Computers in Human Behavior 175 (2026) 108779

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

# AI makes you smarter but none the wiser: The disconnect between performance and metacognition

Daniela Fernandes [a,*], Steeven Villa [b], Salla Nicholls [a], Otso Haavisto [a], Daniel Buschek [c], Albrecht Schmidt [b], Thomas Kosch [d], Chenxinran Shen [e], Robin Welsch [a]

[a] Aalto University, Espoo, 02150, Finland
[b] LMU Munich, Munich, 80539, Germany
[c] University of Bayreuth, Bayreuth, 95440, Germany
[d] HU Berlin, Berlin, 10099, Germany
[e] Independent Researcher, Vancouver, Canada

ARTICLE INFO

*Keywords:*
Human–AI interaction
Human-centered computing
Metacognition
Overconfidence
Generative AI

ABSTRACT

Optimizing human–AI interaction requires users to reflect on their performance critically, yet little is known about generative AI systems' effect on users' metacognitive judgments. In two large-scale studies, we investigate how AI usage is associated with users' metacognitive monitoring and performance in logical reasoning tasks. Specifically, our paper examines whether people using AI to complete tasks can accurately monitor how well they perform. In Study 1, participants (N = 246) used AI to solve 20 logical reasoning problems from the Law School Admission Test. While their task performance improved by three points compared to a norm population, participants overestimated their task performance by four points. Interestingly, higher AI literacy correlated with lower metacognitive accuracy, suggesting that those with more technical knowledge of AI were more confident but less precise in judging their own performance. Using a computational model, we explored individual differences in metacognitive accuracy and found that the Dunning–Kruger effect, usually observed in this task, ceased to exist with AI use. Study 2 (N = 452) replicates these findings. We discuss how AI levels cognitive and metacognitive performance in human–AI interaction and consider the consequences of performance overestimation for designing interactive AI systems that foster accurate self-monitoring, avoid overreliance, and enhance cognitive performance.

## 1. Introduction

Humans have always used technologies to augment their cognitive abilities (Alexandre e Castro, 2024; Clark, 2008; Khettab, 2019). Recent advances have aimed to improve human performance and productivity in a range of contexts (Hou et al., 2024; Perera, 2024; Wang et al., 2020; Zulfikar et al., 2024). While there is evidence for an improvement in human performance with AI (Bansal et al., 2021; Steyvers et al., 2022; Zulfikar et al., 2024), the integration of AI also brings challenges related to how users perceive, interact, and rely on these systems. Specifically, it is crucial to understand how AI influences individuals' ability to accurately assess their own competence and make informed decisions, particularly in situations where overconfidence or underestimation determines the success and efficacy of AI applications in real-world settings (Buçinca et al., 2021; Fleming, 2024). A core issue within this scope is the impact of AI on human metacognition — the ability to monitor and regulate one's cognitive processes. From a psychological perspective, people commonly rely on AI to boost their cognitive processes, raising fundamental questions about how people perceive their augmented performance when collaborating with AI, and whether they remain aware of potential errors (Buçinca et al., 2021; Fleming, 2024). Fundamental biases, such as overtrust and overreliance, impair performance (Inkpen et al., 2023) up to the point that the interaction decreases overall performance as compared to having no AI at all (Bastani et al., 2024; Vaccaro et al., 2024).

Psychological research on metacognition has shown that people typically estimate themselves to be better than average (Brown, 1986), also called the "better-than-average effect" (see Zell et al. (2020)). In the context of AI, people believe AI improves performance (Kloft et al., 2024), that AI predictions outperform professionals (Shekar et al., 2024) and hope AI will improve their lives (Cave & Dihal, 2019).

* Corresponding author.
*E-mail address:* daniela.dasilvafernandes@aalto.fi (D. Fernandes).
*URL:* https://aalto.fi (D. Fernandes).

https://doi.org/10.1016/j.chb.2025.108779
Received 13 March 2025; Received in revised form 5 August 2025; Accepted 20 August 2025
Available online 9 October 2025

Research offers some scattered evidence of deficiencies in metacognitive monitoring: users are largely unaware of their performance and their performance improvement with AI. Concretely, when using AI systems, users tend to overestimate their benefits, even when using a sham AI system (Kloft et al., 2024; Kosch et al., 2023; Villa et al., 2023). Moreover, when studying with AI support, people are unaware of their learning process, which leads to low exam scores. However, accurate metacognitive monitoring is crucial for optimizing human–AI interaction (HAI). Inaccurate evaluation of human–AI composite performance (Engelbart, 1962) can lead to an overreliance on the system, resulting in suboptimal outcomes.

From a rational decision-making perspective in HAI (Oulasvirta et al., 2022), optimal interaction with AI requires that users possess a clear understanding of their performance to adjust their behavior. Similarly, metacognitive judgments exhibit considerable individual variability (Ackerman & Thompson, 2017; Toplak et al., 2011), which can relate to cognitive performance (Toplak et al., 2011). The Dunning-Kruger Effect (DKE) describes a cognitive bias where individuals with lower ability overestimate their competence while those with higher ability underestimate it (Kruger & Dunning, 1999). For HAI, a DKE would suggest that low performers may not optimize their interaction with AI due to poorer metacognitive monitoring ("rational-hypothesis"). However, one could argue that if AI interaction improves overall cognitive performance by augmenting intellect (Engelbart, 1962), then metacognitive bias and its link to cognitive performance may disappear ("augmentation-hypothesis").

Despite expanding work on AI-assisted decision-making, few studies systematically examine how people calibrate self-assessments while reasoning with AI. To disambiguate these opposing hypotheses and address the remaining gap in the literature, we must empirically evaluate metacognitive monitoring, including metacognitive bias, individual metacognitive accuracy, metacognitive sensitivity, and its relation to performance (DKE) in HAI. Building on preestablished constructs of metacognition (Fleming, 2024), we focus on whether the DKE persists and whether users' AI literacy mitigates or exacerbates potential overconfidence. We thus conceptually replicate Jansen et al. (2021) in interaction with AI to explore whether AI impacts self-assessments of performance (**RQ1**: Is interaction with AI associated with reduced metacognitive accuracy?), if it reduces the ability to distinguish between correct and incorrect judgments (**RQ2**: Does interaction with AI as compared to no-AI increase or decrease metacognitive sensitivity?), and if it amplifies or reduces self-assessment bias between low- and high-performing individuals (**RQ3**: Does interaction with AI reduce or amplify the DKE pattern?).

We designed an experiment where participants used AI to complete logical reasoning tasks from the Law School Admission Test (LSAT). This setting is analogous to those utilized in prior research on metacognitive abilities and the DKE (Jansen et al., 2021). By analyzing participants' self-assessments after AI interaction, we can describe how AI use is associated with participants' metacognition and study its relation to task performance (DKE).

In Study 1 (N=246), we found that while AI use substantially improves task performance in the LSAT, it also coincides with a large overestimation of users' performance (low metacognitive accuracy). Yet, we show using a computational model that the DKE is not only smaller but disappears entirely in our sample while being present in a comparable large-scale sample without AI use (Jansen et al., 2021). Technological knowledge and critical appraisal of AI, as measured by the "Scale for the assessment of non-experts AI literacy" (SNAIL) (Laupichler et al., 2023), increased confidence but decreased the accuracy of self-assessment. In Study 2 (N = 452), where we incentivize metacognitive monitoring with monetary benefits and collect our own non-AI baseline group, we replicate the pattern of results of Study 1.

To summarize, although AI has the potential to improve performance in cognitive tests such as the LSAT and level individual biases in metacognition, it carries the risk of inflated self-assessments of performance. We discuss how to navigate this trade-off and how to improve metacognitive accuracy to empower users to make better decisions when using interactive AI. Our paper extends our understanding of metacognitive monitoring in HAI by investigating the interplay between metacognition, cognitive performance, and AI literacy. Our contributions and research results are:

1. Empirically examining associations between AI use and metacognitive monitoring.
2. Revealing that while AI can improve task performance, it leads to overestimation of performance.
3. Demonstrating that the DKE is reduced when participants use AI, suggesting that AI can level cognitive and metacognitive deficits.
4. Highlighting a paradox where higher AI literacy relates to less accurate self-assessment, with participants being more confident yet less precise in their performance evaluations.
5. Offering design recommendations for interactive AI systems to enhance metacognitive monitoring by empowering users to critically reflect on their performance.

## 2. Background

### 2.1. Human metacognition

Human metacognition research investigates the ability to monitor, evaluate, and regulate our own cognitive processes (Fleming, 2024) and, therefore, has been proposed to be essential in interactive generative AI systems (Tankelevitch et al., 2024).

A key aspect of metacognition is distinguishing between internal cues (i.e., self-generated reasoning) and external feedback (Koriat, 1997). When feedback is immediate or requires little effort, individuals may develop "illusions of knowledge", overestimating how much they truly know (Fiedler et al., 2019; Fisher & Oppenheimer, 2021). In the context of AI, these illusions may become even more pronounced, as high-quality assistance can overshadow users' metacognitive cues about their abilities.

Metacognitive judgments primarily involve accuracy and sensitivity. Metacognitive sensitivity reflects the ability to distinguish between correct and incorrect judgments, often measured by confidence ratings post-decision (Fleming, 2024). Perfect metacognitive sensitivity would entail that an individual's confidence ratings accurately reflect their performance, with high confidence corresponding to correct judgments and low confidence corresponding to incorrect judgments. Metacognitive accuracy is shaped by metacognitive bias (consistent over- or underestimation of one's cognitive abilities or performance) and noise (encompasses random, unintentional fluctuations in self-assessments) (Colombatto & Fleming, 2023; Fiedler et al., 2019; Fleming, 2024). Bias skews evaluations predictably, while noise introduces inconsistency, reducing sensitivity (Fleming, 2024). High metacognitive noise relates to low metacognitive sensitivity (Fleming, 2024).

The DKE appears in the connection between metacognitive accuracy and skill. It suggests that less-skilled individuals overestimate their performance, while highly competent individuals underestimate theirs (Kruger & Dunning, 1999). Despite the existence of debates regarding whether the DKE is a statistical artifact or if it accurately reflects true population trends (Gignac, 2024; Gignac & Zajenkowski, 2020; Jansen et al., 2021), studies like Jansen et al. (2021) and Ehrlinger et al. (2008) replicated these findings with large samples, confirming the DKE in verbal and logical reasoning tasks.

Our approach to studying the DKE shifts the focus to a task-specific context, even though metacognition and the DKE are often examined in educational settings (e.g., see Hansen et al. (2024) in math education and Mahdavi (2014) for an overview).

Following Dunning (2011), we replicate the method of Jansen et al. (2021) by concentrating on task performance and ratings of absolute performance estimates after task completion.

### 2.2. Metacognition in human-AI interaction

As AI technologies continue to become more integrated into daily life, transcending their original scope (Shneiderman, 2020), they offer unprecedented opportunities to augment human capabilities in a broad range of contexts – such as in medical treatment (Moor et al., 2023), drug discovery (Mak et al., 2023), and climate change (Kaack et al., 2022) – as well as in personal contexts (Draxler et al., 2023), enhancing productivity, improving decision-making and supporting learning (Draxler et al., 2024). However, such interactions' effectiveness heavily depends on how users perceive, trust, and engage with AI systems (Omrani et al., 2022).

A recent survey by Vaccaro et al. (2024) distinguishes between human–AI synergy – where combined performance surpasses either humans or AI alone – and human–AI augmentation, where humans aided by AI do better than unassisted humans. They found that when humans already outperform AI, adding AI improves the team's overall performance. However, as AI becomes more powerful, the average performance of these teams declines. Thus, a central challenge in human–AI interaction is achieving synergy when AI models surpass human capabilities.

These issues likely stem from suboptimal interfaces that fail to support metacognition (Kloft et al., 2024; Kosch et al., 2023; Tankelevitch et al., 2024). Research shows that users often overestimate their AI-assisted performance and struggle to monitor or plan interactions effectively (Bosch et al., 2024; Kloft et al., 2024; Kosch et al., 2023; Villa et al., 2023). For instance, Zamfirescu-Pereira et al. (2023) found that users have difficulty crafting effective prompts, while Dang et al. (2023) noted challenges in switching between tasks and writing prompts. Furthermore, explanations from AI systems are often uninformative, ignored, or lead to cognitive biases themselves (Bertrand et al., 2022; Eiband et al., 2019; Vasconcelos et al., 2023; Wang & Yin, 2021). AI literacy, which involves understanding AI concepts and evaluating outputs critically, is also essential for effective interaction (Laupichler et al., 2023). However, its influence on metacognitive judgments in AI-assisted decision-making and interaction optimization is unclear.

In sum, although previous work highlights that people often offload cognition to external supports, the specific interplay between self-assessment and AI assistance remains insufficiently explored (Tankelevitch et al., 2024). In particular, it is unclear how immediate AI help might distort self-assessments of competence or amplify biases such as the DKE. To address this, we examine how users estimate their own performance when interacting with AI, building on established metacognition frameworks (Fleming, 2024) and prior research on human–AI collaboration (Buçinca et al., 2021).

## 3. Research model and hypotheses

We have conducted two studies. Study 1 compares a group of participants using AI to Jansen et al. (2021) data on the LSAT (where the task was the same but had not involve AI). Study 2 replicates and extends Study 1.

Despite expanding work on AI-assisted decision-making, few studies systematically examine how people calibrate self-assessments while reasoning with AI. To disambiguate the opposing hypotheses (“rational-hypothesis” vs. “augmentation-hypothesis”) and address the remaining gap in the literature, we focus on whether the DKE persists and whether users' AI literacy mitigates or exacerbates potential overconfidence. We thus conceptually replicate Jansen et al. (2021) to explore whether interaction with AI is associated with metacognitive accuracy, metacognitive sensitivity, and the DKE pattern. Recent work on AI literacy distinguishes factors like Technical Understanding (TU), Critical Appraisal (CA), and Practical Application (PA) (Laupichler et al., 2023), each of which may differentially shape confidence and calibration. We further examine how AI literacy subscales (TU, CA, PA) relate to these outcomes (see Table 1).

Our research model aims to clarify how AI usage influences both objective and perceived performance in logical reasoning tasks, as well as how users' AI literacy might shape these effects. In this model, AI usage is hypothesized to improve objective performance (i.e., achieved number of correct answers), given that AI systems can offer high-quality outputs. However, we propose that such AI usage can alter key aspects of metacognitive monitoring, namely, metacognitive accuracy, metacognitive sensitivity, and the DKE, in ways that might undermine users' self-awareness.

Metacognitive accuracy refers to the difference between a user's perceived performance and their actual performance. Building on preestablished constructs of metacognition (Fleming, 2024), we explore whether AI usage affects metacognitive accuracy (more or less accurate self-assessments) (**RQ1**).

Metacognitive sensitivity refers to a user's ability to discriminate between correct and incorrect responses, often measured via confidence judgments. In our model, the immediacy of AI outputs could weaken individuals' sensitivity (Fleming, 2024), making them less able to identify potential errors. We therefore explore whether interaction with AI affects metacognitive sensitivity, and if it amplifies or reduces self-assessment bias between low- and high-performing individuals (**RQ2**).

Third, we explore whether the DKE manifests differently when individuals rely on AI. The Dunning–Kruger pattern typically shows that low-performing individuals overestimate their abilities and high-performing individuals underestimate them, reflecting systematic variations in metacognitive accuracy as a function of skill. In our research model, we explore whether AI usage might flatten or even erase such skill-based differences, creating a new distribution of over- and underestimation patterns (**RQ3**).

To motivate these, we consider three possible mechanisms.

First (**H1**), people often misattribute externally generated information to themselves (Johnson et al., 1993). AI may blur self vs. AI output distinction, causing source-monitoring errors (Johnson et al., 1993), and fostering an “illusion of knowledge”. Fisher and Oppenheimer (2021) show that reading fluent explanations inflates perceived understanding. Drawing on prior research (Fisher & Oppenheimer, 2021; Fleming, 2024; Tankelevitch et al., 2024), we therefore hypothesize that users may mistake the AI's capabilities for their own, thus relatively inflating their performance estimates, lowering metacognitive accuracy despite objective performance gains.

Second (**H2**), instant, highly-confident AI responses trigger the processing-fluency heuristic (Alter et al., 2007), impairing deliberate and effortful error checking (Diemand-Yauman et al., 2011). For this reason, and consistent with the model of Fleming (2024), we expect reduced correspondence between confidence and correctness, i.e., reduced metacognitive sensitivity. This effect is likely to arise if people perceive AI's suggestions as highly reliable, thereby reducing their motivation to examine responses closely.

Third (**H3**), the DKE appears in the connection between metacognitive accuracy and skill: low performers overestimate and high performers underestimate their abilities (Kruger & Dunning, 1999). Under the “rational-hypothesis” lens, this implies that low-skill users, poor at metacognitive monitoring, will fail to optimize their interaction with AI, perpetuating or even amplifying their calibration errors. Conversely, the “augmentation-hypothesis” argues that AI use (Engelbart, 1962; Risko & Gilbert, 2016) levels task accuracy, potentially dissolving the bias–skill link entirely. We therefore ask whether AI support will (a) preserve the classic DKE (“rational-monitoring”) or (b) attenuate it by compressing performance insight variance across users (“augmentation”). Moreover, users' AI literacy may moderate these outcomes. Although one might hypothesize that greater AI knowledge fosters more calibrated self-assessment, it is equally possible that higher literacy encourages false confidence and illusions of competence (Fisher & Oppenheimer, 2021). Consequently, our model incorporates AI literacy as a factor that could amplify, mitigate, or otherwise shape the impact of AI usage on both performance and metacognitive processes.

In summary, the proposed research model positions AI usage as a key driver of changes in objective and metacognitive performance

**Table 1**
Descriptive Statistics for all subjective variables for the full sample and split by performance quartile (Q).

| | Full sample | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|---|
| *n* | 246 | 110 | 59 | 55 | 22 |
| Performance | 12.98 (2.88) | 10.82 (3.07) | 14 (-) | 15 (-) | 16 (-) |
| Estimate | 16.50 (3.71) | 15.34 (4.31) | 17.39 (2.93) | 17.40 (3.07) | 17.68 (2.17) |
| Compared to other participants in this study, how would you rate your general logical reasoning ability when using the help of AI? (% rank) | 68.08 (19.3) | 66.02 (20.51) | 70.71 (18.35) | 68.22 (17.87) | 71.0 (19) |
| Using the AI, how many of the 20 logical reasoning problems do you think you will solve correctly? | 15.96 (3.63) | 15.49 (3.83) | 16.17 (3.6) | 16.33 (3.68) | 16.82 (2.13) |
| Without AI use, how many of the 20 logical reasoning problems do you think you would solve correctly? | 11.64(4.53) | 11.25(4.95) | 11.36(4.19) | 12.24(4.31) | 12.91(3.58) |
| Compared to other AI systems, how would you estimate the AI system's logical reasoning ability? (% rank) | 70.02 (17.91) | 69.04 (18.04) | 69.44 (18.01) | 70.0 (18.27) | 76.59 (15.79) |
| On its own, how many of the 20 logical reasoning problems do you think the AI would solve correctly? | 16.54 (7.75) | 15.75 (4.53) | 16.47 (3.74) | 18.31 (14.42) | 16.32 (3.26) |
| Compared to other participants in this study, how well do you think you will do? (% rank) | 66.95 (19.84) | 64.0 (21.93) | 69.31 (17.49) | 68.8 (17.54) | 70.77 (19.3) |
| How difficult is solving logical reasoning problems for you? | 5.38 (2.01) | 5.43 (2.05) | 5.88 (2.03) | 4.93 (1.86) | 4.91 (1.93) |
| How difficult is solving logical reasoning problems for the average participant? | 6.09 (1.56) | 6.14 (1.62) | 6.34 (1.59) | 5.93 (1.53) | 5.64 (1.14) |
| Compared to other participants in this study, how would you rate your general logical reasoning ability when using the help of AI? (% rank) | 69.83 (21.86) | 66.86 (24.44) | 72.9 (19) | 72.2 (20.18) | 70.45 (18.5) |
| Using the AI, how many of the 20 logical reasoning problems do you think you solved correctly? | 16.50 (3.71) | 15.35 (4.31) | 17.39 (2.93) | 17.40 (3.07) | 17.68 (2.17) |
| Without AI use, how many of the 20 logical reasoning problems do you think you would have solved correctly? | 11.61 (4.52) | 11.21 (4.6) | 11.78 (4.47) | 12.00 (4.94) | 12.18 (3.03) |
| Compared to other AI systems, how would you estimate the AI system's logical reasoning ability? (% rank) | 76.29 (18.42) | 74.04 (19.51) | 78.41 (17.03) | 77.2 (19.57) | 79.64 (11.92) |
| On its own, how many of the 20 logical reasoning problems do you think the AI would have solved correctly? | 17.74 (8.65) | 17.25 (10.40) | 17.68 (2.84) | 18.69 (10.45) | 18.0 (2.16) |
| Compared to other participants in this study, how well do you think you performed? (% rank) | 68.63 (21.39) | 65.11 (23.81) | 71.14 (19.1) | 72.13 (19.21) | 70.82 (17.9) |
| How difficult was solving these logical reasoning problems for you? | 5.67 (2.33) | 5.85 (2.35) | 5.64 (2.24) | 5.47 (2.39) | 5.32 (2.36) |
| How difficult was solving these logical reasoning problems for the average participant? | 6.11 (2.09) | 6.3 (2.14) | 6.07 (1.95) | 5.82 (2.15) | 6.0 (2.09) |
| SNAIL: Technical Understanding | 3.83 (1.60) | 3.75 (1.57) | 3.99 (1.62) | 3.77 (1.66) | 3.94 (1.55) |
| SNAIL: Critical Appraisal | 5.03 (1.28) | 5.02 (1.29) | 5.05 (1.25) | 5.01 (1.33) | 5.05 (1.29) |
| SNAIL: Practical Application | 5.02(1.31) | 5.0(1.39) | 4.99(1.27) | 5.02(1.32) | 5.17(1.07) |

Note: M (SD) and the sample size (n). Scale for the assessment of non-experts' AI literacy (SNAIL). According to task context, rank % instructions are scaled as follows: marking 90% means you perform better than 90% of participants, marking 10% means you perform better than only 10% of participants, and marking 50% means you will perform better than half of the participants.- indicates no variation, e.g., when all participants in a quantile had the same value.

while also recognizing that individual differences in AI literacy may interact with such effects. Over two empirical studies, we examine each of these relations, assessing whether AI indeed boosts task performance, whether it influences users' global and local metacognitive judgments, and whether it disrupts or reshapes the classic DKE.

## 4. Study 1: Metacognition in human-AI interaction

### 4.1. Method

In the following, we motivate and document our methodological choices when conducting Study 1. The research software can be found at the following repositories: https://github.com/aaltoengpsy/interface-frontend and https://github.com/aaltoengpsy/interface-backend. Note that both the data and the material of Jansen et al. (2021) are openly available under https://osf.io/er9ms/, which allowed us to closely follow their task environment and sample characteristics for the purpose of Study 1. All data collected for the purpose of our paper and analysis scripts can be found at https://osf.io/svax9/overview. As the two samples were gathered at different times and participants were not randomly assigned to "AI" versus "No-AI" conditions, any differences we observe are descriptive associations rather than causal effects. We therefore interpret Study 1 as providing suggestive, not causal, evidence.

#### 4.1.1. Participants

To explore individual differences in cognitive and metacognitive performance, we recruited a larger sample than typical DKE studies,

allowing us to detect differences in metacognitive accuracy across high and low performers (Dunning, 2011; Gignac & Szodorai, 2024). We powered for the smallest effect of interest, which is the DKE. For power analysis, we used bootstrapped samples of Jansen et al. (2021) with sample sizes ranging from 80 to 400 to assess the ability to detect the DKE through $t$-tests across quartiles. We computed the proportion of $p$-values < .05 to determine the optimal sample size for sufficient statistical power (80%). With this, we found that a sample size of 250–300 participants is optimal for reliably detecting differences between the upper and lower quartiles in metacognitive accuracy.

We recruited 274 English-fluent participants located in the USA through Prolific. We included an attention check, requiring participants to read a short description of the study and task. They then answered two multiple-choice questions, one about the topic (logical reasoning) and another regarding which option to choose when solving the problems (the best one). We excluded thirteen (13) participants due to failing the attention check, as well as two (2) due to erroneous responses (e.g., exceeding the number of possible correct answers in estimating performance) and thirteen (13) due to low completion times.

We further analyzed data from 246 participants (identified as female 114; identified as male 130, identified as non-binary 2; Age: $M = 39.85$, $SD = 14.53$). When asked to estimate their English fluency, 218 participants reported themselves as native English speakers, 25 as fully fluent, two as conversationally fluent and one as understanding basic English. No participants preferred not to disclose their language proficiency. 14 participants in our sample reported their highest educational degree to be a doctoral degree, 58 a higher tertiary education degree (Master's level), 95 a lower tertiary education degree (Bachelor's level), 52 an upper secondary school/high school and 27 a vocational college degree. 13 participants had taken the LSAT before; their performance was slightly lower, $M = 12.38$, $SD = 3.22$, compared to those who have not taken it, $M = 13.01$, $SD = 2.85$, thus they were not excluded from the sample. We collected informed consent from each participant before the study in accordance with the Declaration of Helsinki guidelines of 2013. Each participant was compensated 6.5 pounds per hour. In accordance with the TENK national guidelines (the Finnish National Board on Research Integrity), this study did not require ethics approval as it involved minimal risk to participants, with no intervention beyond standard practice and no collection of sensitive personal data.

In comparison to our sample, Jansen et al. (2021) drew a much larger benchmark cohort—3,543 U.S. adults recruited on MTurk—who completed the items entirely unaided. The study was purely observational, collected before the advent of ChatGPT, and offered a flat participation compensation of 3$.

#### 4.1.2. Quasi-experimental design

Participants' logical reasoning ability was measured with the 20 multiple-choice logical reasoning problems used by Jansen et al. (2021) to approximate the LSAT, a widely recognized, real-world assessment used in high-stakes decision-making, such as law school admissions (Shultz & Zedeck, 2011; Wainer, 1995). It also serves as a benchmark in machine learning research, making it ideal for comparing AI-assisted performance (Katz et al., 2024). An example LSAT question provided to participants was: "It has been proven that the lie detector can be fooled. If one is truly aware that one is lying, when in fact one is, then the lie detector is worthless. The author of this argument implies that: (1) The lie detector is sometimes worthless. (2) The lie detector is a useless device. (3) No one can fool the lie detector all of the time. (4) A good liar can fool the device. (5) A lie detector is often inaccurate.".

Using the same items as Jansen et al. (2021) enabled us to compare our results to a representative sample of participants who did not use AI in the task and replicate the results of the original study by Kruger and Dunning (1999).

In addition to participants' actual logical reasoning performance with AI use, we measured perceived performance with and without AI, and AI's system performance on its own using the items presented in Table 1. Lastly, participants' AI literacy was measured using the SNAIL (Laupichler et al., 2023) at the end of the study, allowing us to evaluate AI literacy comprehensively among non-experts. The scale features 31 items to assess participants' technical understanding, critical appraisal, and practical application of AI systems. The scores can be found at the end of Table 1.

#### 4.1.3. Task

Participants completed 20 LSAT logical reasoning items in a randomized order. Each problem was displayed on the left-hand side of the screen, while a ChatGPT interface was presented on the right (see Fig. 1). Participants were required to interact with ChatGPT for assistance, ensuring at least one prompt per problem, before submitting their final answers and rating their confidence in their response ("How confident are you that your response is correct?"; from "unsure" to "certain" on a 100-step slider), see Table 1. Unlimited text chat interaction was enabled during the task, allowing participants to engage with ChatGPT as much as they deemed necessary. The LSAT problems were intended to assess logical reasoning abilities and did not require any prior knowledge of law to solve.

#### 4.1.4. Procedure

Upon entering the study on Prolific, participants were redirected to our application. After consenting to participate, we quantified user expectations before interaction with the systems. A series of studies (Bosch et al., 2024; Kloft et al., 2024; Kosch et al., 2023; Villa et al., 2023) found that users hold high expectations regarding their performance with AI systems, yet are largely unaware of their actual performance when completing tasks with AI-assistance – in other words, they fail to monitor their performance. Participants estimated how many of the 20 items they expected to answer correctly on a 0-100 numeric scale (pre-task expectation). After completing the task, they provided the same estimate again (post-task expectation). Collecting users' expectations both before and after interaction serves two purposes. First, pre-task expectations capture anticipatory beliefs that may influence subsequent behavior (a placebo-like mechanism that has been documented in HCI research on AI systems (Bosch et al., 2024; Kosch et al., 2023). Second, the difference between post- and pre-task scores provides an individualized index of expectation, which we use as a predictor of performance-monitoring accuracy models (see Analysis section). This also aligns with the original study of Jansen et al. (2021).

After measuring expected performance, they were briefly introduced to the task and allowed to test the chat interaction. Afterwards, participants engaged in a task to assess their logical reasoning skills by solving a series of LSAT problems. Before submitting their final answers, they were asked to interact with ChatGPT, ensuring they provided at least one prompt per problem. After completing each question, participants rated their confidence in their response using a 100-step slider ranging from "unsure" to "certain". Participants were permitted unlimited text-based interaction with ChatGPT, allowing them to seek as much assistance as they felt necessary during the problem-solving process. After solving the problems in a randomized order (see Section 4.1.3), participants were again asked to complete the expectations questionnaire in the past tense. They also responded to the SNAIL questionnaire (Laupichler et al., 2023) and filled in their demographic information, including age, gender, occupation, education level, English proficiency, and whether they had taken the LSAT before. Study 1 took, on average, 42 min to complete.

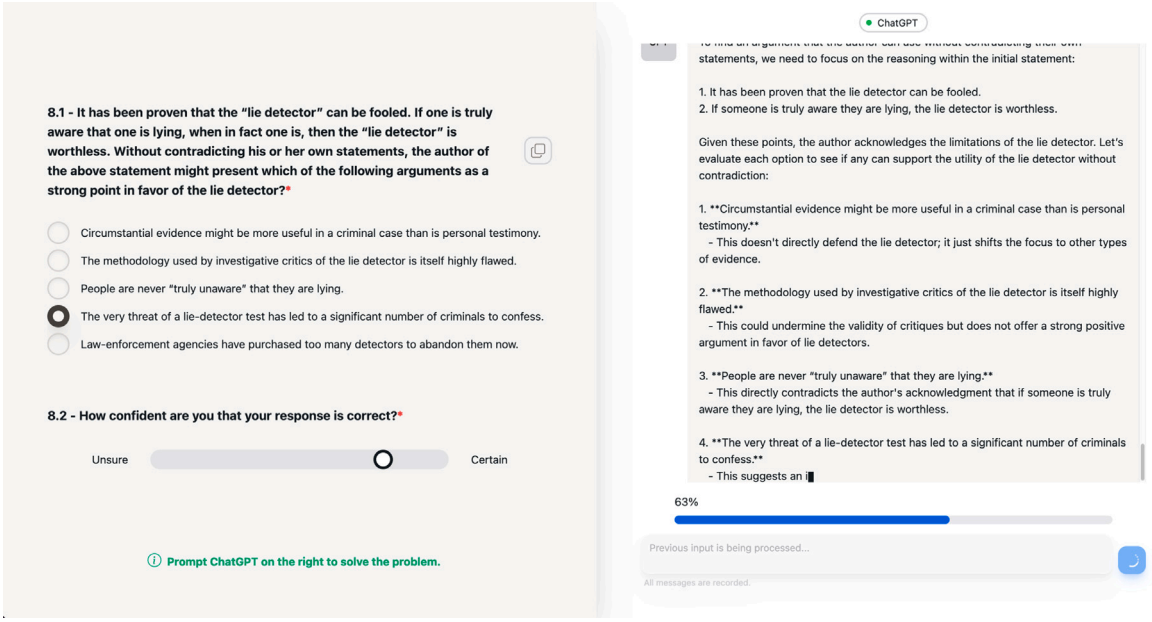



**Fig. 1.** Our online study application featured a horizontally split interface, with survey items and logical reasoning problems presented on the left and a ChatGPT interface on the right.

#### 4.1.5. Apparatus

We inspected the software and flow of Jansen et al. (2021) and carefully replicated it, integrating a side-by-side view of ChatGPT and the survey interface (Fig. 1). We used ChatGPT-4o (`gpt-4o-2024-05-13`) due to its widespread use in enhancing cognitive task performance (Bastani et al., 2024; Draxler et al., 2023, 2024). A custom interface (Fig. 1) was built to log user interactions, enabling qualitative chat analysis. The application automatically collected survey responses and chat logs and recorded them at the end of the study. We included a button to copy each problem and its answer options to the clipboard so they could be easily pasted into the chat.

#### 4.1.6. Analysis

We analyzed the data in five steps. First, we compared our sample performance to that of Jansen et al. (2021) and to the performance of AI alone. This allowed us to analyze where AI augments human performance (i.e., human–AI interaction outperforms a no-AI group) and human–AI synergy (i.e., human–AI interaction outperforms AI). We analyze both average performance and its distribution. Second, we analyzed metacognitive (**RQ1**) accuracy on a task level – the difference between objective and estimated performance[1] comparing human–AI interaction to a no-AI group – and trial-level confidence ratings to assess metacognitive sensitivity (**RQ2**). With this, we can analyze the metacognitive performance of users when interacting with AI globally (accuracy) after the tasks and locally (sensitivity) after each decision. Note that for Jansen et al. (2021), the question was ("How many of the 20 logical reasoning problems do you think you solved correctly?") while in our sample using AI, we asked: "Using the AI, how many of the 20 logical reasoning problems do you think you solved correctly?". Next, we correlated metacognitive performance metrics with performance and AI literacy measures to explore what predicts low metacognitive performance in human–AI interaction, for a similar analysis approach, see McIntosh et al. (2019). Fourth, we used a computational model of performance and performance assessments to compare our sample and Jansen et al. (2021) to estimate how AI affects the DKE (**RQ3**). Lastly, we qualitatively analyzed participant strategies and how they conceptualize the human–AI relation.[2] We use frequentist statistics ($\alpha$= 5%) for simple statistical tests (e.g., paired and unpaired $t$-test and Pearson correlation) in the first four analyses and Bayesian statistical modeling for the computational model. Note that we focus only on these analyses in the exploration of our data for the current paper; however, we encourage the re-analysis and further exploration in the openly available dataset: https://osf.io/svax9/overview.

### 4.2. Findings

#### 4.2.1. Human-AI composite performance

To see whether there is a synergy effect of using ChatGPT in the LSAT (i.e., Human–AI performance > AI performance), we compared the average ChatGPT performance (100 runs at $M = 13.65$) to our users' average performance ($M = 12.98$, $SD = 2.88$). We find a significant difference, with participants performing slightly worse than ChatGPT alone $t(245) = -3.66, p < .001, d = -0.23$ in the task. Next, we compared our sample's performance to Jansen et al. (2021) representative sample of 3543 participants, who completed the same task without any assistance ($M = 9.45$, $SD = 3.59$). We find that in our sample, participants performed significantly better with ChatGPT assistance $t(245) = 19.23, p < .001, d = 1.23$ as compared to the Jansen et al. (2021) sample. Therefore, while, on average, there is no human–AI synergy, we do find that ChatGPT use can augment human performance for solving the LSAT logical problems (i.e. human–AI performance > human performance).

Looking at individual performance, we can find indications of human–AI synergy. The difference between our sample's and ChatGPT's performance is rather small, at less than one point. 55.28% (136 of 246) of our participants performed better than ChatGPT. However, 89.43%

[1] In line with Ehrlinger et al. (2008), we focus on numeric estimates of performance after the task, not relative performance in comparison to others.

[2] Given the large samples in our study and the ceiling effects encountered Fig. 4(a), we do not test for the normality of residuals of our variables. Instead, we model the data using a final Bayesian computational framework, which allows for more flexible assumptions and can account for ceiling effects in our main analysis. This approach provides more robust estimates by incorporating uncertainty in a probabilistic manner rather than relying on strict parametric assumptions.

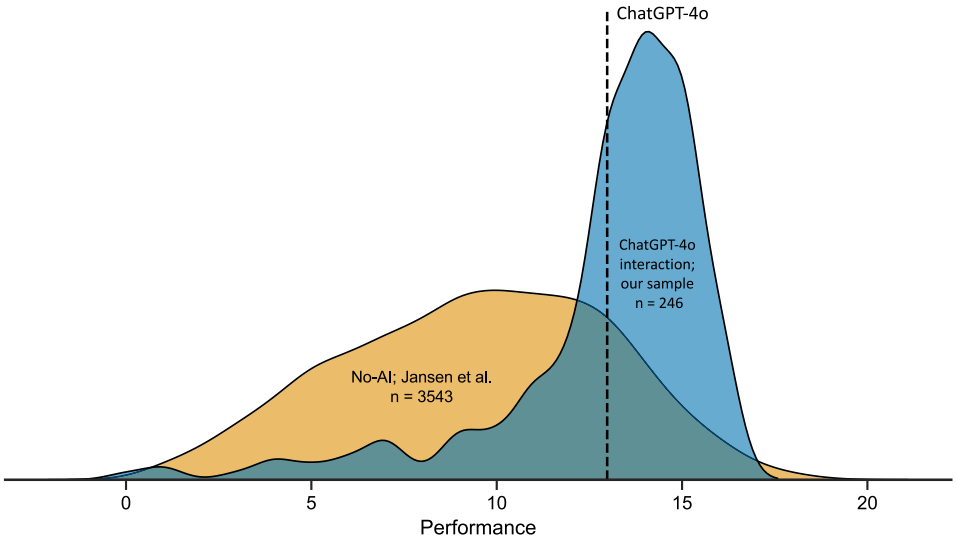



**Fig. 2.** Comparison of performance scores between participants interacting with ChatGPT and a dataset without AI (Jansen et al., 2021). The blue density curve represents the distribution of performance scores in a sample of 246 participants using ChatGPT, showing a peak in performance at around 14 points. The yellow density curve corresponds to a larger sample (n = 3543) from the Jansen et al. (2021) dataset, with lower overall performance scores. The vertical dashed line indicates the mean performance score in the ChatGPT simulation.

(220 of 246) in our sample performed better than the average score of the Jansen et al. (2021) sample (see also Fig. 2).

Therefore, while overall performance increased with the use of ChatGPT augmenting the human ability to solve LSAT problems, on average, we do not find a human–AI synergy. The composite performance of ChatGPT and the participant overtook the performance of ChatGPT alone for only slightly more than half of the participants in our sample. With our chat-based AI-assistance's ability to enhance performance established in human augmentation but not synergy, we can now focus on investigating metacognitive abilities.

#### *4.2.2. Metacognitive accuracy and sensitivity*

RQ1 investigated whether interaction with AI affects metacognitive accuracy, that is, how closely participants' estimated performance aligns with their actual performance. Our data shows that participants were inaccurate in assessing their performance after task completion, as indicated in the item "Using the AI, how many of the 20 logical reasoning problems do you think you solved correctly?", see also Table 1. On average, they estimated solving about 17 out of 20 items ($M$ = 16.50, $SD$ = 3.72). This overestimation of about 4 points could be distinguished from 0, $t(245) = 14.14, p < .001, d = 0.9$.

To test whether AI use amplifies overestimation, we applied the same metric to the (Jansen et al., 2021) dataset (no-AI condition; $N$ = 3543 *vs.* $N$ = 246 in our AI group). Comparing estimates of performance and actual performance, we find that the no-AI participants also significantly overestimated their performance, $t(3542) = 17.35, p < .001, d = 0.29$, but the effect size was substantially smaller than in the AI group ($d$ = 0.93). Note that for a more thorough investation of RQ1 a true experiment is needed comparing the biased estimation more closely.

To see whether participants track information in each trial and answer RQ2, we turn to metacognitive sensitivity that we estimate from confidence ratings when making the decision. After removing participants with only one level of correctness, i.e., all correct or all incorrect, we further analyzed data from 245 participants. The mean confidence (rated on a scale from 0–100) for correct answers was 82.49 (14.24) and for incorrect answers 77.00 (16.52), $t(244) = 8.21, p < .001$, $d = 0.52$. To evaluate sensitivity more granularly, we conducted a Receiver Operating Characteristic (ROC) analysis. ROC analysis is a technique used to assess the performance of a judgment by plotting the true positive rate against the false positive rate across different thresholds. Here, we applied ROC analysis to understand how well participants' confidence scores predicted whether their responses were correct (see Fig. 3A).

By using ROC analysis, we obtain a metric, the area under the curve (AUC), that quantifies how effectively a participant's confidence ratings differentiate between correct and incorrect responses. An AUC value of 0.5 indicates no better-than-chance discrimination, while higher values reflect greater sensitivity, meaning the participant's confidence reliably tracks correctness. Thus, the ROC analysis provides a nuanced individual, trial-level measure of metacognitive accuracy beyond simple average confidence or aggregate performance estimates.

The mean AUC was .62 ($SD$ = 11.2) which could be distinguished from 0.5 ($t(244) = 16.02, p < .001, d = 1.02$). Most participants' (210 out of 246; 85.37%) metacognitive AUC values are above .50 (random guessing). This means that confidence scores indicate participants' metacognitive sensitivity on a trial level in human–AI interaction.

Prior metacognition work treats AUC values above .7 as "moderate sensitivity", where participants are able to separate correct from incorrect answers (Ais et al., 2016; Clayton et al., 2023; Fleming & Lau, 2014). Our AI-assisted group achieved a mean AUC of .62, significantly above chance, yet noticeably attenuated compared to the benchmark. This value is sufficiently high to indicate that participants engaged in metacognitive processing.

Although participants' mean AUC (.62) exceeded chance, it fell significantly short of the commonly used "benchmark" of .7, $t(244) = -11.73, p < .001, d = -0.75$, indicating a deficit in trial-level sensitivity relative to that standard. Additionally, a mean AUC of .62 is small enough to confirm our prediction that AI support would temper participants' ability to distinguish right from wrong answers. This measured attenuation aligns with our hypothesis that interacting with the AI improves accuracy but impairs the ability that underlies effective self-monitoring. Against those reference points, our participants' mean AUC = .62 indicates they are relatively worse at monitoring their accuracy (above chance), yet below the level typically considered acceptable. Note that for the remaining 36 participants, confidence ratings could not distinguish between correct and incorrect trials (for the distribution of AUC values, refer to Fig. 3B). For these participants, confidence judgments were effectively random or worse than random chance, indicating that they tended to be as confident or even more confident about incorrect responses than correct ones. This pattern suggests a miscalibration in their metacognitive judgments, where confidence fails

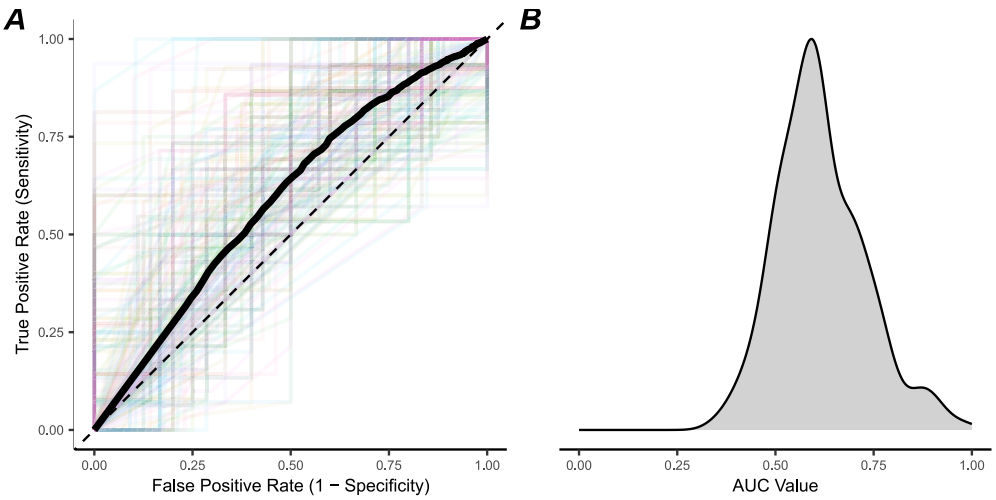



**Fig. 3.** (A) Receiver Operating Characteristic (ROC) curves showing the relationship between True Positive Rate (Sensitivity) and False Positive Rate (1 – Specificity) for participants (color) and pooled across participants (black). The dashed diagonal line indicates the line of no discrimination (random guessing). For each participant, we generated a ROC curve, colored lines, as well as one from pooled responses (black line), which illustrates the trade-off between the true positive rate (i.e., the proportion of correct judgments identified as correct) and the false positive rate (i.e., the proportion of incorrect judgments identified as correct) across various confidence thresholds. High metacognitive sensitivity would position the curves close to the $y$-axis and the top of the graph, and low metacognitive sensitivity to the dashed line. We can compute the area under the curve (AUC) as an estimate of metacognitive sensitivity. (B) Distribution of AUC values across all participants, with a peak around 0.6, suggesting variability in metacognitive sensitivity, with most participants performing above chance level (AUC = .5).

to serve as a reliable indicator of actual performance. Thus, our sample exhibits very low metacognitive sensitivity and, in consequence, low metacognitive monitoring. Note, however, that for a robust thorough investigation of RQ2 a true experiment is again needed comparing sensitivity across groups.

#### *4.2.3. Correlation of metacognitive ability, performance, and AI literacy (SNAIL)*

A number of significant relationships were found when correlating several metacognitive indices with LSAT performance and AI literacy, see Table 2. We observed a positive relation between performance and participants' average confidence estimates. Participants who performed well were also more confident on average. However, those who were, on average, more confident also overestimated their performance due to increased metacognitive bias. This is probably due to the relationship between SNAIL factors and performance estimates, where participants who expressed more technical knowledge and more critical appraisal also estimated their performance to be relatively higher. However, those with high technical understanding were also less accurate in their metacognitive judgments. All SNAIL factors correlated positively to average confidence. Note that these correlations are rather small and should thus be interpreted with caution. Metacognitive sensitivity (AUC and $\Delta conf$) was not related to AI literacy, performance, or metacognitive accuracy.

#### *4.2.4. AI use cancels the Dunning-Kruger effect*

RQ3 aimed to determine whether AI interaction would affect the classic DKE pattern in which lower performers overestimate their abilities while higher performers underestimate them. The correlation between estimated performance and actual performance is small to medium-sized (see Table 2, and for visual representation Fig. 4(a)). While some participants were very accurate in estimating their performance, some participants were considerably off in their estimates (Fig. 4(a)). This suggests the possibility of a DKE-like pattern, where ability in a task is related to the metacognitive ability to judge one's task performance. For the classical quantile plot, refer to Fig. 4(b).[3]

We calculated the difference between quantiles to test whether metacognitive accuracy was worse in the low-scoring quantile than in the high-scoring quantile. Both quantile's metacognitive accuracy differed from 0, (Q1: $t(109) = -10.15, p < .001, d = -0.97$, Q4: $t(21) = -3.64, p = .002, d = -0.78$), probably due to the overall bias. However, we found that the difference for Q1 is larger than for Q4, $t(130) = 2.79, p = .006, d = 0.49$ (see also Fig. 4(b) and Table 1). Note that this pattern of effect could be driven by metacognitive bias alone. To establish a DKE, we must first quantify the metacognitive noise in our sample. To do so, we employ a Bayesian computational model,[4] a hierarchical Bayesian model to jointly estimate participants' objective and perceived performance while accounting for latent skill, metacognitive bias, and metacognitive noise.[5] To allow for a baseline comparison of the DKE, we modeled the data jointly with that of Jansen et al. (2021), whose study did not involve AI. This approach enables us to compare the Dunning-Kruger effect in our sample, where participants used AI, with the non-AI sample of Jansen et al. (2021). The model accounts for ceiling effects in performance estimates, treating scores of 20 as censored observations.

Specifically, our hierarchical Bayesian model accounts for the presence of AI ($k$ = "AI" or "no AI") in estimating both the participants' achieved performance and their estimated performance. It also integrates latent skill, metacognitive bias, and group-level metacognitive noise, which scales the bias and latent skill. Metacognitive bias, in the model, reflects each person's over- or under- estimation of their skill, while metacognitive noise reflects the lack of information regarding their own skill.

Let $\theta_i$ represent the relative latent skill for participant $i$ with the prior $\theta_i \sim \mathcal{N}(0, 2)$.

The objective performance $y_{\text{obj},i}$ is modeled as:

$$y_{\text{obj},i} \sim \text{Binomial}\left(n_{\text{obj}}, \Phi_{\text{approx}}(\theta_i)\right),$$

[3] Note, however, that this plot can be misleading (Gignac & Zajenkowski, 2020).

[4] For a guide on Bayesian techniques, see Bürkner (2017), Dix (2022), Kay et al. (2016), Schad et al. (2021), van de Schoot et al. (2021), we used the tutorial of Nathaniel Haine's as a starting point for our modeling efforts: http://haines-lab.com/post/2021-01-10-modeling-classic-effects-dunning-kruger/.

[5] For a theoretical model, see Burson et al. (2006).

**Table 2**
Correlation table of metacognitive measures and AI literacy as measured by the SNAIL.

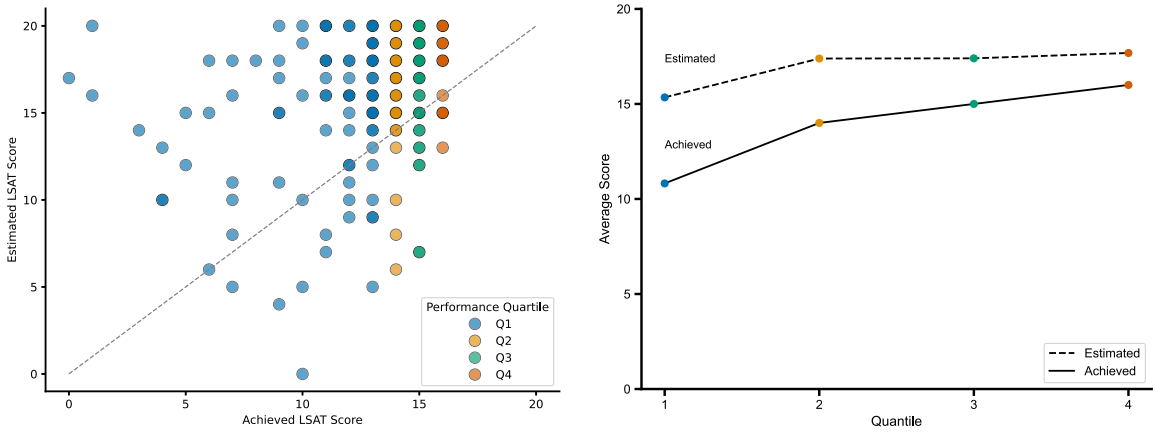
| | $\Delta EP$ | Estimate | Performance | $\Delta conf$ | $\mu conf$ | AUC | SNAIL TU | SNAIL CA | SNAIL PA |
|---|---|---|---|---|---|---|---|---|---|
| $\Delta EP$ | | | | | | | | | |
| Estimate | 0.72*** | | | | | | | | |
| Performance | −0.43*** | 0.32*** | | | | | | | |
| $\Delta conf$ | −0.04 | −0.03 | 0.01 | | | | | | |
| $\mu conf$ | 0.24*** | 0.46*** | 0.27*** | −0.10 | | | | | |
| AUC | 0.03 | 0.05 | 0.03 | 0.59*** | −0.08 | | | | |
| SNAIL TU | 0.21** | 0.17** | −0.06 | −0.12 | 0.13* | −0.10 | | | |
| SNAIL CA | 0.10 | 0.14* | 0.06 | 0.04 | 0.25*** | 0.03 | 0.49*** | | |
| SNAIL PA | 0.05 | 0.10 | 0.06 | 0.01 | 0.24*** | 0.05 | 0.57*** | 0.81*** | |

**Note**. $df = 243$, $\Delta EP$ represents the difference between performance and estimated performance (metacognitive accuracy). Performance refers to the achieved task performance. $\Delta conf$ is the difference between predicted and actual confidence, while $\mu conf$ is the mean confidence (average confidence ratings). AUC refers to Area Under the Curve, with a higher AUC value indicating a more reliable confidence score in reflecting participants' correctness; SNAIL TU stands for the Technical Understanding score, SNAIL CA represents the Critical Appraisal score, and SNAIL PA is the Practical Application score.

* Indicates $p < .05$.

** Indicates $p < .01$.

*** Indicates $p < .001$.




(a) *Scatter-plot of Estimated LSAT Scores as a function of Achieved LSAT Scores for the AI group in Study 1.*

(b) *Classical Dunning-Kruger Quartile-plot. Comparison of Mean Estimated and Mean Achieved Scores Across Quartiles for the AI group in Study 1.*

**Fig. 4.** Correlation of estimated and achieved LSAT score from different perspectives, individual Fig. 4(a) vs. quartile-level Fig. 4(b) for the AI group in Study 1.

where $n_{\text{obj}}$ is the total number of items and $\Phi_{\text{approx}}(\cdot)$ is the approximation for the cumulative standard normal distribution.

Perceived performance $y_{\text{per},i}$ is influenced by group level bias $b_k$, latent skill $\theta_i$ and noise $\sigma_k$, which scales the difference of bias $b_k$ and latent skill $\theta_i$:

Here, $n_{\text{per}}$ is the total number of perceived items. The priors for bias $b_k$ and noise $\sigma_k$ are the following:

$$b_k \sim \mathcal{N}(0, 2), \quad \sigma_k \sim \text{LogNormal}(0, 2).$$

Our model mitigates the issue around regression to the mean by explicitly modeling latent skill $\theta_i$ as a continuous variable with a flexible distribution. By incorporating noise $\sigma_k$ that scales skill level $\theta_i$ and bias $b_k$, the model allows for greater variability in judgment among low-skilled participants. This scaling effect, coupled with a bias term $b_k$ and hierarchical priors, reduces the tendency for all participants to regress toward a single mean. For a DKE to exist, bias that is $b_k > 0$ and noise that is $\sigma_k > 1$ has to be satisfied. If only metacognitive bias is driving a DKE pattern, then $\sigma_k$ will be centered at 1 (values of $\sigma_k$ under 1 and close to zero would mean that high-performers would be less accurate in estimating their performance[6]). The bias parameter for our AI-interaction sample, $b_{AI}$, showed a median of 0.45 (95% HDI [0.32, 0.60]). The consistently positive bias indicates that individuals, when using AI, tend to overestimate their abilities. We also find a bias $b_{noAI}$ for the non-AI group of Jansen et al. (2021) (Median = 0.23 (95% HDI [0.21, 0.25], $p_b = 0.0\%$) although when comparing posterior samples (see Fig. 5), 99% of posterior samples were larger in the AI group as compared to the non-augmented sample.

To understand how metacognitive noise affected self-assessment, we can turn to $\sigma_k$. For the non-AI group, we find a $\sigma_{noAI}$ above 1, indicating noise affecting self-assessment. This group had a median of 1.78 (95% HDI [1.69, 1.88], $p_b = 0.0\%$, see Fig. 5B), indicating noise in judgment for the sample of Jansen et al. (2021). Combined with bias, this contributes to the DKE (see also Fig. 5C for posterior predictions from the model). In comparison, our sample, which used AI to complete

[6] To fit our data into the model, we used the STAN-sampler (Carpenter et al., 2017). Four Hamilton-Monte-Carlo chains were computed, each with 15,000 iterations and a 30% warm-up. Trace plots of the Markov-chain Monte-Carlo permutations were inspected for divergent transitions and autocorrelation, and we checked for local convergence. All Rubin-Gelman statistics (Gelman & Rubin, 1992) were well below 1.1, and the Effective Sampling Size was over 1000.

We then analyzed the posterior of the model. To investigate a parameter's distinguishability from zero, we utilized $p_b$, which resembles the classical $p$-value but quantifies the effect's likelihood of being zero (for $b$) and one (for $\sigma$) or opposite (Hoijtink & van de Schoot, 2018; Shi & Yin, 2020). Effects with $p_b \leq 2.5\%$ were deemed distinguishable. We also calculated the 95% High-Density Interval (HDI) for each model parameter; for visualization of prior and posterior, see Fig. 5.

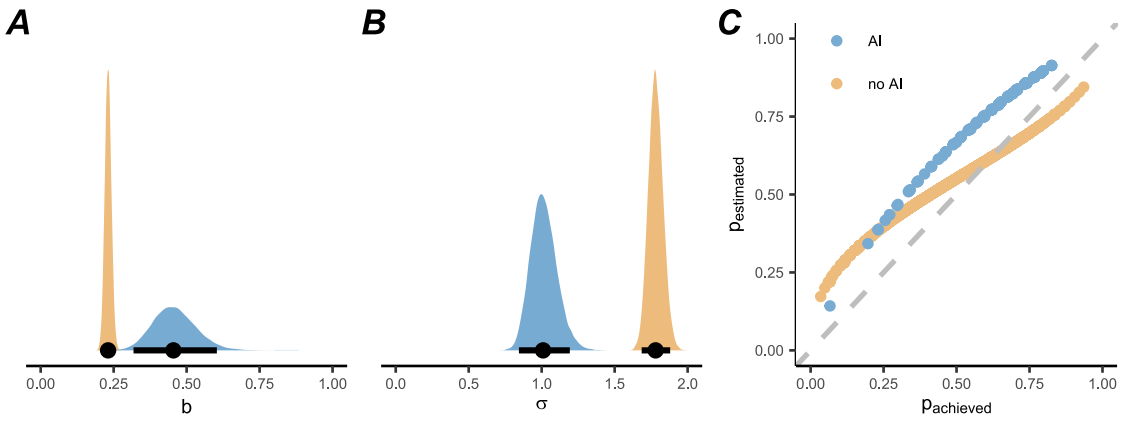



**Fig. 5.** Comparison of posterior distributions with median and 95% HDI for the model parameters $b$ for bias in each group (Plot A) and $\sigma$ (Plot B). The posterior distributions of the AI group (in blue) and no-AI group (in yellow). In Plot A, kernel-density curves show the full posterior for each group (blue = AI, yellow = no-AI); vertical ticks mark the posterior median and the shaded band the 95 HDI. All mass lies to the right of zero, but the AI density is clearly shifted further right, indicating stronger overestimation bias. In plot B, densities are plotted on the same scale as in (A). The no-AI posterior peaks well above the neutral point of $\sigma$, signifying noise that scales bias by skill and thus sustains a Dunning–Kruger gradient. In contrast, the AI posterior is centered almost exactly on $\sigma = 1$, implying that bias no longer increases as skill decreases. Plot C shows the average posterior predicted values for percent correct achieved (x-axis) and percent correct expected (y-axis) for each group. The s-shape around ideal metacognitive accuracy (gray line) indicates a DKE with low-performers overestimating their performance more than high-performers (yellow; no AI group).

the task, was not affected by the DKE (Median = 1.01, 95% HDI [0.84, 1.19], $p_b$ = 45.66%). Given the non-overlapping distributions (0% overlap) and the small HDI's, we can assert that with $\sigma_{AI}$ being around 1, scaling of the equation of bias and skill is not present in our sample. This finding aligns with our "augmentation-hypothesis": as AI's outputs levels individual skill differences, even lower performers achieve a higher performance level, resulting in uniform overestimation rather than the classic Dunning–Kruger pattern. Hence, when augmented with AI, we observe no DKE (see again Fig. 5C).

#### 4.2.5. Qualitative data

In addition to quantitative measurements, we analyzed the qualitative data collected during Study 1 using an inductive thematic approach (Clarke & Braun, 2017). This included both the prompts participants entered into the AI chatbot interface and their responses to an open-ended question at the end of the questionnaire. The analysis of the prompts provided insight into how participants interacted with the AI chatbot and their perceptions of it. The prompts were filtered to exclude those that were a direct copy from the task, ensuring that only meaningful interactions were kept for this analysis. The remaining prompts were then inductively analyzed to identify recurring themes (see Section 4.2.6). Responses to the open-ended question were analyzed to explore differences in AI perception during the interactions, where recurrent themes were found (see Section 4.2.6).

#### 4.2.6. Analysis of prompts

In our study, we collected 6629 prompts from 246 participants, each answering 20 logical reasoning questions using the AI chatbot. Although participants could use as many prompts as they wished (with a minimum of 1 per question), in practice, they seldom did. Across the 246 participants, the mean number of prompts per question was $M$ = 1.15 ($SD$ = 0.34). Table 3 shows the maximum number of prompts per participant across all items: 46% of participants prompted the system only once per question, and only 8% exceeded three prompts.

*Analysis of open-ended questions.* A qualitative analysis of the open-ended question "Please describe how you used the AI Chatbot" revealed diverse types of perceptions and interactions with AI among participants, highlighting varying degrees of user reliance, collaboration, and trust.

The majority of participants demonstrated a high level of trust in AI, often accepting its suggestions without further inquiry. This behavior raises concerns about overreliance on AI, as noted by Lu and Yin (2021). 12.60% of participants perceived AI as a collaborative partner, using inclusive language and viewing it as part of a joint effort rather than just a tool. Another 21.54% of participants viewed AI strictly as a complementary tool, using it cautiously for verification while maintaining control of the problem-solving process. Finally, 6.5% either provided inconclusive responses regarding the strategy used or did not find the AI tool useful. The data reveals diverse ways participants perceived AI, providing insights into HAI dynamics and individual variability.

**Table 3**
Maximum number of prompts per participant across all pages. Participants tended to prompt ChatGPT only one time more frequently. Only 8% of the time participants exceeded 3 prompts.

| Maximum number of prompts | Count |
|---|---|
| 1 | 113 (46%) |
| 2 | 90 (37%) |
| 3 | 23 (9%) |
| 4 | 10 (4%) |
| 5 | 4 (2%) |
| >5 | 6 (2%) |

### 4.3. Interim discussion

We found that using ChatGPT augmented our sample beyond a no-AI benchmark (i.e. AI augmentation) but that only a little more than half of our sample could surpass the AI alone (i.e. human–AI synergy). We found that most people overestimated their performance with AI and that there was no indication of a DKE when using the AI system. This may be due to participants' tendency not to reflect on their performance and low metacognitive sensitivity (on an absolute level, participants do not perform well), which is corroborated by our qualitative reports of people copying and pasting questions to the chat interfaces and then taking the AI's answer without reflection.

## 5. Study 2: Incentivizing metacognitive thinking

Our descriptive data revealed that most users rarely prompted ChatGPT more than once per question. This shallow level of engagement

**Table 4**
Participant approaches to AI use in study 1.

| Category | Description | Actual ($M \pm SD$) | Perceived ($M \pm SD$) |
|---|---|---|---|
| High level of trust | Participants relied heavily on AI ("blindly trusted"), copying and pasting questions without critically assessing AI's outputs or further inquiry. 58.94% (145 out of 246) | 13.014 ± 3.062 | 16.844 ± 3.679 |
| Collaborative Partner | Participants perceived AI as a collaborative partner rather than a mere tool, engaging in joint problem-solving and using inclusive language ("we did this") when describing their interactions. 12.60% (31 out of 246) | 13.065 ± 3.176 | 18.161 ± 1.881 |
| Complementary Tool | Participants used AI strictly as a complementary tool for verification of their independently formulated answers, maintaining control of the process. 21.54% (53 out of 246) | 10.5 ± 3.317 | 16.000 ± 6.733 |
| Inconclusive/Did Not Use | Participants either provided inconclusive responses regarding the strategy used during the experience or did not find the AI tool useful. 6.91% (17 out of 246) | 13.302 ± 2.729 | 16.302 ± 4.012 |

may have limited the cues needed to calibrate confidence and allow for accurate self-monitoring. It is therefore plausible that encouraging or experimentally requiring multiple prompts could provide better feedback loops, enhancing users' metacognitive sensitivity.

To address the potential confound of a lack of motivation in our sample to engage in metacognitive monitoring, we conducted a second study in which participants received a monetary incentive for accurate judgments across the task; for a DKE study employing incentives, see Ehrlinger et al. (2008) Study 3. If participants monitor their performance when incentivized, the DKE could resurface. Given that Jansen et al. (2021) did not incentivize participants for accurate metacognitive judgments, this also mandated the sampling of a no-AI group within our study setup. We thus sampled another 250 participants for each the AI and the no-AI group to see if an incentive can motivate metacognitive monitoring and analyze the quantitative data. All data and analysis scripts for Study 2 can be found at https://osf.io/svax9/overview.

### 5.1. Method

We recruited 500 English-speaking participants located in the USA through Prolific. The sample was split into two groups: 250 participants completed the task without AI assistance (no-AI group) and 250 participants completed the task with AI assistance (AI group). Participants solved the same 20 logical reasoning problems used in Study 1. We did not collect the SNAIL for the no-AI group. Study 2 took, on average, 25 min to complete for the no-AI group and around 52 min for the AI group. Each participant was compensated 7 pounds per hour.

To motivate accurate self-assessment, participants in both groups were informed they would receive monetary compensation based on the accuracy of their performance estimates (compensation would be given to participants whose presumed number of correct answers closely matched their actual score). This incentive aimed to motivate participants to engage critically with the task and closely monitor their performance. All participants received full benefit compensation of 0.50 pounds (around 8% increase) regardless of their achieved performance. Similarly to Study 1, we included an attention check where participants were required to read a brief study and task description. They then answered two multiple-choice questions, one about the topic (logical reasoning) and another regarding how they could receive additional compensation (good judgment).

From 500 participants, we analyzed 452 participants (age: $M = 37.24, SD = 13.36$): 245 in the AI group and 207 in the no-AI group. Across both groups, 48 participants were excluded (3 for missing data, 3 for invalid performance estimates (exceeding 20 correct answers), and 42 for too low completion times.

202 participants identified as female, 242 as male, 6 as non-binary, 1 as two-spirit, and 1 who preferred not to disclose. Their highest educational degrees included 21 doctoral, 132 Master's-level, 194 Bachelor's-level, 64 upper secondary school, and 41 vocational qualifications. Regarding English proficiency, 391 participants identified as native speakers, 56 as fully fluent, 4 as conversationally fluent, and 1 as having basic proficiency.

For the AI group, a subset of participants (26) reported prior experience taking the LSAT. Their performance in this study ($M = 13.50, SD = 1.98$) was comparable to those without prior LSAT experience (n = 219; $M = 13.25, SD = 2.55$), and they were not excluded from the analysis. For the no-AI group, 25 participants reported prior LSAT experience, performing similarly ($M = 9.50, SD = 4.06$) to those without LSAT experience (n = 182; $M = 9.52, SD = 3.60$).

### 5.2. Results and discussion

Participants in the AI group performed on average slightly worse as compared to AI alone ($M = 13.31$, $SD = 2.44$, $t(244) = -2.17, p = .031, d = -0.14$) but better than the no-AI group ($M = 9.71$, $SD = 3.59$, $t(450) = 12.60, p < .001, d = 1.18$). Therefore, we can assert that, on average, using AI has augmented performance but not that there is a synergy effect. In the AI group, 59.18% of participants scored higher than ChatGPT, with a total of 145 participants out of 245 surpassing its performance. In the no-AI group, 14.49% of participants scored higher than ChatGPT, corresponding to 30 participants out of 207, see also Fig. 6. Therefore, performance in Study 2 mirrors Study 1.

Investigating metacognitive accuracy for each sample, we find that in the AI group, participants overestimate their performance ($M = 17.13$, $SD = 3.16$), which differed significantly from zero when subtracting individual performance ($t(244) = 18.33, p < .001, d = 1.17$). The same was found for the no-AI group ($M = 13.62$, $SD = 4.14$, $t(206) = 11.81, p < .001, d = 0.82$). Both quantile's metacognitive accuracy differed from 0 for AI (Q1: $t(99) = -12.78, p < .001, d = -1.28$, Q4: $t(23) = -3.83, p < .001, d = -0.78$) as well as the no-AI group (Q1: $t(51) = -10.73, p < .001, d = -1.49$, Q4: $t(51) = -3.43, p = .001, d = -0.48$). Comparing estimates of estimated performance and performance of the first and the fourth quartile for each group, we find that the lowest quartile overestimates their performance relatively more when compared to the best-performing quartile (AI: $t(122) = 4.06, p < .001, d = 0.73$, no-AI: $t(102) = 7.25, p < .001, d = 1.42$), see Fig. 7. Therefore, the pattern of results in Study 2 regarding metacognitive accuracy also closely resembles Study 1. Comparing the confidence for correct and incorrect responses for the AI group, we find that, on average, participants are more confident for correct ($M = 85.95$, $SD$

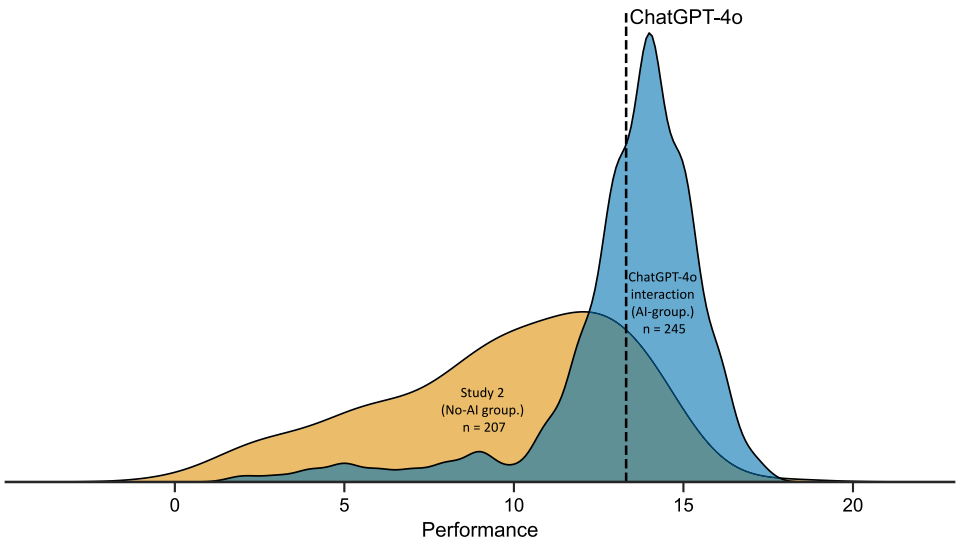



**Fig. 6.** Comparison of performance scores between the sample of participants interacting with ChatGPT and the sample of participants without AI. The blue density curve represents the distribution of performance scores in a sample of 245 participants using ChatGPT, showing a peak in performance at around 14 points. The yellow density curve corresponds to the sample of participants in the No-AI group, with lower overall performance scores. The vertical dashed line indicates the mean performance score in the ChatGPT simulation.

**Table 5**
Correlation table of metacognitive measures and AI literacy as measured by the SNAIL in study 2.

| | $\Delta EP$ | Estimate | Performance | $\Delta conf$ | $\mu conf$ | AUC | SNAIL TU | SNAIL CA | SNAIL PA |
|---|---|---|---|---|---|---|---|---|---|
| $\Delta EP$ | | | | | | | | | |
| Estimate | 0.71*** | | | | | | | | |
| Performance | −0.42*** | 0.34*** | | | | | | | |
| $\Delta conf$ | −0.05 | −0.05 | −0.01 | | | | | | |
| $\mu conf$ | 0.39*** | 0.62*** | 0.28*** | −0.09 | | | | | |
| AUC | −0.16* | −0.09 | 0.09 | 0.49*** | −0.23*** | | | | |
| SNAIL TU | 0.20** | 0.18** | −0.04 | −0.26*** | 0.21** | −0.21*** | | | |
| SNAIL CA | 0.14* | 0.21*** | 0.08 | −0.08 | 0.14* | −0.12 | 0.68*** | | |
| SNAIL PA | 0.09 | 0.19** | 0.12 | −0.13* | 0.22*** | −0.15* | 0.75*** | 0.83*** | |

**Note**. $df = 243$, $\Delta EP$ represents the difference between performance and estimated performance (metacognitive accuracy). Performance refers to the achieved task performance. $\Delta conf$ is the difference between predicted and actual confidence, while $\mu conf$ is the mean confidence (average confidence ratings). AUC refers to Area Under the Curve, with a higher AUC value indicating a more reliable confidence score in reflecting participants' correctness; SNAIL TU stands for the Technical Understanding score, SNAIL CA represents the Critical Appraisal score, and SNAIL PA is the Practical Application score.

* Indicates $p < .05$.

** Indicates $p < .01$.

*** Indicates $p < .001$.

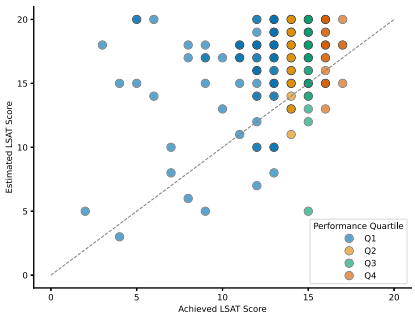
= 13.71) as compared to incorrect responses ($M = 82.57$, $SD = 15.61$; $t(244) = 5.39, p < .001, d = 0.34$). While confidence was descriptively lower for the no-AI group, we find the same pattern (correct: $M = 77.57$, $SD = 14.85$; incorrect: $M = 73.04$, $SD = 16.73$; $t(205) = 6.33, p < .001, d = 0.44$). A slight increase in confidence when accurate metacognition is incentivized is consistent with Ehrlinger et al. (2008).

Conducting a ROC analysis for each group, we found that most participants could distinguish between correct and incorrect answers. The mean AUC for the AI group ($M = .62$, $SD = 0.12$) and the no-AI group ($M = .61$, $SD = 0.11$) differed from .5 (AI: $t(244) = 15.30, p < .001, d = 0.98$; no-AI: $t(205) = 14.26, p < .001, d = 0.99$), with most people exceeding the threshold of .5 AUC (AI: 196 of 245; no-AI: 172 of 207). Consistent with Study 1, participants' mean AUC in Study 2 also fell significantly below the "acceptable" .70 benchmark (see Section 4.2.2) (AI group: $t(244) = -10.56, p < .001, d = -0.67$, no-AI group: $t(205) = -11.15, p < .001, d = -0.78$), showing a decrease of metacognitive sensitivity.

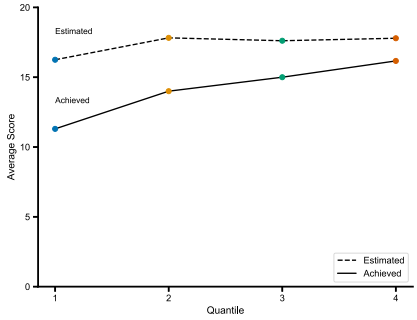
The pattern of correlations of performance, metacognitive measures, and AI literacy also resembled Study 1 in the AI group, see Table 5, and in the no-AI group, see Table A.1.

Applying our computational model using the same priors and sampler configuration as in Study 1, we find that both AI (Median = 0.63 (95% HDI [0.45, 0.85], $p_b = 0.0\%$) and no-AI (Median = 0.75 (95% HDI [0.58, 0.95], $p_b = 0.0\%$)) show a metacognitive bias without distinguishing clearly between the AI group as compared to the non-AI sample; 18.1% of posterior samples were larger in the AI group as compared to the no-AI group, see also Fig. 8A. Note that the discrepancy to Study 1, likely comes from the relatively lower precision, given the smaller sample size in Study 2. $\sigma_k$ indicating metacognitive noise was found to be above 1 for the no-AI group (Median = 1.53 (95% HDI [1.18, 1.96], $p_b = 0.1\%$), resembling the DKE pattern in Study 1, but centered around 1 for the AI group (Median = 1.13 (95% HDI [0.93, 1.36], $p_b = 10.1\%$)), see also Fig. 8B. 2.46% of posterior samples for $\sigma$ in the no-AI group exceed the AI group. Therefore, metacognitive noise does not scale the bias for the AI group, but it does for the no-AI group. We replicate the pattern of results in Study 1 again; see also Fig. 8C. The difference in shape for Fig. 8C and Fig. 5C can be explained by the difference in range, especially regarding high performance, see Fig. 2 and compare to Fig. 6.

Overall, we can replicate the results of Study 1 in Study 2. Giving an incentive for accurate metacognitive judgments did not activate a DKE pattern for participants using AI. Notably, despite the added 0.50 pounds (≈ 8% of overall compensation) performance bonus, we observed no improvement in metacognitive accuracy relative to Study




(a) *Scatter-plot of Estimated LSAT Scores as a function of Achieved LSAT Scores for the AI group in Study 2.*




(b) *Classical Dunning-Kruger Quartile-plot. Comparison of Mean Estimated and Mean Achieved Scores Across Quartiles for the AI group in Study 2.*

**Fig. 7.** Correlation of estimated and achieved LSAT score from different perspectives, individual Fig. 7(a) vs. quartile-level Fig. 7(b) for the AI group in Study 2.

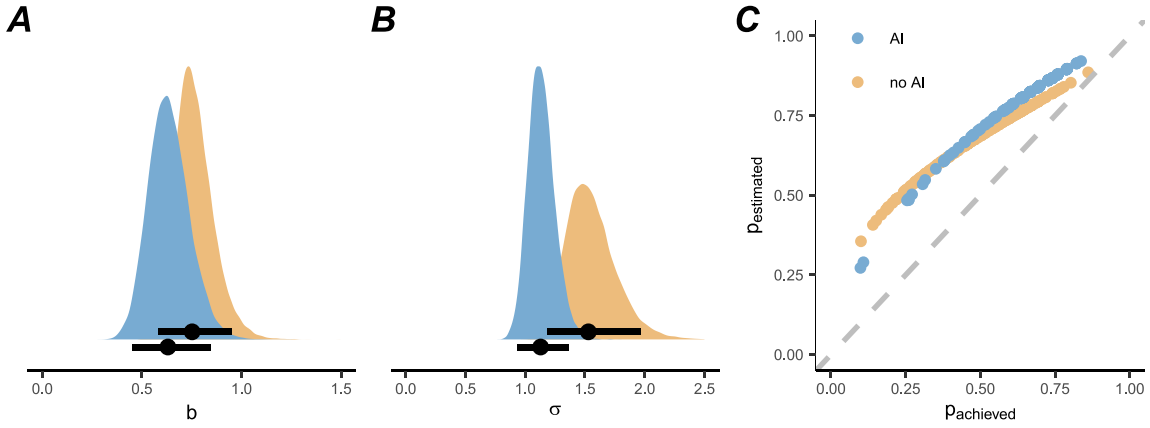



**Fig. 8.** Comparison posterior distributions with median and 95% HDI for the model parameters $b$ for bias in each group for bias (Plot A) and $\sigma$ (Plot B) for the second study. The posterior distributions of the AI group (in blue) and no AI group (in yellow). Plot C shows the average posterior predicted values for percent correct achieved (x-axis) and percent correct expected (y-axis) for each group. The s-shape around ideal metacognitive accuracy (gray line) indicates a DKE with low-performers overestimating their performance more than high-performers (yellow; no AI group).

1. Both the AI and no-AI groups continued to overestimate their performance by similar margins, suggesting that even with incentive alone, it did not substantially improved self monitoring. This pattern implies that participants were already sufficiently engaged in Study 1, and that incentive-driven effort is unlikely to be the primary driver of metacognitive calibration. Nevertheless, we can see that the absolute levels of performance overestimation are slightly larger for the no-AI group in our sample (e.g., comparing estimated performance across studies).

## 6. Discussion

This paper offers insights into metacognitive monitoring in a human–AI interaction context by examining how users with varying competence interacted with AI during logical reasoning tasks. We explored the impact of AI on metacognitive accuracy, focusing on the DKE, user confidence, and AI literacy in two studies. Our findings reveal a significant inability to assess one's performance accurately when using AI equally across our sample.

### 6.1. Effect of AI literacy on metacognition in human-AI interaction

While AI users in our sample outperformed those in Jansen et al. (2021), they consistently overestimated their performance by about four points, aligning with previous research (Kloft et al., 2024; Kosch et al., 2023; Villa et al., 2023). The moderate correlation between estimated and actual scores (Table 2), with many participants estimating their joint performance with AI higher than the most skilled in the sample (Fig. 4(a)), suggests that AI improves performance but leads to highly biased self-assessments.

This disconnect between actual and perceived performance mirrors earlier findings on overtrust and overreliance in AI systems (Kloft et al., 2024; Lu & Yin, 2021; Shekar et al., 2024). Overconfidence may impair users' ability to evaluate their performance without AI, posing challenges for designing balanced human–AI interfaces. The classic DKE, where lower performers overestimate and higher performers underestimate their performance, disappeared with AI use, suggesting that while AI levels performance, it does not correct inflated self-assessments. We found that participants, regardless of their skill level, exhibited significant overestimation. While such leveling might seem beneficial for the lowest performing quartile, who are often unskilled yet unaware (Ehrlinger et al., 2008), it raises concerns about accurate self-awareness across all performance levels. In particular, this uniform overestimation aligns with our "augmentation-hypothesis", in which AI's consistently correct outputs overshadow skill-based differences, improving low performers to a higher baseline at the cost of leading to generalized overconfidence. Metacognitive bias was doubled for the entire sample compared to Jansen et al. (2021) in Study 1, although there was a lack of a large difference in overestimation between groups in Study 2.

A skeptic might attribute the observed metacognitive distortion to the quasi-experimental sample in Study 1. Yet, the randomized replication (Study 2) still shows robust overestimation. Its smaller magnitude arises from a ceiling in self-ratings – censored in our hierarchical

Bayesian model – which nevertheless recovers a clear AI-linked bias. AI assistance also shifts the noise parameter toward unity, eliminating the skill-based damping that normally curbs high-performers' miscalibration and thus flattening the DKE slope. Converging evidence – few prompts used, the effect of AI literacy (performance estimates and confidence rise without better calibration), and many participants claiming near-perfect scores – supports the conclusion that AI use in itself affects metacognitive monitoring in our studies.

We have also found an unexpected link between AI literacy and metacognitive accuracy across both studies. Participants with higher AI literacy were less accurate in self-assessments, contradicting the assumption that higher AI literacy improves metacognitive monitoring and calibration. Familiarity with AI may enhance the better-than-average effect (Brown, 1986; Zell et al., 2020), leading to the overestimation of both relative and absolute performance.

Metacognitive sensitivity further explains these effects. Our ROC, which examines how confidence ratings are distributed between correct and incorrect responses, showed that while participants were generally confident, they tended to overestimate the correctness of incorrect responses, indicating low metacognitive sensitivity. This suggests that while participants felt assured in their answers, their metacognitive sensitivity (*i.e.* how well their confidence distinguished correctness on the trial level) was consistently low. Therefore, we can find further evidence that participants did not monitor their performance when using the AI system to complete the task.

Qualitative data revealed varied perceptions of AI's role, from a tool to a teammate, but these differences did not descriptively show an effect on performance or metacognitive accuracy, see Table 4, contradicting theories that suggest interaction framing impacts outcomes (Pataranutaporn et al., 2023; Villa et al., 2023). Regardless of user perception, the core metacognitive challenges in HAI persist. These findings suggest that while AI assistance can improve task performance, it does not proportionally enhance individuals' metacognitive abilities, highlighting a disconnect between cognitive performance gains and self-evaluative insight.

Decomposing overall AI-literacy into its three sub-scales (see Table 1) reveals that Technical Understanding (TU) (*i.e.* familiarity with prompting, parameter settings, and API workflows) is significantly associated with greater mean overestimation. By contrast, Critical Appraisal (CA) and Practical Application (PA) show small correlations with overestimation. Notably, CA and PA instead predict higher mean confidence without improving the ability to discriminate between correct and incorrect judgments (AUC), suggesting a disconnect between global self-belief and local monitoring (disconnect between global and local knowledge). This pattern aligns with the illusion of explanatory depth (Fisher & Oppenheimer, 2021), where procedural fluency provides a misleading sense of ability. In our context, users who feel technically proficient while interacting with AI tend to overestimate their performance (even though their trial-level sensitivity remains unchanged). Targeting this gap by, for example, prompting users to justify AI suggestions, may help bridge global confidence and local accuracy.

### 6.2. Integration into theory

High metacognitive bias leads users to overestimate their performance and over-rely on AI systems (Ma et al., 2024), reducing their ability to critically monitor HAI outcomes (Tankelevitch et al., 2024). From a computational rationality perspective (Oulasvirta et al., 2022), this bias may be an adaptive response to AI presence, as participants may optimize perceived utility (e.g., efficiency) rather than monitoring HAI. This is supported by low metacognitive sensitivity and participants' reliance on copy-pasting rather than higher-level metacognitive strategies (e.g., AI as a collaborator). This aligns with Villa et al. (2023), who found reduced error-processing when participants believed they were using a sham AI. Therefore, our study provides further evidence of diminished metacognitive monitoring in HAI.

A lack of HAI monitoring explains several effects. First, it clarifies why AI use has been linked to adverse learning outcomes (Abbas et al., 2024; Bastani et al., 2024); users are overly optimistic and fail to monitor evolving joint performance. Second, while AI tools offer perceived empowerment and efficiency (Kloft et al., 2024; Kosch et al., 2023; Villa et al., 2023), the lack of reflection hinders users' ability to assess real benefits (placebo effect). Additionally, it explains persistent overreliance and overtrust (Klingbeil et al., 2024), and why AI explanations are rarely integrated into behavior (Bansal et al., 2021; Ghassemi et al., 2021; Wang & Yin, 2021). Though we can offer methods to improve metacognitive judgments grounded in HCI and psychological research (refer to Table 6), they may provide only temporary solutions. Rafner et al. (2022) calls for a systemic, long-term strategy, considering cognitive and metacognitive deskilling risks at individual, team, and organizational levels. This emphasizes the need for strategies that foster cognitive resilience and critical engagement with AI over short-term fixes targeting immediate metacognitive deficits.

### 6.3. Limitations

While our study provides valuable insights into metacognitive monitoring in HAI, several limitations may affect the generalizability and interpretation of our findings.

As Study 1 contrasted two independent datasets, it should be considered a quasi-experimental comparison, not a true randomized experiment. The AI group was recruited for the present study, whereas the no-AI benchmark relies on the open dataset of Jansen et al. (2021). As the two samples were gathered at different times and participants were not randomly assigned to "AI" versus "No-AI" conditions, any differences we observe are descriptive associations rather than causal effects. We therefore interpret Study 1 as providing suggestive, not causal, evidence, and use Study 2's randomized design to probe the effect of AI support more rigorously.

Secondly, the apparent absence of the DKE ("being unskilled and unaware" (Dunning, 2011) in our study may not fully reflect the underlying cognitive dynamics. AI interaction may have made participants equally skilled yet (still) unaware of their performance rather than eliminating the DKE. AI-enhanced performance might have decoupled metacognitive judgments from cognitive performance (e.g., high confidence, low sensitivity, and high bias).

Third, we relied on LSAT questions as our logical reasoning task. While these assess logical reasoning, they may not capture the diversity of real-world reasoning skills or domains. Furthermore, these tasks might overlap with the AI's training data, limiting generalizability. However, since ChatGPT-4o's performance was imperfect (68.25%), we believe our findings still apply to other tasks.

Fourth, our focus on LSAT-based reasoning limits the scope of metacognitive biases across domains. Future research should use diverse tasks (e.g., writing creative texts with an LLM) to examine how AI interaction affects metacognitive monitoring and whether improvements in accuracy generalize across tasks.

Fifth, we found little effect of different AI strategies on metacognitive accuracy and performance. The AI's role, whether as a tool, collaborator, or teammate, did not impact participants' performance evaluation (see again Table 4). Future research should explore how different AI roles (e.g., tool vs. collaborator) influence metacognitive accuracy and task performance.

Sixth, AI literacy was self-reported, introducing the possibility of over- or underestimation due to psychological biases. In particular, people may experience illusions of explanatory depth, believing they understand AI better than they actually do, akin to the better-than-average effect (Fisher & Oppenheimer, 2021; Zell et al., 2020). This is closely related to the DKE, where individuals with low competence often remain unaware of their limitations (Kruger & Dunning, 1999).

**Table 6**
Issues, consequences, and design principles to address impaired metacognitive monitoring in human–AI interaction.

| Metacognitive issue | Consequences | Design principles |
|---|---|---|
| Overreliance on AI outputs | Users trust AI outputs without critical assessment, leading to reduced self-reflection and failure to notice AI errors. | • Confidence calibration to align user confidence with AI output uncertainty (Ma et al., 2024)<br>• AI uncertainty visualization to make AI output reliability transparent (Beauxis-Aussalet et al., 2021; Prabhudesai et al., 2023)<br>• Explanatory AI interfaces to clarify AI decision-making processes and enable users to assess validity (Karran et al., 2022) |
| Loss of metacognitive monitoring | Users are unable to accurately assess their own performance or monitor task progress, especially in complex decision-making tasks. | • Post-task reflection to encourage users to evaluate their performance after interacting with AI (for a starting point, see Tankelevitch et al. (2024))<br>• Cognitive forcing strategies such as prompts to promote critical thinking and reduce automatic reliance on AI outputs (Buçinca et al., 2021) |
| Illusion of understanding | AI-literate users tend to over-rely and over-trust on AI outputs. | • "Explain-back" micro-task before submission to help calibrate illusions of knowledge (Fisher & Oppenheimer, 2021). |

In our context, participants who self-rated higher AI literacy might also have been more prone to overconfidence, conflating AI's outputs with their own abilities. Consequently, the correlation between AI literacy and performance overestimation could be amplified simply because people who think they know AI better also tend to inflate their self-assessment. In line with this, self-reported AI literacy might not directly translate into effective interactions with AI. Thus, how AI literacy links to performance and metacognition deserves to be further explored in future studies.

Seventh, participants were required to prompt ChatGPT at least once and then proceed with their decision, which may not reflect naturalistic usage patterns. In real-world scenarios, users might consult AI tools multiple times, ignore them entirely, or encounter more proactive AI systems offering unsolicited suggestions. Future studies should examine how varying degrees of user engagement with the AI and the AI's proactiveness influence metacognitive accuracy, sensitivity, and bias.

Finally, exploring long-term learning scenarios, such as those examined in Bastani et al. (2024), could illuminate how metacognitive processes evolve as individuals repeatedly interact with AI systems. Over extended periods, accurate metacognitive monitoring may become increasingly vital for achieving sustained performance gains (e.g., learning calculus with AI assistance and then taking an exam); our study cannot speak to the role of metacognition in AI-mediated learning contexts.

### 6.4. Implications

Following Van Berkel and Hornbæk (2023), we identify three types of implications of our work: theory, design, and methodology.

Regarding theory, while biases in human–AI interaction have been studied (Bertrand et al., 2022; Haliburton et al., 2024; Kloft et al., 2024; Liu et al., 2019), little research has empirically examined the metacognitive mechanisms driving these biases. Even Tankelevitch et al. (2024), who underscore the importance of metacognition in generative AI contexts, do not provide a dedicated empirical investigation into these underlying processes. Applying metacognition theory to HAI (e.g., see Colombatto & Fleming, 2023) thus offers a new perspective for improving metacognitive monitoring in interaction design. Our findings additionally suggest that AI literacy alone is insufficient for achieving optimal metacognitive abilities in HAI.

For the design of HAI interaction and its behavioral analysis, we propose that research needs to develop a new HAI interaction model that integrates metacognition and new design concepts. In the short term, designers should adjust interaction models, similar to Buçinca et al. (2021), and consider broader sociotechnical risks (Rafner et al., 2022). In the long term, HCI must develop a specific HAI interaction model that enhances metacognitive functions, e.g., to support knowledge-based interactions (Rasmussen, 1983). Based on the results from study 2, participants who used the LLM solved, on average, three more items, yet overestimated their score by four points. We thus present a design implication (see Table 6), introducing a simple "explain-back" task before accepting the model's answer, requiring users to briefly re-state its logic in simple language. Prior HCI work shows that this lowers overconfidence and directly targets the illusion of understanding (Fisher & Oppenheimer, 2021). Future work should systematically vary prompt frequency to test whether deeper interaction with AI improves users' ability to discriminate correct from incorrect responses. Furthermore, interface designers should calibrate assistance to those empirical variations, e.g., surface real-time accuracy feedback once overestimation exceeds four points, progressively adjusting assistance level through real-time sensing of metacognitive states (e.g., user agency) and adapting interfaces to cognitive states (Chiossi et al., 2023; Villa et al., 2024). Such mechanisms could be crucial for optimizing HAI.

Methodologically, our study introduces a new approach to analyzing biased decision-making in repeated AI interactions. Using computational methods, researchers can distinguish general biases from those emerging in post-decision processes (e.g., increased metacognitive noise), allowing for a more precise analysis of bias development in HAI.

## 7. Conclusion

We found that participants using an LLM had improved logical reasoning performance as compared to no AI, but that cognitive performance gains did not scaffold metacognition. With AI use, the DKE was eliminated. Based on these findings, we suggest developing new interfaces for interactive AI that are designed to enhance metacognition, allowing users to monitor their performance more accurately.

**Table A.1**
Correlation table of metacognitive measures in study 2 for the no-AI group.

| | $\Delta EP$ | Estimate | Performance | $\Delta conf$ | $\mu conf$ | AUC |
|---|---|---|---|---|---|---|
| $\Delta EP$ | | | | | | |
| Estimate | 0.68*** | | | | | |
| Performance | −0.54*** | 0.25*** | | | | |
| $\Delta conf$ | −0.22** | −0.15* | 0.11 | | | |
| $\mu conf$ | 0.38*** | 0.60*** | 0.20** | −0.18** | | |
| AUC | −0.12 | −0.13 | 0.01 | 0.57*** | −0.24*** | |

**Note.** $df = 205$, $\Delta EP$ represents the difference between performance and estimated performance (metacognitive accuracy). Performance refers to the achieved task performance. $\Delta conf$ is the difference between predicted and actual confidence, while $\mu conf$ is the mean confidence (average confidence ratings). AUC refers to Area Under the Curve, with a higher AUC value indicating a more reliable confidence score in reflecting participants' correctness.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

## CRediT authorship contribution statement

**Daniela Fernandes:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Steeven Villa:** Writing – review & editing, Visualization, Software, Data curation. **Salla Nicholls:** Writing – review & editing, Visualization, Validation, Software, Methodology, Data curation. **Otso Haavisto:** Writing – review & editing, Investigation, Software, Data curation, Conceptualization. **Daniel Buschek:** Writing – review & editing, Software. **Albrecht Schmidt:** Writing – review & editing, Conceptualization. **Thomas Kosch:** Writing – review & editing, Visualization, Software, Methodology. **Chenxinran Shen:** Writing – review & editing, Visualization. **Robin Welsch:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors do not declare any conflict of interest with the submitted work.

## Acknowledgments

This work was supported by The Finnish Doctoral Program Network in Artificial Intelligence, Finland, AI-DOC [decision number VN/3137/2024-OKM-6]; This work relates to the upcoming ERC project AmplifAI (grant agreement No. 101217557).

## Appendix A. Tables

See Table A.1.

## Appendix B. Figures

See Fig. B.1.

## Data availability

The research software can be found under the following repositories: https://github.com/aaltoengpsy/interface-frontend and https://github.com/aaltoengpsy/interface-backend All data collected for the purpose of our paper and analysis scripts can be found at https://osf.io/svax9/overview.

## References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, *21*(1), 10. http://dx.doi.org/10.1186/s41239-024-00444-7.

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. http://dx.doi.org/10.1016/j.tics.2017.05.004.

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386. http://dx.doi.org/10.1016/j.cognition.2015.10.006, URL https://www.sciencedirect.com/science/article/pii/S0010027715300846.

Alexandre e Castro, P. (2024). What neurohacking can tell us about the mind: Cybercrime, mind upload and the artificial extended mind. In P. Alexandre e Castro (Ed.), *Challenges of the technological mind: between philosophy and technology* (pp. 43–62). Cham: Springer Nature Switzerland, http://dx.doi.org/10.1007/978-3-031-55333-2_4.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569–576. http://dx.doi.org/10.1037/0096-3445.136.4.569, URL https://doi.apa.org/doi/10.1037/0096-3445.136.4.569.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (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* (pp. 1–16). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3411764.3445717.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. *4895486*, http://dx.doi.org/10.2139/ssrn.4895486, Available At SSRN 4895486.

Beauxis-Aussalet, E., Behrisch, M., Borgo, R., Chau, D. H., Collins, C., Ebert, D., El-Assady, M., Endert, A., Keim, D. A., Kohlhammer, J., et al. (2021). The role of interactive visualization in fostering trust in AI. *IEEE Computer Graphics and Applications*, *41*(6), 7–12. http://dx.doi.org/10.1109/MCG.2021.3107875.

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 78–91). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3514094.3534164.

Bosch, E., Welsch, R., Ayach, T., Katins, C., & Kosch, T. (2024). The illusion of performance: The effect of phantom display refresh rates on user expectations and reaction times. In *Chi ea '24*, *Extended abstracts of the 2024 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3613905.3650875.

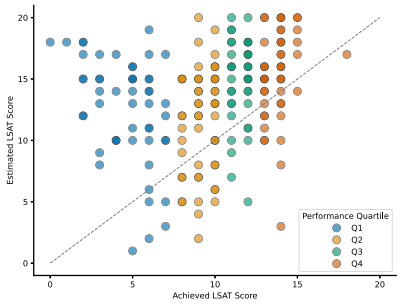
Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition*, *4*(4), 353–376. http://dx.doi.org/10.1521/soco.1986.4.4.353.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, *5*(Cscw1), http://dx.doi.org/10.1145/3449287.

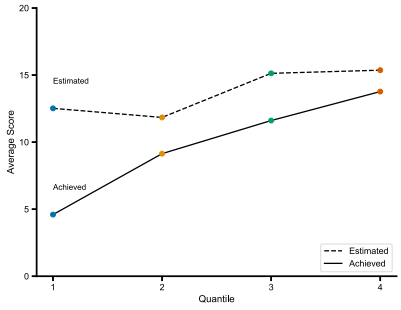
Bürkner, P.-C. (2017). Brms: An r package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), http://dx.doi.org/10.18637/jss.v080.i01.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons.. *Journal of Personality and Social Psychology*, *90*(1), 60. http://dx.doi.org/10.1037/0022-3514.90.1.60.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), http://dx.doi.org/10.18637/jss.v076.i01.




(a) *Scatter-plot of Estimated LSAT Scores as a function of Achieved LSAT Scores for the no-AI group in Study 2.*




(b) *Classical Dunning-Kruger Quartile-plot. Comparison of Mean Estimated and Mean Achieved Scores Across Quartiles for the no-AI group in Study 2.*

**Fig. B.1.** Correlation of estimated and achieved LSAT score from different perspectives, individual Fig. B.1(a) vs. quartile-level Fig. B.1(b) for the no-AI group in Study 2.


Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, *1*(2), 74–78. http://dx.doi.org/10.1038/s42256-019-0020-9.

Chiossi, F., Turgut, Y., Welsch, R., & Mayer, S. (2023). Adapting visual complexity based on electrodermal activity improves working memory performance in virtual reality. *Proc. ACM Hum.-Comput. Interact.*, *7*, http://dx.doi.org/10.1145/3604243.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, http://dx.doi.org/10.1093/acprof:oso/9780195333213.001.0001.

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology*, *12*(3), 297–298. http://dx.doi.org/10.1080/17439760.2016.1262613.

Clayton, D. A., Eguchi, M. M., Kerr, K. F., Miyoshi, K., Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2023). Are Pathologists Self-Aware of Their Diagnostic Accuracy? Metacognition and the Diagnostic Process in Pathology. *Medical Decision Making*, *43*(2), 164–174. http://dx.doi.org/10.1177/0272989X221126528, _eprint: DOI: 10.1177/0272989X221126528.

Colombatto, C., & Fleming, S. (2023). Illusions of confidence in artificial systems. http://dx.doi.org/10.31234/osf.io/mjx2v,

Dang, H., Goller, S., Lehmann, F., & Buschek, D. (2023). Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3544548.3580969.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115. http://dx.doi.org/10.1016/j.cognition.2010.09.012.

Dix, A. (2022). Bayesian statistics. In J. H. Williamson, A. Oulasvirta, P. O. Kristensson, & N. Banovic (Eds.), *Bayesian methods for interaction and design* (pp. 81–114). Cambridge University Press, http://dx.doi.org/10.1017/9781108874830.004.

Draxler, F., Buschek, D., Tavast, M., Hämäläinen, P., Schmidt, A., Kulshrestha, J., & Welsch, R. (2023). Gender, age, and technology education influence the adoption and appropriation of LLMs. arXiv:2310.06556.

Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., & Welsch, R. (2024). The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Trans. Comput.-Hum. Interact.*, *31*(2), http://dx.doi.org/10.1145/3637875.

Dunning, D. (2011). The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology: vol. 44*, (pp. 247–296). Elsevier, http://dx.doi.org/10.1016/B978-0-12-385522-0.00005-6.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121.

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (pp. 1–6). http://dx.doi.org/10.1145/3290607.3312787.

Engelbart, D. C. (1962). Augmenting human intellect: A conceptual framework. (p. 21). Menlo Park, CA.

Fiedler, K., Ackerman, R., & Scarampi, C. (2019). Metacognition: Monitoring and controlling one's own knowledge, reasoning and decisions. In R. J. Sternberg, & J. Funke (Eds.), *The psychology of human thought: an introduction* (pp. 89–111). Heidelberg University Publishing, http://dx.doi.org/10.17885/heiup.470.c6669.

Fisher, M., & Oppenheimer, D. M. (2021). Harder than you think: How outside assistance leads to overconfidence. *Psychological Science*, *32*(4), 598–610. http://dx.doi.org/10.1177/0956797620975779.

Fleming, S. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*, 241–268. http://dx.doi.org/10.1146/annurev-psych-022423-032425.

Fleming, S., & Lau, H. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. http://dx.doi.org/10.3389/fnhum.2014.00443.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. http://dx.doi.org/10.1214/ss/1177011136.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. http://dx.doi.org/10.1016/S2589-7500(21)00208-9.

Gignac, G. E. (2024). Rethinking the Dunning-Kruger effect: Negligible influence on a limited segment of the population. *Intelligence*, *104*, Article 101830. http://dx.doi.org/10.1016/j.intell.2024.101830.

Gignac, G. E., & Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives. *Intelligence*, *104*, Article 101832. http://dx.doi.org/10.1016/j.intell.2024.101832.

Gignac, G. E., & Zajenkowski, M. (2020). The dunning-kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, *80*, Article 101449. http://dx.doi.org/10.1016/j.intell.2020.101449.

Haliburton, L., Ghebremedhin, S., Welsch, R., Schmidt, A., & Mayer, S. (2024). Investigating labeler bias in face annotation for machine learning. In *HHAI 2024: hybrid human AI systems for the social good* (pp. 145–161). IOS Press, http://dx.doi.org/10.3233/FAIA240191.

Hoijtink, H., & van de Schoot, R. (2018). Testing small variance priors using prior-posterior predictive p values. *Psychological Methods*, *23*(3), 561–569. http://dx.doi.org/10.1037/met0000131.

Hou, Y., Tamoto, H., & Miyashita, H. (2024). "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents. In *chi ea '24, Extended abstracts of the 2024 CHI conference on human factors in computing systems* (pp. 1–7). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3613905.3650839.

Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L. H., & Quinn, G. (2023). *30, Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making* (5), (pp. 71:1–71:29). http://dx.doi.org/10.1145/3534561,

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756–763. http://dx.doi.org/10.1038/s41562-021-01057-0.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. http://dx.doi.org/10.1037/0033-2909.114.1.3, Publisher: American Psychological Association (APA).

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, *12*(6), 518–527. http://dx.doi.org/10.1038/s41558-022-01377-7.

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P.-M. (2022). Designing for confidence: The impact of visualizing artificial intelligence decisions. *Frontiers in Neuroscience*, *16*, Article 883385. http://dx.doi.org/10.3389/fnins.2022.883385.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society, Series A*, *382*(2270), Article 20230254. http://dx.doi.org/10.1098/rsta.2023.0254.

Kay, M., Nelson, G. L., & Hekler, E. B. (2016). Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4521–4532). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2858036.2858465.

Khettab, S. A. (2019). David chalmers -the extended mind. http://dx.doi.org/10.13140/RG.2.2.14997.14560.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, Article 108352. http://dx.doi.org/10.1016/j.chb.2024.108352.

Kloft, A. M., Welsch, R., Kosch, T., & Villa, S. (2024). "Ai enhances our performance, I have no doubt this one will do the same": The placebo effect is robust to negative descriptions of AI. In *Proceedings of the CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3613904.3642633.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning.. *Journal of Experimental Psychology: General*, *126*(4), 349.

Kosch, T., Welsch, R., Chuang, L., & Schmidt, A. (2023). The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer-Human Interaction*, *29*(6), 1–32. http://dx.doi.org/10.1145/3529225.

Kruger, J., & Dunning, D. (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), 1121. http://dx.doi.org/10.1037/0022-3514.77.6.1121.

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the "scale for the assessment of non-experts' AI literacy"–an exploratory factor analysis. *Computers in Human Behavior Reports*, *12*, Article 100338. http://dx.doi.org/10.1016/j.chbr.2023.100338.

Liu, Y., He, F., Zhang, H., Rao, G., Feng, Z., & Zhou, Y. (2019). How well do machines perform on IQ tests: a comparison study on a large-scale dataset. In *Ijcai* (pp. 6110–6116). http://dx.doi.org/10.24963/ijcai.2019/846.

Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–16). http://dx.doi.org/10.1145/3411764.3445562.

Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. (2024). "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI conference on human factors in computing systems* (pp. 1–20). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3613904.3642671.

Mahdavi, M. (2014). An overview: Metacognition in education. *International Journal of Multidisciplinary and Current Research*, *2*(6), 529–535.

Mak, K.-K., Wong, Y.-H., & Pichika, M. R. (2023). Artificial intelligence in drug discovery and development. *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*, 1–38. http://dx.doi.org/10.1016/j.drudis.2020.10.010.

McIntosh, R. D., Fowler, E. A., Lyu, T., & Della Sala, S. (2019). Wise up: Clarifying the role of metacognition in the dunning-kruger effect.. *Journal of Experimental Psychology: General*, *148*(11), 1882.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259–265. http://dx.doi.org/10.1038/s41586-023-05881-4.

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., & Agreda, S. G. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change*, *181*, Article 121763.

Oulasvirta, A., Jokinen, J. P., & Howes, A. (2022). Computational rationality as a theory of interaction. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1–14). http://dx.doi.org/10.1145/3491102.3517739.

Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, *5*(10), 1076–1086. http://dx.doi.org/10.1038/s42256-023-00720-7, Publisher: Nature Publishing Group.

Perera, M. (2024). Enhancing Productivity Applications for People who are Blind using AI Assistants. In *chi ea '24*, *Extended abstracts of the 2024 CHI conference on human factors in computing systems* (pp. 1–6). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3613905.3638180.

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V., & Banovic, N. (2023). Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th international conference on intelligent user interfaces* (pp. 379–396). http://dx.doi.org/10.1145/3581641.3584033.

Rafner, J., Dellermann, D., Hjorth, A., Veraszto, D., Kampf, C., MacKay, W., & Sherson, J. (2022). Deskilling, upskilling, and reskilling: a case for hybrid intelligence. *Morals & Machines*, *1*(2), 24–39. http://dx.doi.org/10.5771/2747-5174-2021-2-24.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, (3), 257–266. http://dx.doi.org/10.1109/TSMC.1983.6313160.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688.

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science.. *Psychological Methods*, *26*(1), 103–126. http://dx.doi.org/10.1037/met0000275.

Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A., & Maes, P. (2024). People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. http://dx.doi.org/10.48550/arXiv.2408.15266.

Shi, H., & Yin, G. (2020). Reconnecting p-value and posterior probability under one-and two-sided tests. *The American Statistician*, *75*, 265–275. http://dx.doi.org/10.1080/00031305.2020.1717621.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504.

Shultz, M. M., & Zedeck, S. (2011). Predicting lawyer effectiveness: Broadening the basis for law school admission decisions. *Law & Social Inquiry*, *36*(3), 620–661. http://dx.doi.org/10.1111/j.1747-4469.2011.01245.x.

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), Article e2111547119. http://dx.doi.org/10.1073/pnas.2111547119, URL https://www.pnas.org/doi/abs/10.1073/pnas.2111547119. Publisher: Proceedings of the National Academy of Sciences.

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. *57*, In *Proceedings of the CHI conference on human factors in computing systems* (pp. 1–24). ACM, http://dx.doi.org/10.1145/3613904.3642902.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. http://dx.doi.org/10.3758/s13421-011-0104-1.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1–11.

Van Berkel, N., & Hornbæk, K. (2023). Implications of human-computer interaction research. *Interactions*, *30*(4), 50–55. http://dx.doi.org/10.1145/3600103.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), http://dx.doi.org/10.1038/s43586-020-00001-2.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(Cscw1), 1–38. http://dx.doi.org/10.1145/3579605.

Villa, S., Kosch, T., Grelka, F., Schmidt, A., & Welsch, R. (2023). The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior. *Computers in Human Behavior*, *146*, Article 107787. http://dx.doi.org/10.1016/j.chb.2023.107787.

Villa, S., Welsch, R., Denisova, A., & Kosch, T. (2024). Evaluating interactive AI: Understanding and controlling placebo effects in human-AI interaction. In *Chi ea '24*, *Extended abstracts of the 2024 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3613905.3636304.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education*, *8*(2), 157–186. http://dx.doi.org/10.1207/s15324818ame0802_4.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *chi ea '20*, *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–6). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3334480.3381069.

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. (pp. 318–328). http://dx.doi.org/10.1145/3397481.3450650.

Yang Hansen, K., Thorsen, C., Radišić, J., Peixoto, F., Laine, A., & Liu, X. (2024). When competence and confidence are at odds: a cross-country examination of the dunning–kruger effect. *European Journal of Psychology of Education*, 1–23.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3544548.3581388.

Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis.. *Psychological Bulletin*, *146*(2), 118. http://dx.doi.org/10.1037/bul0000218.

Zulfikar, W. D., Chan, S., & Maes, P. (2024). Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI conference on human factors in computing systems* (pp. 1–18). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3613904.3642450.