Accurate but Not Confident or Confident but Not Accurate? Cognitive Offloading Impairs Confidence Calibration in Human-AI Teams

Charlotte Cheatle (c.cheatle@surrey.ac.uk)[1] & Adrian P. Banks (a.banks@surrey.ac.uk)[1]
[1]School of Psychology, University of Surrey

Abstract

Appropriate AI reliance requires users to accurately evaluate their own performance in order to discern whether to retain or defer responsibility. While underconfidence is a known driver of automation bias, little is known about how collaboration with AI itself influences users' metacognition. We investigated how cognitive offloading different stages of the decision-making process affects confidence calibration. Participants completed diagnostic decision-making tasks with varying levels of memory and judgement support. Overall, cognitive offloading impaired confidence calibration, with unaided decision makers showing the greatest alignment between confidence and accuracy. Importantly, offloading different cognitive processes produced distinct metacognitive biases: judgement offloading led to *overconfidence*, whereas combined offloading of memory and judgement processes led to *underconfidence*. These findings demonstrate that AI support can disrupt user confidence calibration in systematic ways, depending on the type and extent of cognitive delegation. The results highlight metacognitive miscalibration as a critical and underexplored consequence of human-AI collaboration.

Keywords: cognitive offloading; artificial intelligence; confidence calibration; AI augmented decision making; metacognition

Introduction

The rapid integration of artificial intelligence (AI) into everyday life is reshaping how people approach cognitive tasks. Promising greater efficiency and productivity, AI systems increasingly support with information retrieval, problem solving, and decision making. Nevertheless, inappropriate reliance remains a barrier to synergistic team performance, with human-AI teams often failing to outperform the best of either agent alone (Vaccaro et al., 2024). Appropriate reliance requires an accurate mental model of the distributed cognitive system (Bansal et al., 2019; Srivastava et al., 2025). That is, an individual must have an awareness of their own and the AI's capabilities and constraints to discern whether to retain or defer responsibility. Accordingly, much empirical research has been dedicated to understanding and improving user awareness of AI behaviour, for example, through explanations (Vasconcelos et al., 2023) or AI confidence levels (Ishizu et al., 2024; D. Lee et al., 2025). However, these interventions have yielded limited benefits (Li et al., 2025; Schemmer et al., 2022). Less attention has been paid to the humans' ability to monitor and evaluate their own capabilities in augmented decision making, known as metacognition. Metacognitive evaluations play a crucial role in cognitive offloading strategies, with underconfidence driving overreliance on AI systems (Ma et al., 2024). Critically however, how offloading and reliance on cognitive aids in turn shape metacognition is not well understood, despite the profound implications this has for fostering appropriate reliance and optimising human-AI collaboration. Therefore, the present study addresses this gap by investigating the effect of offloading on metacognition. We focus on confidence calibration, the degree of alignment between subjective confidence and objective accuracy, in an AI-assisted decision-making task.

Cognitive Offloading

Cognitive offloading refers to the process of shifting cognitive demands onto external tools (Risko & Gilbert, 2016). This phenomenon, traditionally examined in the context of working memory, reduces cognitive load and allows resources to be redistributed to other cognitive tasks. Thus, externalising memory enhances both immediate recall (Burnett & Richmond, 2025) and secondary task performance (Grinschgl et al., 2023; Storm & Stone, 2015). The advent of generative AI extends offloading from simple information storage to the dynamic delegation of cognitive tasks. AI systems can now support with judgement and reasoning, reducing mental load and improving decision accuracy (Buettner, 2013; Schemmer et al., 2022). Benefits which are amplified under higher task difficulty (Peper et al., 2023; Vasconcelos et al., 2023). Previous research, however, has examined offloading memory or judgement in isolation. Based on a normative model of optimal multi-attribute choice, decision making involves sequential stages: identifying, recalling, weighting, integrating, and comparing cues (Payne et al., 1993). Both memory and judgement therefore represent core cognitive components: accurate recall of cues and the effective weighting of their importance jointly determine decision quality. Thus, the significance of offloading in decision-making contexts is twofold: external aids can support with memory and/or judgement. Accordingly, we investigate increasing levels of cognitive delegation, comparing memory offloading, judgement offloading, and combined memory and judgement offloading, in a decision-making task, to identify their distinct contributions. This study provides the first empirical test of how distributing different cognitive processes between humans and AI influences confidence calibration.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Metacognition

Metacognition, often termed 'thinking about thinking', refers to how individuals evaluate their own cognition. This comprises general, long-term assumptions about one's cognitive abilities (metacognitive beliefs) and concurrent, task-specific evaluations of cognitive performance (metacognitive experiences) (Flavell, 1979). In offloading research, metacognitive beliefs have primarily been considered a determinant of reliance on external aids (Gilbert et al., 2020; Küper et al., 2025; Ma et al., 2024). Indeed, self-confidence is a greater predictor of AI reliance than confidence in the AI model itself (Chong et al., 2022). However, Fügener and colleagues (2022) maintain that humans lack sufficient metaknowledge needed to accurately assess their own abilities when deciding whether to delegate tasks to AI, with overconfidence a well-established bias in decision making (Soll & Klayman, 2004; Tidwell et al., 2016). Thus, erroneous metacognitive beliefs often lead to inappropriate reliance strategies (Chong et al., 2022; He et al., 2023). Therefore, accurate confidence calibration is a necessary prerequisite to achieving appropriate reliance.

What remains unclear is the inverse relationship: how external support reciprocally influences metacognition. There is evidence that the use of external aids, like the internet, distort general metacognitive beliefs about one's cognitive abilities. Individuals encode less information when they expect it to be easily retrievable (Sparrow et al., 2011), yet view this searchable information as their own knowledge (Fisher et al., 2015) and consequently overestimate their future unaided abilities (Ward, 2021). This effect, termed the illusion of knowledge, has also been observed when humans partner with AI systems, with individuals overestimating their abilities even when the AI is harmful to performance (Fisher & Oppenheimer, 2021). The authors attribute this poor calibration to the ambiguity around who deserves credit for success or blame for failure when an individual receives external assistance. With accountability a significant ethical concern for human-AI teaming, understanding how users perceive responsibility and assess team performance is of critical importance.

Despite this, very little is known about the influence of cognitive support on metacognitive experiences, that is, task-specific evaluations present while performing the task. Taudien and colleagues (2022, p. 251) claim that AI recommendations provide "false comfort", finding elevated confidence for both correct and incorrect AI-supported judgements. Nevertheless, it is unclear from their analyses whether these increases in confidence are justified due to coinciding increases in accuracy provided by AI. This lack of evidence surrounding the influence of offloading on confidence calibration presents a critical gap in our understanding of human-AI teaming and informs the basis of the present experiment. Here, we offer two competing hypotheses for the potential effects of cognitive offloading on confidence calibration. The purpose of the present study is to test these opposing predictions.

On the one hand, evidence for offloading's role in inflating metacognitive beliefs would lead to us to believe that offloading would similarly inflate metacognitive experiences. Likewise, literature on group decision making argues that individuals exhibit poorer confidence calibration when working in teams due to inflated confidence in incorrect decisions, termed the *two heads are worse than one effect* (Puncochar & Fox, 2004). M. D. Lee and Dry (2006) also found that receiving advice elevates confidence, but the degree of this inflation depends on both the accuracy and the frequency of the advice. Interestingly, overconfidence peaks when advice is only moderately accurate but reasonably frequent, rather than extremely accurate but less frequent. Based on these findings, we could predict that participants offloading to AI would exhibit overconfidence.

An alternative perspective, grounded in research on automation bias, would suggest that individuals may over-rely on AI recommendations and under-rely on their own judgement. Rather than engaging analytically, users often uncritically accept AI outputs, displaying passive cognitive engagement and minimal involvement in the decision process (Buçinca et al., 2021; Gajos & Mamykina, 2022). This tendency reflects a broader pattern of cognitive disengagement observed when external systems augment cognition: access to external information storage diminishes motivation to engage with new information (Kleinberg & Lau, 2021), encode information (Sparrow et al., 2011) and engage in deep thinking processes (Gerlich, 2025). Preliminary evidence also suggests that collaboration with AI reduces task engagement at a neural level, as well as perceived ownership over the final outcome (Kosmyna et al., 2025). On this basis, in delegating significant components of the decision-making process, individuals may lose insight into the underlying reasoning, and consequently, lose confidence that the decision is correct. Thus, while they may readily endorse AI's judgements, without this self-generated reasoning they may not be able to substantiate them when asked to evaluate the judgement's accuracy. Supporting this view, Fisher and Oppenheimer (2021) found that individuals only conflate team success with their own when they have actively contributed to the task. Therefore, while AI advice may improve objective accuracy, this may not bring with it increased confidence, translating into poor confidence calibration in the form of underconfidence.

The Present Study

Together, these competing perspectives capture a critical question in human-AI teaming about the effects that augmentation is having on the way we make decisions and evaluate those decisions. The present study directly evaluates these opposing hypotheses to determine whether cognitive offloading of memory and judgement processes produces overconfidence or underconfidence - a question with significant implications for how humans calibrate reliance on AI systems in future decision-making contexts. To test these predictions, we conducted a study using a diagnostic decision-making paradigm. Participants were randomly

assigned to one of four conditions manipulating offloading (memory; judgement; combined memory and judgement; no offloading/control). They each completed tasks at varying levels of difficulty (high; low) and rated their confidence in their final decision. Confidence calibration, calculated as the difference between subjective confidence and objective accuracy, was then compared between groups.

Method

Participants and Design

We recruited 153 participants through Prolific, compensating participants £2.50 for their time. Data from 47 participants were removed due to pre-defined exclusion criteria: failing at least one attention check, ReCAPTCHA score < 0.3, or completion time exceeding 2SD above the mean. The final sample comprised 106 participants (67% female; M_{age} = 39.1, SD_{age} = 13.7). An a priori power analysis conducted using G*Power (Faul et al., 2007) suggested a minimum sample size of N = 92 to detect a small-to-medium effect size (Cohen's f=.18) in a mixed factorial ANOVA (α = .05 and power = .80). The obtained sample size was therefore deemed acceptable to test study hypotheses. In a 4 (cognitive offloading; between subjects) × 2 (task difficulty; within subjects) mixed design, we investigated differences in decision accuracy, confidence, and confidence calibration.

Dependent Variables

Confidence calibration was calculated using two metrics. First, we computed an absolute confidence calibration statistic, *cc**, denoting deviation from perfect calibration (Weber & Brewer, 2004). It is calculated as the weighted mean of the squared difference between confidence, *c*, and accuracy, *p*, for each confidence level, *k,* summed across all trials*, n*. The statistic ranges from 0 (perfect calibration) to 1 (worst possible calibration).

$$cc^* = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)(c_k - p_k)^2$$

Second, we computed an over/underconfidence (O/U) calibration index, taking the mean difference between subjective confidence, *c*, and objective accuracy, *p*, averaged across trials after confidence was rescaled to reflect a 0-1 scale. This O/U statistic indexes the extent to which participants respond with more or less confidence than the accuracy of their decision warrants (Weber & Brewer, 2004). Scores of zero indicate perfect calibration, with positive and negative values reflecting overconfidence and underconfidence, respectively.

$$\overline{o/u} = \left(\frac{\bar{c} - 1}{6}\right) - \bar{p}$$

Stimuli

Sixteen (8 easy, 8 hard) medical diagnostic vignettes were developed for the purpose of this study. Each vignette presented information across six medical categories[1] and included two to four salient decision cues indicative of the correct medical condition. Difficulty was manipulated by increasing memory load, with hard vignettes containing ten numerical lab results compared to five in the easy vignettes, and by increasing judgement demands, with hard vignettes including red herrings to suggest multiple plausible medical diagnoses, unlike easy tasks. Pilot testing confirmed the effectiveness of the difficulty manipulation (N = 22).

Procedure and Materials

The experiment was conducted via Qualtrics. Participants were instructed about the task, provided consent, and were randomly assigned to one of four offloading conditions. An example vignette was shown to familiarise participants with the task, before they completed eight experimental trials (4 easy, 4 hard) presented in random order. Each of the six sections of the vignettes were displayed on a separate page and participants could not return to previous pages, creating a deliberate memory burden. The subsequent procedure varied based on the offloading condition.

In the no offloading condition (n = 28), participants read all the information and made a diagnosis without any support. In the memory-only offloading condition (n = 28), participants selected key pieces of information they deemed relevant while reading, which were presented again at the point of diagnosis to aid memory. In the judgement-only offloading condition (n = 23), participants read the vignettes, made an initial diagnosis, and recalled from memory key information. This was submitted to GPT-4o (temperature 0.2), which returned a percentage likelihood for the participant's diagnosis, along with two alternative diagnoses and their respective likelihoods. Participants then submitted a final diagnosis. Finally, in the combined memory and judgement offloading condition (n = 27), participants selected relevant information while reading, which was then passed to GPT-4o alongside their initial diagnosis. Again, GPT-4o then returned likelihood estimates for the initial and two alternative diagnoses, before participants submitted a final decision. If the participant's initial diagnosis was not a valid medical condition (e.g., "I don't know") the model returned a third alternative diagnosis, so that participants always saw three recommendations ranked by likelihood.

Final diagnoses were selected from 10 multiple-choice options, specific to that vignette. Participants rated their confidence in their decision on a 7-point scale (1=not at all confident, 7=extremely confident) and rated the difficulty of the task (0=extremely easy, 100=extremely difficult). Finally, participants provided demographics (age, gender) and were debriefed.

[1] Patient details, medical history, symptoms, physical examination, vitals, lab results.

Results

The manipulation of task difficulty was successful. The perceived difficulty of the tasks was higher in the hard condition (M = 63.3, SD = 19.8) than the easy condition (M = 42.6, SD = 22.9), $t(105)$ = 10.57, p < .001, d = 1.03. Performance and confidence were then compared between offloading groups using 2 × 4 mixed-model ANOVAs.

Decision accuracy was significantly higher for easy tasks compared to hard tasks ($F(1, 102)$ = 250.49, p < .001, ηp2 = .711), but comparable between offloading conditions ($F(3, 102)$ = 2.22, p = .09, ηp2 = .061). However, a significant interaction effect revealed that offloading improved performance under high task difficulty ($F(3, 102)$ = 3.86, p = .012, ηp2 = .102). Specifically, those in the combined memory and judgement offloading condition (M = .56, SD = .26) outperformed unaided decision makers (M = .32, SD = .21) on hard tasks (p = .009; Bonferroni corrected).

Analyses of confidence revealed a significant main effect of difficulty ($F(1, 102)$ = 156.67, p < .001, ηp2 = .606) and of offloading condition ($F(3, 102)$ = 3.44, p = .020, ηp2 = .092). Post hoc tests revealed this was driven by a greater confidence in the judgement offloading condition relative to the memory offloading condition (p = .019). A significant interaction effect was also observed ($F(3, 102)$ = 3.01, p = .034, ηp2 = .081). Under low task difficulty, confidence levels of those in the judgement offloading condition (M = 5.4, SD = 1.1) exceeded the memory offloading condition (M = 4.5, SD = 1.3) (p = .044). Under high task difficulty, confidence was significantly higher for judgement offloading (M = 4.4, SD = 1.1) compared to unaided decision makers (M = 3.3, SD = 1.1) (p = .018), and this difference approached significance compared to those who offloaded memory (M = 4.5, SD = 1.3) (p = .051) (Figure 1).

Analyses of confidence calibration revealed that the unaided group were the most accurately calibrated for both the O/U calibration index (M = -.05, SD = .19) and the absolute calibration statistic (M = .19, SD = .13) ($F(3, 102)$ = 3.02, p = .033, ηp^2 = .082). This control group also showed the only significant correlation between accuracy and confidence for easy tasks, $r(26)$ = .40, p = .037. No such relationship was observed for any of the offloading conditions, suggesting miscalibration.

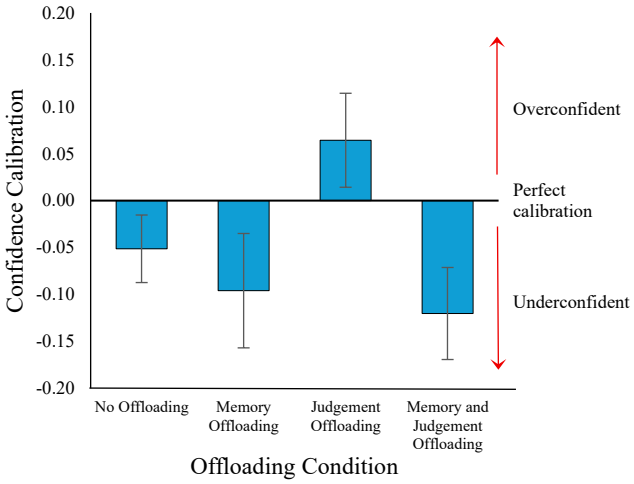

Figure 2: Confidence calibration across offloading conditions. Error bars represent standard error.

Interestingly, the direction of O/U miscalibration differed between offloading conditions. Participants in the judgement offloading condition exhibited overconfidence (M = .06, SD = .24), whereas all other groups reported underconfidence, with the combined offloading group showing the greatest underconfidence (M = -.12, SD = .25) (see Figure 2). Furthermore, participants tended to be overconfident for hard tasks (M = .05, SD = .35) and underconfident for easy tasks (M = -.16, SD = .25), reflecting the well-established hard-easy effect (Gigerenzer et al., 1991).

We hypothesised that the underconfidence observed in the combined offloading condition, but not seen in the judgement offloading group, may reflect reduced cognitive engagement arising from overreliance on AI support. Participants who offloaded both memory and judgement may have delegated so much of the task that, despite achieving higher final accuracy through AI feedback, they engaged in less self-generated reasoning. As a result, their improved performance

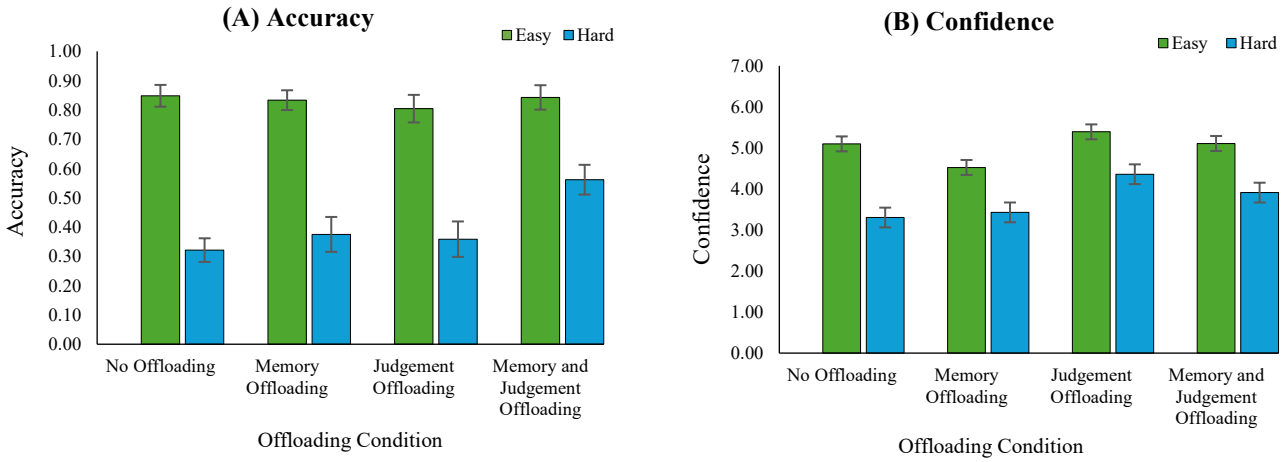

Figure 1: Accuracy (A) and confidence (B) across offloading conditions and difficulty. Error bars represent standard error.

was not accompanied by a corresponding increase in certainty, leading to underconfidence.

To test this assumption, we examined whether participants' initial (pre-AI feedback) accuracy differed between the AI-assisted conditions, as a more accurate unaided judgement would be consistent with a greater engagement in the task. Indeed, we found that initial accuracy was higher in the judgement offloading condition (M = .52, SD = .20) than in the combined offloading condition (M = .44, SD = .24), consistent with our prediction and suggesting that over-delegation may have both performance and metacognitive consequences. Whilst this difference did not reach significance ($t(48) = 1.26$, $p = .22$, $d = .36$), a post-hoc power analysis indicated our data lacked adequate power to detect a significant effect for this exploratory test ($1 - \beta = .23$).

Finally, AI reliance was assessed by comparing participants' initial diagnoses, the AI recommendations, and final decisions. Overall, final decisions matched the AI recommendation in 90.4% of cases. However, this often reflected agreement between participants and the AI prior to feedback. Across all trials, 47.3% involved alignment between the initial diagnosis, AI recommendation, and final decision that resulted in a correct outcome, while 28.6% involved aligned but incorrect outcomes. Such convergence occurred more frequently in the judgement offloading condition (89.0%) than in the combined offloading condition (63.7%).

Focusing on cases where the participant's initial judgement differed from the AI recommendation (n = 90) we then analysed reliance strategies. Figure 3 outlines the coding system for each AI reliance strategy (based on the 'Appropriateness of Reliance' measurement concept proposed by Schemmer et al. (2023)).

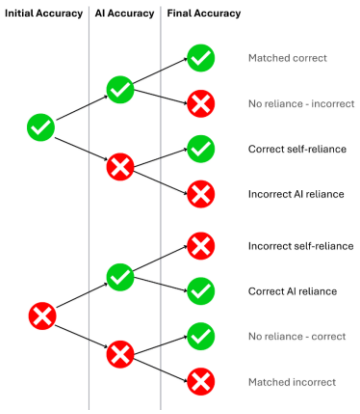

Figure 3: AI reliance coding system

Correct AI reliance was the most common strategy (60.0%), where participants correctly changed their decision to follow the AI's recommendation. This was followed by incorrect self-reliance (24.4%), where participants ignored AI's correct recommendation, instead submitting their own incorrect decision. 11.1% of participants began with an incorrect diagnosis, received an incorrect AI recommendation, yet still arrived at the correct final judgement ("no reliance, correct"). These cases often involved selecting the AI's second recommendation after it repeated the participant's incorrect initial diagnosis as the first. A small number of participants (3.3%) were initially correct, received the correct AI recommendation, but still submitted an incorrect final answer ("no reliance, incorrect"). Finally, there was only one case (1.1%) of correct self-reliance, and no instances of incorrect AI-reliance.

After excluding matched cases, the combined offloading condition showed more AI reliance (68.6%) than the judgement-only condition (30.0%), but AI reliance was comparable between easy (59.4%) and hard tasks (60.3%).

Discussion

Our results offer the first evidence that AI decision support impairs confidence calibration. Specifically, participants in the no offloading condition showed the only alignment between confidence and accuracy, while all offloading groups demonstrated miscalibration. Interestingly, the direction of miscalibration differed by AI condition: offloading of the evaluative process (judgement-only) produced overconfidence, whereas combined offloading of both cue storage and evaluation (memory and judgement) produced underconfidence. In terms of performance, combined offloading yielded the greatest accuracy gains under high task difficulty, reflecting the benefit of AI augmentation for complex decision making. Yet, these improvements did not translate into greater decision confidence. This dissociation between objective accuracy and subjective certainty underscores a critical metacognitive cost of AI support. Together, these findings reveal that offloading affects not only how people decide, but also how well they evaluate their decision making - a dynamic that carries significant implications for augmented human cognition.

One explanation for the underconfidence observed in the combined offloading condition lies in cognitive engagement. As presented earlier, extensive offloading may lead individuals to disengage in the decision-making process, attenuating their sense of ownership and understanding of the reasoning behind the decision. When both memory and judgement were delegated, participants likely relied heavily on the AI to retrieve, evaluate, and integrate diagnostic cues. Indeed, participants in this condition showed far more AI reliance (68.6%) than their peers in the judgement-only condition (30.0%). Anticipating AI support throughout the process may have reduced their attention to relevant information and subsequent internal evaluative processes (Kleinberg & Lau, 2021; Sparrow et al., 2011). Consequently, while participants in the combined offloading condition achieved higher accuracy through this higher rate of delegation, they may have lacked a sense of why their decision was correct. This absence of self-generated reasoning may have undermined their ability to verify the accuracy of the decision, and so, relative to their improved performance, they appear underconfident. Furthermore,

participants in this condition offloaded a larger proportion of the diagnostic cues to the model. While this maximised the AI accuracy, it may also have made it harder for users to reconstruct the model's reasoning. With limited transparency into how the AI reached its conclusions, participants may have accepted the output passively without genuine understanding, reinforcing uncertainty. This mechanism parallels research showing that automation bias can foster complacency, whereby users trust AI outputs without maintaining sufficient cognitive engagement (Gajos & Mamykina, 2022).

By contrast, participants in the judgement-only condition, who stored information internally and relied solely on the AI for weighting and integrating diagnostic cues which they themselves selected, were forced to engage more actively in the decision process. They had to encode and prioritise cues, hold them in working memory, and synthesise them into an initial hypothesis before seeking AI feedback. Given that the more effort required to process and react to an advisor's recommendations, the more overconfidence increases (Bonaccio & Dalal, 2006), this effortful engagement may explain why confidence is higher in the judgement-only group, and by contrast, why less effort results in lower confidence in the combined group. This effortful processing and cognitive engagement may explain why participants achieved greater independent accuracy for their initial decision when offloading the judgement process alone. However, as this analysis was exploratory and underpowered, the results should be interpreted as preliminary evidence rather than a definitive test of the disengagement account. Nevertheless, it raises a compelling direction for future research.

A second mechanism that may account for the contrasting patterns of over- and underconfidence concerns differences in how participants provided information to the AI. Because participants produced the AI's input themselves, the cues they selected shaped the quality and direction of the model's recommendations. For the judgement-only group, constrained by working memory limitations, participants would have retained only a small subset of cues - likely those consistent with their emerging hypothesis. Under this high cognitive load and subsequent selective cue sampling, participants could have biased the AI's recommendations toward confirming their initial judgement. Indeed, the AI agreed with the judgement-only group's initial diagnosis 90% of the time, suggesting that participants were inadvertently steering the model toward their own hypothesis. The resulting agreement between the AI and user likely reinforced participants' sense of certainty, amplifying confidence without improving accuracy, thus, resulting in overconfidence. In contrast, participants in the combined offloading condition were not constrained by working memory. They could supply the AI with more comprehensive, balanced cue sets, resulting in higher overall AI accuracy. However, this also meant the AI disagreed with participants' initial judgements more often. Such contradictions, while improving team performance, may have undermined participants' confidence in their decision-making abilities. This account is consistent with human-human advice-taking research showing that advisor agreement inflates confidence even when advice is incorrect, and similarly, that advisor disagreement reduces confidence even when the advice is correct (Budescu et al., 2003; Soll et al., 2022). Thus, our findings suggest that it may be AI agreement, rather than actual team performance, that drives metacognitive evaluations.

Taken together, our findings suggest two potential routes by which offloading distorts metacognitive calibration. First, combined offloading may produce underconfidence through cognitive disengagement. By delegating both storage and reasoning to AI, users are relatively uninvolved in the decision process, losing insight into how conclusions are reached, enhancing accuracy without increasing confidence accordingly. Second, judgement-only offloading may have fostered overconfidence by reinforcing participants' beliefs. Under higher cognitive load, users may have fed selective, confirmatory cues to the AI, and received validating feedback, which in turn inflated confidence without equivalent gains in accuracy. Both patterns reflect poor metacognitive calibration - one group accurate but not confident, the other confident but not accurate. These results underscore the complexity of human-AI collaboration, with performance gains potentially coming at the cost of distorted metacognition.

Limitations

While the present study provides novel insights, there are limitations to consider. Although we propose cognitive disengagement and confirmation bias as explanatory mechanisms, these are speculatory rather than directly observed. Future work should look to empirically verify these accounts, employing process-tracing methods such as eye-tracking or concurrent verbal protocols to objectively assess cue selection strategies and cognitive engagement throughout the task.

Conclusions

This study provides the first empirical evidence that cognitive offloading decision-making processes to AI disrupts metacognitive calibration, with the direction of distortion depending on the type and extent of cognitive delegation. Offloading judgement alone inflated confidence, while combined offloading of memory and judgement diminished it. Both patterns reveal that AI support, while enhancing performance, can impair users' ability to accurately assess their own competence. We propose two mechanisms - biased cue input through cognitive overload, and disengagement through automation bias - which may underlie these effects, warranting further investigation and verification. Findings carry important implications for designing human-AI systems. Interventions to foster appropriate reliance must account for the dynamics of the distributed cognitive system, targeting overconfidence under certain conditions and underconfidence for others.

References

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

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. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Budescu, D. V., Rantilla, A. K., Yu, H.-T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178–194. https://doi.org/10.1016/S0749-5978(02)00516-2

Buettner, R. (2013). Cognitive workload of humans using artificial intelligence systems: Towards objective measurement applying eye-tracking technology. In I. J. Timm & M. Thimm (Eds.), *KI 2013: Advances in Artificial Intelligence* (Vol. 8077, pp. 37–48). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40942-4_4

Burnett, L. K., & Richmond, L. L. (2025). Meta-analytic investigations of the effect of cognitive offloading on memory-based task performance and interindividual variability. *Memory & Cognition*. https://doi.org/10.3758/s13421-025-01743-8

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*(3), 674–687. https://doi.org/10.1037/xge0000070

Fisher, M., & Oppenheimer, D. M. (2021). Who knows what? Knowledge misattribution in the division of cognitive labor. *Journal of Experimental Psychology: Applied*, *27*(2), 292–306. https://doi.org/10.1037/xap0000310

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, *33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Gajos, K. Z., & Mamykina, L. (2022). Do people engage cognitively with AI? Impact of AI Assistance on incidental learning. *27th International Conference on Intelligent User Interfaces*, 794–806. https://doi.org/10.1145/3490099.3511138

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), 6. https://doi.org/10.3390/soc15010006

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Gilbert, S. J., Bird, A., Carpenter, J. M., Fleming, S. M., Sachdeva, C., & Tsai, P.-C. (2020). Optimal use of reminders: Metacognition, effort, and cognitive offloading. *Journal of Experimental Psychology: General*, *149*(3), 501–517. https://doi.org/10.1037/xge0000652

Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2023). Mutual interplay between cognitive offloading and secondary task performance. *Psychonomic Bulletin & Review*, *30*(6), 2250–2261. https://doi.org/10.3758/s13423-023-02312-3

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3544548.3581025

Ishizu, N., Yeoh, W. L., Okumura, H., & Fukuda, O. (2024). The effect of communicating AI confidence on human decision making when performing a binary decision task. *Applied Sciences*, *14*(16), 7192. https://doi.org/10.3390/app14167192

Kleinberg, M. S., & Lau, R. R. (2021). Googling politics: How offloading affects voting and political knowledge. *Political Psychology*, *42*(1), 93–110. https://doi.org/10.1111/pops.12689

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.08872

Küper, A., Lodde, G. C., Livingstone, E., Schadendorf, D., & Krämer, N. (2025). Psychological factors influencing appropriate reliance on AI-enabled clinical decision support systems: Experimental web-based study among dermatologists. *Journal of Medical Internet Research*, *27*, e58660. https://doi.org/10.2196/58660

Lee, D., Pruitt, J., Zhou, T., Du, J., & Odegaard, B. (2025). Metacognitive sensitivity: The key to calibrating trust and

optimal decision making with AI. *PNAS Nexus*, *4*(5). https://doi.org/10.1093/pnasnexus/pgaf133

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science*, *30*(6), 1081–1095. https://doi.org/10.1207/s15516709cog0000_71

Li, J., Yang, Y., Liao, Q. V., Zhang, J., & Lee, Y.-C. (2025). As confidence aligns: Understanding the effect of AI confidence on human self-confidence in human-AI decision making. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3706598.3713336

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. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3613904.3642671

Payne, J. W., Johnson, E. J., & Bettman, J. R. (1993). *The adaptive decision maker*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173933

Peper, P., Alakbarova, D., & Ball, B. H. (2023). Benefits from prospective memory offloading depend on memory load and reminder type. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(4), 590–606. https://doi.org/10.1037/xlm0001191

Puncochar, J. M., & Fox, P. W. (2004). Confidence in individual and group decision making: When 'two heads' are worse than one. *Journal of Educational Psychology*, *96*(3), 582–591. https://doi.org/10.1037/0022-0663.96.3.582

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626. https://doi.org/10.1145/3514094.3534128

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422. https://doi.org/10.1145/3581641.3584066

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299–314. https://doi.org/10.1037/0278-7393.30.2.299

Soll, J. B., Palley, A. B., & Rader, C. A. (2022). The bad thing about good advice: Understanding when and how advice exacerbates overconfidence. *Management Science*, *68*(4), 2949–2969. https://doi.org/10.1287/mnsc.2021.3987

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Srivastava, D., Lilly, J. M., & Feigh, K. M. (2025). Exploring the role of judgement and shared situation awareness when working with AI recommender systems. *Cognition, Technology & Work*, *27*(1–2), 1–18. https://doi.org/10.1007/s10111-024-00771-9

Storm, B. C., & Stone, S. M. (2015). Saving-enhanced memory: The benefits of saving on the learning and remembering of new information. *Psychological Science*, *26*(2), 182–188. https://doi.org/10.1177/0956797614559285

Taudien, A., Fügener, A., Gupta, A., & Ketter, W. (2022). *The effect of AI advice on human confidence in decision-making*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2022.029

Tidwell, J. W., Buttaccio, D., Chrabaszcz, J. S., Dougherty, M. R., & Thomas, R. P. (2016). Sources of bias in judgment and decision making. *The Oxford Handbook of Metamemory.*, 109–125.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

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. https://doi.org/10.1145/3579605

Ward, A. F. (2021). People mistake the internet's knowledge for their own. *Proceedings of the National Academy of Sciences*, *118*(43). https://doi.org/10.1073/pnas.2105061118

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. https://doi.org/10.1037/1076-898X.10.3.156