimation of themselves). In order to further promote appropriate reliance in human-AI collaboration, we need to develop more effective human-centered tutorials. Meanwhile, participants who show lower performance in our scenario have significantly higher probability to overestimate their performance, which aligns with DKE properties. Thus, we can leverage overestimation of individual performance as an indicator of such a meta-cognitive bias and further mitigate it with personalized or appropriate interventions.

Guidelines for Tutorial Designs to Promote Appropriate Reliance. While our tutorial intervention proved to be effective in helping users calibrate their self-assessment, accurate self-assessment does not necessarily translate to optimal appropriate reliance. Compared with participants with accurate self-assessment, the participants with underestimation showed a significantly better performance in **RSR** (see Table 3), and calibrating such underestimation may even lead to decreased appropriate reliance (see Table 4), which indicates accurate self-assessment does not necessarily lead to optimal appropriate reliance. One possible cause is that while the tutorial makes such users aware that they underestimated themselves and they can make correct decisions when the AI system is wrong in the task, users may have an illusion of superior capability than the AI system. As a result, on some tasks where AI systems are more capable, users make mistakes by exhibiting under-reliance on the AI system due to recalibrated overestimation of their own competence. Our findings suggest that we should pay attention to avoiding such side effects of making users overestimate themselves in comparison to the AI system. To avoid such side effects, tutorials designed to mitigate a specific kind of bias should be carefully checked before subjecting them to broad participant pools. This also implies that tutorials designed for promoting appropriate reliance should not only reveal the shortcomings of users or AI systems (*i.e.,* when they are less capable of making the right decision), but also their strengths (*i.e.,* when they are capable or more capable). This has useful implications for the future design of interventions to mitigate cognitive biases in human-AI decision making.

In previous work on mitigating over-reliance with a tutorial intervention, researchers focused on revealing the AI systems' brittleness [8, 9]. Combined with their findings, we argue that a more effective tutorial to promote appropriate reliance can be one that helps users understand both themselves and AI systems, and not only revealing the weakness but also showing the strengths of each. With such a comprehensive understanding, human decision makers can potentially have a better chance to understand when they should rely on AI systems, and when they should rely on themselves, ultimately leading to (more) appropriate reliance. More work is required to understand whether and how explanations can mediate this process of creating a better understanding among users of AI system capabilities in comparison to their own. This resonates with recent work exploring human-AI complementarity [3, 44, 52].

6.3 Caveats and Limitations

Potential Biases. Our research questions focused on DKE and reliance and how to mitigate such impact. As we cannot pre-identify which participants have DKE, we recruit the participants and determine it with performance assessment. However, such assessment may be affected by other factors, which can lead to biased results. For example, although we relied on a pilot study to inform our task selection while creating two batches of tasks with comparable difficulty levels, we cannot be certain that they would be perceived the same way on average across the participants.

As pointed out by Draws *et al.* [15], cognitive biases introduced by task design and workflow may have a negative impact on crowdsourcing experiments. With the help of Cognitive Biases Checklist introduced [15], we analyzed potential bias in our study. *Self-interest bias* is possible, because crowd workers we recruited from the Prolific platform are motivated by monetary compensation. To alleviate any participants with low effort results, we put attention checks to remove ineligible participants from our study. As the question and context in Reclor dataset may be something participants familiar with, *familiarity bias* and *availability bias* can also affect our results.

Transferability Concern. In our study, all analyses are based on the logical reasoning task, which most laypeople are capable of dealing with. However, in practice, the application scenarios may be affected by more factors (like user expertise, familiarity, and input modality). This gap can be a potential threat to the transferability of our findings and implications. However, Dunning and Kruger [43] showed that participants suffer from DKE across multiple scenarios: "participants scoring in the bottom quartile on tests of humor, grammar, and logic grossly overestimated their test performance and ability." These effects were replicated in a number of other tasks, like human-AI collaboration [65] and crowdsourcing [63, 76]. Our findings are therefore highly relevant and can play an important role in informing the design for appropriate reliance in the context of human-AI interaction, collaboration, and teaming.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we present a quantitative study to understand the impact of the Dunning-Kruger effect (DKE) on reliance behavior of participants in a human-AI decision making context. We propose a tutorial intervention and explore its effectiveness in mitigating such an effect. Our results suggest that participants who overestimate their own performance tend to rely less on the AI system. Combined with the findings that participants with DKE show a much higher probability of overestimating their performance, we conclude that participants with DKE rely less on AI systems, and such

under-reliance hinders them in achieving better performance on average (RQ1). Through a rigorous experimental setup and statistical analysis, we found the effectiveness of our tutorial intervention in mitigating DKE (RQ2). However, we found that the tutorial may mislead some participants (*i.e.,* participants who underestimated themselves) to overestimate their performance or exhibit algorithm aversion, which in turn harms their appropriate reliance on the AI system. Our findings suggest that, to fully mitigate the negative impact of the Dunning-Kruger effect and achieve appropriate reliance, more comprehensive, insightful, and personalized user tutorials are required. We reflected on guidelines for better tutorial designs based on our key findings.

We found that our tutorial intervention failed to make a difference in participants' subjective trust in the AI systems. Instead, we found that users' general propensity to trust has a significant impact on shaping their subjective trust in the AI system. Future work can further look into how user trust can be reshaped with different interventions or by using more effective explanations (*e.g.,* contrastive explanations or logical explanations in natural language). We hope the key findings and implications reported in this work will inspire further research on promoting appropriate reliance.

ACKNOWLEDGMENTS

This work was partially supported by the TU Delft Design@Scale AI Lab and the 4TU.CEE UNCAGE project. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-3888. We thank all participants from Prolific.

REFERENCES

[1] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.
[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] Cynthia Sherraden Bradley, Kristina Thomas Dreifuerst, Brandon Kyle Johnson, and Ann Loomis. 2022. More than a Meme: The Dunning-Kruger Effect as an Opportunity for Positive Change in Nursing Education. *Clinical Simulation in Nursing* 66 (2022), 58–65.
[5] 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.
[6] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239.
[7] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.
[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] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022), 107018.
[11] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.
[12] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.
[13] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.
[14] Nicole Dillen, Marko Ilievski, Edith Law, Lennart E Nacke, Krzysztof Czarnecki, and Oliver Schneider. 2020. Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[15] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 48–59.
[16] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*. Vol. 44. Elsevier, 247–296.
[17] David Dunning, Chip Heath, and Jerry M Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest* 5, 3 (2004), 69–106.
[18] Joyce Ehrlinger, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. 2008. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes* 105, 1 (2008), 98–121.
[19] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
[20] 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.
[21] 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.
[22] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
[23] 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.
[24] 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.
[25] 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.
[26] Jennifer Fereday and Eimear Muir-Cochrane. 2006. The role of performance feedback in the self-assessment of competence: a research study with nursing clinicians. *Collegian* 13, 1 (2006), 10–15.
[27] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.
[28] 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.
[29] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 4 (2017), 1–26.
[30] 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.
[31] Jan Gogoll and Matthias Uhl. 2018. Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics* 74 (2018), 97–103.
[32] 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.
[33] 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.
[34] 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).
[35] 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.
[36] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
[37] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
[38] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[39] Rachel A Jansen, Anna N Rafferty, and Thomas L Griffiths. 2021. A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour* 5, 6 (2021), 756–763.
[40] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion. In *28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020, Marrakech, Morocco, June 15-17, 2020*, Frantz Rowe, Redouane El Amrani, Moez Limayem, Sue Newell, Nancy Pouloudi, Eric van Heck, and Ali El Quammah (Eds.).
[41] 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.
[42] Max F Kramer, Jana Schaich Borg, Vincent Conitzer, and Walter Sinnott-Armstrong. 2018. When do people want AI to make decisions?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 204–209.
[43] 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.
[44] 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).
[45] 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.
[46] 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.
[47] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[48] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[49] 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.
[50] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).
[51] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements* 27 (1990), 247–266.
[52] 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.
[53] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[54] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650.
[55] 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.
[56] Catherine C Marshall and Frank M Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*. 234–243.
[57] Conor Thomas McKevitt. 2016. Engaging students with self-assessment and tutor feedback to improve performance and support assessment capacity. *Journal of University Teaching & Learning Practice* 13, 1 (2016), 2.
[58] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.
[59] Tim Miller. 2021. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review* 36 (2021).
[60] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15, 5 (2010), 625–632.
[61] 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.
[62] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233.
[63] Ioannis Petros Samiotis, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju, and Alessandro Bozzon. 2021. Exploring the Music Perception Skills of Crowd Workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 108–119.
[64] James Sawler. 2021. Economics 101-ism and the Dunning-Kruger effect: Reducing overconfidence among introductory macroeconomics students. *International Review of Economics Education* 36 (2021), 100208.
[65] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251.
[66] 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)*.
[67] 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.
[68] Andrew Selbst and Julia Powles. 2018. "Meaningful Information" and the Right to Explanation. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 48.
[69] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. 2021. Knowledge-intensive language understanding for explainable AI. *IEEE Internet Computing* 25, 5 (2021), 19–24.
[70] Donghee Shin. 2022. Expanding the role of trust in the experience of algorithmic journalism: User sensemaking of algorithmic heuristics in Korean users. *Journalism Practice* 16, 6 (2022), 1168–1191.
[71] Donghee Shin. 2022. How do people judge the credibility of algorithmic sources? *AI & Society* 37, 1 (2022), 81–96.
[72] Donghee Shin, Kerk F Kee, and Emily Y Shin. 2022. Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? *International Journal of Information Management* 65 (2022), 102494.
[73] Donghee Shin, Azmat Rasul, and Anestis Fotiadis. 2021. Why am I seeing this? Deconstructing algorithm literacy through the lens of users. *Internet Research* (2021).
[74] Donghee Shin, Bouziane Zaid, Frank Biocca, and Azmat Rasul. 2022. In Platforms We Trust? Unlocking the Black-Box of News Algorithms through Interpretable AI. *Journal of Broadcasting & Electronic Media* (2022), 1–22.
[75] Donghee Shin, Bouziane Zaid, and Mohammed Ibahrine. 2020. Algorithm appreciation: Algorithmic performance, developmental processes, and user interactions. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 1–5.
[76] 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.
[77] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020), 100049.

[78] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[79] 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.

[80] Fangzhi Xu, Jun Liu, Qika Lin, Yudai Pan, and Lingling Zhang. 2022. Logiformer: A Two-Branch Graph Transformer Network for Interpretable Logical Reasoning. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 1055–1065.

[81] Hanqi Yan, Lin Gui, and Yulan He. 2022. Hierarchical Interpretation of Neural Text Classification. *arXiv preprint arXiv:2202.09792* (2022).

[82] 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.

[83] 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.

[84] Sangseok You, Cathy Liu Yang, and Xitong Li. 2022. Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems* 39, 2 (2022), 336–365.

[85] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations (ICLR)*.

[86] Muhammad Bilal Zafar, Philipp Schmidt, Michele Donini, Cédric Archambeau, Felix Biessmann, Sanjiv Ranjan Das, and Krishnaram Kenthapadi. 2021. More Than Words: Towards Better Quality Interpretations of Text Classifiers. *arXiv preprint arXiv:2112.12444* (2021).

[87] 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.