Dunning-Kruger Effect Can Hinder Appropriate Reliance on AI Systems

GAOLE HE, Delft University of Technology, The Netherlands
LUCIE KUIPER, Delft University of Technology, The Netherlands
UJWAL GADIRAJU, Delft University of Technology, The Netherlands

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**.

ACM Reference Format:
Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Dunning-Kruger Effect 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, 13 pages. https://doi.org/978-1-4503-XXXX-X/18/06

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 [29, 40]. 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 [17], medical domains [24]). 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

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

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

© 2023 Copyright held by the owner/author(s).

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/978-1-4503-XXXX-X/18/06

explanations for AI advice are important for trust in the AI system and for human reliance on the system [1, 2, 10, 17, 26, 31, 35, 40].

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 [34], AI literacy [2], risk perception [17, 18], and performance feedback [27, 30] 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 [1, 26, 35, 40]. 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 [1]. 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 [9, 35, 37]. In this work, we explore whether Dunning-Kruger effect (DKE) [21] – 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 [3, 17].

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 [6, 8], 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 [14]. 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 [21] and promoting appropriate reliance [2, 22, 35], 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 [7], 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.* [21] 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 [4], where users are less tolerant of the mistakes made by AI systems. 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.

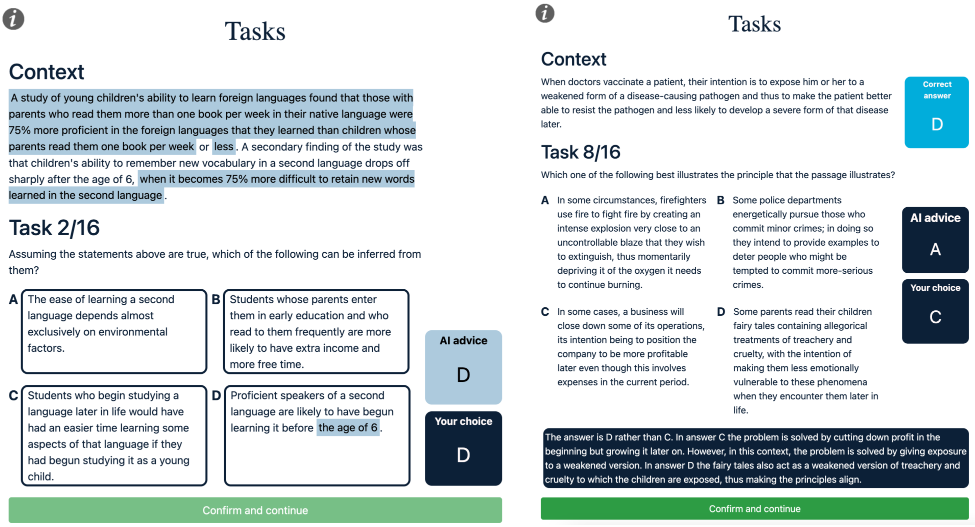
2 STUDY DESIGN

This section describes our experimental setup, variables, procedure, and participants in our main study.

2.1 Logical Reasoning Task

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(a)). We use the publicly available Reclor[1] [39] 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 [1]. 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 [6, 21].

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 1(a).



Tasks

Context

A study of young children's ability to learn foreign languages found that those with parents who read them more than one book per week in their native language were 75% more proficient in the foreign languages that they learned than children whose parents read them one book per week or less. A secondary finding of the study was that children's ability to remember new vocabulary in a second language drops off sharply after the age of 6, when it becomes 75% more difficult to retain new words learned in the second language.

Task 2/16

Assuming the statements above are true, which of the following can be inferred from them?

A The ease of learning a second language depends almost exclusively on environmental factors.

B Students whose parents enter them in early education and who read to them frequently are more likely to have extra income and more free time.

C Students who begin studying a language later in life would have had an easier time learning some aspects of that language if they had begun studying it as a young child.

D Proficient speakers of a second language are likely to have begun learning it before the age of 6.

AI advice: D

Your choice: D

Confirm and continue

(a) Logical question answering page with AI advice.

Tasks

Context

When doctors vaccinate a patient, their intention is to expose him or her to a weakened form of a disease-causing pathogen and thus to make the patient better able to resist the pathogen and less likely to develop a severe form of that disease later.

Correct answer: D

Task 8/16

Which one of the following best illustrates the principle that the passage illustrates?

A In some circumstances, firefighters use fire to fight fire by creating an intense explosion very close to an uncontrollable blaze that they wish to extinguish, thus momentarily depriving it of the oxygen it needs to continue burning.

B Some police departments energetically pursue those who commit minor crimes; in doing so they intend to provide examples to deter people who might be tempted to commit more-serious crimes.

C In some cases, a business will close down some of its operations, its intention being to position the company to be more profitable later even though this involves expenses in the current period.

D Some parents read their children fairy tales containing allegorical treatments of treachery and cruelty, with the intention of making them less emotionally vulnerable to these phenomena when they encounter them later in life.

AI advice: A

Your choice: C

The answer is D rather than C. In answer C the problem is solved by cutting down profit in the beginning but growing it later on. However, in this context, the problem is solved by giving exposure to a weakened version. In answer D the fairy tales also act as a weakened version of treachery and cruelty to which the children are exposed, thus making the principles align.

Confirm and continue

(b) Tutorial page with manual explanation.

Fig. 1. 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.

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. After that, we showed the same task with AI advice (and *explanations* depending on the experimental condition)

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

and provided an opportunity for the participants to alter their initial choice. An example of second stage is shown in Figure 1(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 [17], and in line with findings that people first make a decision on their own and only then decide whether to incorporate system advice [16]. It also fits with the research of Dietvorst *et al.* [5] on trust in two-stage decision making.

Logic Units-based Explanations. 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. [36], 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 top-5 logic units. One example of such explanation is shown in figure 1(a).

2.2 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 [21]. 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 [12, 28], 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.

At the same time, explanations have been shown to improve the human understanding and interpretation of AI advice [1, 26, 35], 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.

2.3 Experimental Conditions

In our study, all participants worked on logical reasoning tasks with two-stage decision making process (described in Sec. 2.1). 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 tasks selected are presented as normal tasks without any performance feedback or explanation for correct answers, to prevent 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.

2.4 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 [38, 40]. 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.* [32] presented four conditions of appropriate reliance patterns 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. To provide an overview of participants' appropriate reliance under initial disagreement, we considered **Accuracy-wid** (*i.e.,* accuracy with initial disagreement).

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 [25, 33, 34]: Trust in Automation (TiA) questionnaire [20] and Affinity for Technology Interaction Scale (ATI) [13].

2.5 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 [11]. 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[2], 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 [23] 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).

2.6 Procedure

The full procedure that participants followed in our study is illustrated in Figure 2. 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.

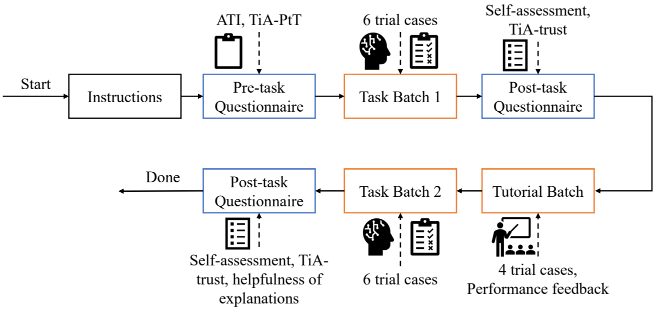


Fig. 2. Illustration of the procedure participants followed within our study. This flow chart describes the experimental condition ✓ **`Tutorial`**, ✓ **`XAI`**. Blue boxes represent the questionnaire phase, orange boxes represent the task phase.

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.

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

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). 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 [15].

3 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.[3] 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. No significant difference is found in the time spent across experimental conditions.

3.1 Hypothesis Tests

3.1.1 *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 1.

Table 1. 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* ± *SD*(Under)	*M* ± *SD*(Accurate)	*M* ± *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

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

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

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**, 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.

3.1.2 ***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**.

3.1.3 ***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 2. 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**.

Table 2. 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	↑

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.

3.1.4 *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, 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.

4 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 DKE and achieve appropriate reliance, more comprehensive, insightful, and personalized user tutorials are required.

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 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. This work is encore part of [19], specifically used for CHI TRAIT workshop submission. For more details about related work, experimental setup, results and analysis, please refer to the original paper [19].

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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).

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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)*.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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).*

[40] 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.