Too Sure for Our Own Good: A User Study on AI Confidence and Human Reliance

Caterina Fregosi[1], Lucia Vicente[1], Andrea Campagner[1,2], Federico Cabitza[1,2]

[1]University of Milano-Bicocca, Milan, Italy
[2]IRCCS Ospedale Galeazzi–Sant'Ambrogio, Milan, Italy
caterina.fregosi@unimib.it, lucia.vicenteholgado@unimib.it, andrea.campagner@unimib.it, federico.cabitza@unimib.it

Abstract

Achieving appropriate human reliance on Artificial Intelligence (AI) systems remains a central challenge in Human-Computer Interaction. Confidence scores–indicators of an AI system's certainty in its recommendations–have been proposed as a means to help users calibrate their trust and reliance on AI Decision Support Systems (DSS). However, limited research has explored how well-calibrated versus miscalibrated confidence scores affect human decision-making. We report a study examining the effects of confidence calibration on user reliance, decision accuracy, and perceived utility of an AI DSS. In a within-subjects experiment involving 184 participants solving logic puzzles, we found that well-calibrated confidence scores significantly improved decision accuracy (+20%, 95% CI: [0.18, 0.23]), whereas miscalibrated scores yielded minimal accuracy gains (+2%, 95% CI: [-0.00, 0.04]) and increased vulnerability to automation bias and conservatism bias. Participants were more likely to accept AI recommendations when high confidence was expressed, even when those recommendations were incorrect, resulting in errors. Conversely, miscalibrated and low-confidence recommendations increased conservatism bias, leading users to reject even accurate AI suggestions. Perceived utility of the AI system was higher when confidence levels were high ($p < 0.001$) and when confidence was well-calibrated ($p = 0.002$). These findings underscore the importance of designing AI systems with properly calibrated confidence cues to improve human-AI collaboration and mitigate reliance-related biases.

Datasets — https://github.com/cfregosi/Confidence_score

Introduction

Achieving appropriate human reliance on artificial intelligence (AI) systems has emerged as a central challenge in human-computer interaction (HCI) and human-AI collaboration. In high-stakes decision-making, users must discern when to trust AI-generated recommendations and when to rely on their own judgment (Schemmer et al. 2023). Effective human-AI collaboration requires users to dynamically calibrate their reliance based on the AI system's actual reliability (Huang and Bashir 2017). Indeed, inappropriate reliance can have serious consequences: over-reliance on AI, a phenomenon known as automation bias (Goddard, Roudsari, and Wyatt 2012, 2014), can lead humans to mistakes that could be avoided with careful supervision. Conversely, excessive distrust of the machines, an effect known as algorithm aversion (Dietvorst, Simmons, and Massey 2015), can be an obstacle to leveraging the multiple benefits of this technology. Although the relevance of promoting appropriate human reliance on AI systems is widely recognized, empirical research addressing this need remains limited (Schemmer et al. 2023) and there is still insufficient knowledge about the most effective approaches to help users calibrate their trust in AI systems.

Explainable AI (XAI) techniques have been proposed to address this challenge by offering explanations that make AI systems more transparent, helping users make informed decisions (Schemmer et al. 2023; Schoeffer, De-Arteaga, and Kuehl 2024; Vasconcelos et al. 2023). However, explanations themselves are not without risk. While increased transparency in AI systems is intended to enhance user understanding, it can sometimes produce unintended effects. For example, the "white-box" paradox occurs when explanations mislead users into believing that incorrect AI advice is actually reliable, fostering inappropriate reliance on the system (Bansal et al. 2021; Cabitza et al. 2023b). Similarly, the "halo effect" arises when misleading explanations–those lacking coherence or contextual relevance–cause users to become wary of AI advice and ultimately disregard it, even when the advice is correct, thereby impairing decision accuracy (Cabitza et al. 2024). These findings underscore the importance of designing XAI strategies that not only foster understanding but also mitigate the risks of inappropriate reliance (Famiglini, Campagner, and Cabitza 2023).

One promising approach, combining XAI with research on uncertainty quantification (UQ) (Marx et al. 2023; Seuß 2021), is the use of confidence scores, which indicate the AI's confidence in specific predictions. Confidence scores offer an interpretable method for users to evaluate AI performance dynamically, enabling more precise adjustments in reliance (Banerji et al. 2023; Zhang, Liao, and Bellamy 2020). Compared to general accuracy metrics, task-specific confidence indicators help users assess the reliability of individual recommendations, making them particularly effective in mitigating automation bias. Confidence scores are also cognitively less demanding and suitable for non-expert

users (Okamura and Yamada 2020; Zhang, Liao, and Bellamy 2020).

However, very few studies have examined how confidence scores influence users' acceptance of correct and incorrect AI advice (Bussone, Stumpf, and O'Sullivan 2015; Zhang, Liao, and Bellamy 2020; Jiang, Kahai, and Yang 2022; Cao, Liu, and Huang 2024; Li and Steyvers 2025). For example, one study found that confidence scores can help participants better evaluate whether to rely on AI predictions (Zhang, Liao, and Bellamy 2020). Participants were more likely to trust AI recommendations when the model expressed high confidence (scores above 80%) and were more skeptical when confidence was low (scores below 60%). These confidence indicators helped users identify scenarios in which following AI advice carried a high risk of error. Other work has similarly shown that the system's expressed confidence–whether conveyed through probability estimates, confidence intervals, or verbal statements–affects how human decision-makers adopt, question, or disregard automated recommendations (Bussone, Stumpf, and O'Sullivan 2015; Kocielnik, Amershi, and Bennett 2019). Specifically, research has examined how different forms of confidence disclosure (e.g., numeric vs. textual) impact trust calibration and the likelihood that clinicians will rely on a system's recommendations (Bussone, Stumpf, and O'Sullivan 2015), as well as how AI confidence, along with other system cues, shapes user decision-making regarding the acceptance or rejection of AI-generated suggestions (Kocielnik, Amershi, and Bennett 2019). More recently, a study focusing on medical image diagnosis found that how model uncertainty is presented can significantly influence user reliance, particularly in high-stakes contexts (Cao, Liu, and Huang 2024).

While the above mentioned studies provide important cues, their conclusions typically assume that the AI's expressed confidence is well-calibrated (Silva Filho et al. 2023). In reality, confidence scores do not necessarily align with actual performance (Cao, Liu, and Huang 2024). When confidence scores are miscalibrated, failing to accurately reflect the true probabilities of correct predictions, they can mislead users, undermining decision-making processes. Miscalibration can lead users to overtrust high-confidence predictions that are incorrect or disregard low-confidence predictions that are accurate (Price and Stone 2004; Pulford et al. 2018). This risk is exacerbated by human tendencies to treat confidence as a proxy for accuracy, especially when direct accuracy information is unavailable or ambiguous (Sah, Moore, and MacCoun 2013). Consequently, poorly calibrated confidence scores may unintentionally perpetuate the very reliance issues they are aimed to address. By contrast, recent work has suggested that miscalibrated yet strategically amplified confidence cues may in some cases enhance performance (Vodrahalli, Gerstenberg, and Zou 2022), although such approaches raise concerns about transparency and potential over-reliance, especially in critical settings where human accountability remains central.

This highlights a critical and underexplored question at the intersection of human cognition and AI design: how do expressions of confidence, calibrated or miscalibrated, affect not just trust, but the quality of human decision-making?

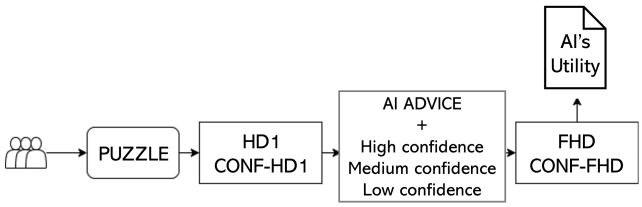

Figure 1: Participants first solved a series of puzzles independently, recording their initial decision (HD1) along with their confidence level (CONF-HD1). They then received AI-generated advice, which was accompanied by confidence ratings labeled as High, Medium, or Low. After reviewing the AI's suggestion, participants could either revise their answers (FHD) and update their confidence levels (CONF-FHD) or confirm both as they were. Additionally, for each puzzle, participants evaluated the perceived usefulness of the AI's advice.

To our knowledge, no work in HCI has investigated the risks associated with miscalibrated scores, that is, to what in cognitive psychology is called poor metacognitive sensitivity (Li and Steyvers 2025).

To address this gap, we conducted a within-subjects user study based on a logic question-answering scenario to investigate how different combinations of AI advice correctness, expressed confidence, and confidence calibration affect users' reliance behavior, decision accuracy, and perceived utility (see Figure 1). We adopted a human-first (Cabitza et al. 2023b) (or sequential (Tejeda et al. 2022)) study design, in which participants first answered a set of logic puzzles before receiving AI-generated advice accompanied by confidence scores labeled as high, medium, or low. They then had the opportunity to revise their initial answers based on this input.

This approach allowed us to examine how confidence calibration affects users' decision-making, specifically investigating whether different levels of calibration (namely good calibration and poor calibration) and different levels of expressed AI confidence (high, medium, low) do affect decision accuracy, appropriate reliance, as well as the related cognitive biases (namely automation and conservatism bias), and perceived utility.

Methods

Participants A total of 184 participants (mean age = 29.46, SD = 14.37) took part in our study. Of these, 38.6% self-identified as female and 58.7% as male (2.7% other/preferred not to answer). No specific inclusion or exclusion criteria were established a priori, as we aimed for a sample representative of the general population. The only requirement for participation was being at least 18 years old.

Materials In our study, participants were required to solve a set of 20 logic puzzles of moderate to high difficulty, supported by a simulated AI system.

The selected puzzles, taken from Eysenck's intelligence test (Eysenck 1976) and Youmath[1] , were cognitively challenging and required a variety of skills, including mathematical, verbal and visual-spatial abilities. The logic puzzles were selected based on complexity criteria validated in prior studies to ensure comparable difficulty across conditions. These logic questions were not used to assess participants' capabilities, but rather to present them with challenging problems to solve in collaboration with an AI system. Our main aim was to examine the participants' reliance on the AI-generated advice.

Data collection for the study was supported by the LimeSurvey platform[2]. Participants accessed the study via a web link using their own devices.

Procedure In line with human-first protocols, participants were first presented with each one of the 20 puzzles and they were required to make an initial unaided decision, without any support, by selecting one of five possible answers. Subsequently, the same logic question was presented again, this time accompanied by the AI-generated advice and its corresponding confidence level. Participants then had the opportunity to either revise their initial decision based on the AI's suggestion or confirm their original choice (see Figure 1).

The simulated AI supporting participants in our study was designed according to the Wizard of Oz approach (Dahlbäck, Jönsson, and Ahrenberg 1993), and scripted to suggest the correct answer in 10 out of 20 cases, thus, exhibiting an overall accuracy of 50%, which was higher than the average accuracy of human participants (approximately 26%) as observed in a prior pilot study that we conducted to test our experimental paradigm. We decided to adopt a Wizard of Oz study design in order to have more control on the interaction between the human participants and the AI support.

Similarly, the AI's confidence levels were pre-determined to be either calibrated or miscalibrated, so as to ensure controlled accuracy and confidence calibration. Different approaches to measure calibration have been proposed in the literature (Silva Filho et al. 2023). In our experiment, we considered the notion of confidence-calibration (Mortier et al. 2023), commonly employed in the definition of the Expected Calibration Error (ECE) used to evaluate the calibration of machine learning models (Naeini, Cooper, and Hauskrecht 2015; Nixon et al. 2019), which requires the AI support confidence scores to be aligned with the correctness of the corresponding predictions. Thus, a calibrated confidence level meant that either AI provided a correct response with high confidence or an incorrect response with low confidence: these configurations occurred in 8 out of 20 cases in our study. Conversely, the AI was miscalibrated when it provided an incorrect response with high confidence or a correct response with low confidence, in 7 out of 20 cases. In the remaining five cases, the AI showed a level of intermediate confidence in its answers, in order to make its behavior less extreme and more believable. Thus, each participant encountered both well-calibrated and miscalibrated confidence levels throughout the problem-solving task.

In our study, AI confidence was communicated using categorical labels - high, medium, or low - rather than numerical percentages. We chose this approach because categorical labels are generally easier to interpret. There is no consensus on the exact percentage ranges that define high, medium, or low confidence; for instance, a 75% confidence level has been interpreted as high confidence in some studies (Bussone, Stumpf, and O'Sullivan 2015) but as medium confidence in others (Zhang, Liao, and Bellamy 2020). These findings suggest that presenting confidence as a numerical percentage may lead to inconsistent interpretations among participants, as individuals may perceive the same value differently (Fischhoff and Bruine De Bruin 1999). By contrast, in our study we only distinguished between polarized expressions of confidence (i.e., high/highly confident and low/slightly confident) and intermediate ones (medium/-moderately confident)[3].

Along with participants' initial decision (HD1) and final decision (FHD) for each logic puzzle, we also recorded their self-perceived confidence in their responses, both before and after viewing the AI's advice. Thus, in addition to selecting an initial response for each puzzle, participants rated their confidence on a 4-point ordinal scale (ranging from 1 = "Not at all confident" to 4 = "Totally confident"). After viewing the AI's suggested response and its associated confidence level (high, medium, or low), participants provided their final response and had the opportunity to update their confidence rating using the same 4-point scale. In addition to their final decision and confidence, participants also rated how useful they found the AI support in solving each case on a 4-point ordinal scale, ranging from "Not at all useful" to "Extremely useful".

All participants completed the 20 logic questions in the same predetermined order. Incorrect and correct AI advice, whether it was given with calibrated or miscalibrated confidence levels, was presented to all participants in a fixed order sequence. The order of each type of AI output was determined randomly, except for the first three cases out of the 20. In these initial instances, we avoided presenting any incorrect or miscalibrated AI responses to prevent excessive distrust in the system. Results are produced by means of the Human–AI Interaction Assessment tool available online[4]. Statistical analyses of the results were conducted using a 95% confidence level and a significance level of 5%. Accuracy rates were compared using a Two-Proportion Z Test, given the independence of responses (ICC for HD1: 0.04; ICC for FHD: 0.01). Perceived utility levels were tested using a Mann-Whitney U Test. Automation bias and conservatism bias were computed using the metrics proposed in (Cabitza et al. 2023a). Automation bias quantifies users' tendency to follow incorrect AI advice rather than rejecting it. It is computed as an odds ratio comparing the detrimental over-reliance pattern (accepting wrong AI advice) with the

[1]https://www.youmath.it/

[2]https://www.limesurvey.org/

[3]The exact wordings were Italian phrases equivalent to, respectively "the system is highly confident", "the system is moderately confident", "the system is slightly confident".

[4]https://www.entechne.com/metimeter/haiassessment

beneficial self-reliance pattern (rejecting wrong AI advice):

$$AB = \frac{\text{Odds}(FHD{=}0,\ AI{=}0,\ HD1{=}1)}{\text{Odds}(FHD{=}1,\ AI{=}0,\ HD1{=}1)} \quad (1)$$

By contrast, conservatism bias quantifies users' tendency to reject correct AI advice, reflecting detrimental self-reliance. It is computed as an odds ratio comparing the detrimental self-reliance pattern (rejecting correct AI advice) with the beneficial over-reliance pattern (accepting correct AI advice):

$$CB = \frac{\text{Odds}(FHD{=}0,\ AI{=}1,\ HD1{=}0)}{\text{Odds}(FHD{=}1,\ AI{=}1,\ HD1{=}0)} \quad (2)$$

Finally, appropriate reliance was computed as defined in (Cabitza et al. 2025):

$$\begin{aligned} AR = {} & P(\text{FHD correct, AI wrong}) \\ & + P(\text{FHD correct, AI correct, HD1 wrong}) \\ & + P(\text{FHD, AI, HD1 all correct, confidence} \uparrow) \\ & + P(\text{FHD, AI, HD1 all wrong, confidence} \downarrow) \end{aligned} \quad (3)$$

which quantifies the rate of positive reliance patterns[5], that is the proportion of cases in which humans trust the machine and follow its advice when it is correct, and distrust the machine and disregard its advice when this is incorrect.

Results

Results about the participants' accuracy are reported in Figures 2 and 3, while the appropriate reliance, stratified by confidence and calibration levels, as well as its components automation bias and conservatism bias, are reported in Table 1 and Figures 4; perceived utility is also reported in Table 1.

Discussion

We conducted an empirical study to examine the impact of different confidence and calibration levels on users' reliance on AI systems. Using a within-subjects design, our primary dependent variable was the change in accuracy between participants' initial (pre-AI) and final (post-AI) decisions.

We then evaluated participants' appropriate reliance on the AI, linking it to well-known cognitive biases such as automation bias and conservatism bias. Finally, we assessed the perceived utility of the AI after participants received its advice, treating this as a secondary dependent variable to provide further insight into user perceptions and attitudes toward AI systems that disclose their confidence levels.

Accuracy

In Figure 2 we observe that the mean accuracy improvement, measured as the average difference FHD - HD1, is significantly higher in the calibrated condition (0.20, 95% CI: [0.18, 0.23]) compared to the miscalibrated condition (0.02, 95% CI: [-0.00, 0.04]). This indicates that properly calibrated confidence levels lead to a more substantial improvement in user performance. Figure 3 provides a clear illustration of how the expression and calibration of AI confidence influence decision accuracy. Notably, right-calibrated advice–where the system provides correct recommendations with high confidence–yields the highest accuracy, suggesting that users effectively integrate AI suggestions when the confidence level accurately reflects the correctness of the advice. In contrast, right-miscalibrated advice, where correct recommendations are expressed with low confidence, results in reduced accuracy, indicating that underconfident AI may undermine trust in its own guidance. A particularly striking observation is the detrimental impact of wrong-miscalibrated advice: when incorrect recommendations are expressed with high confidence, accuracy drops significantly, likely because participants are more inclined to accept and act upon confidently stated misinformation.

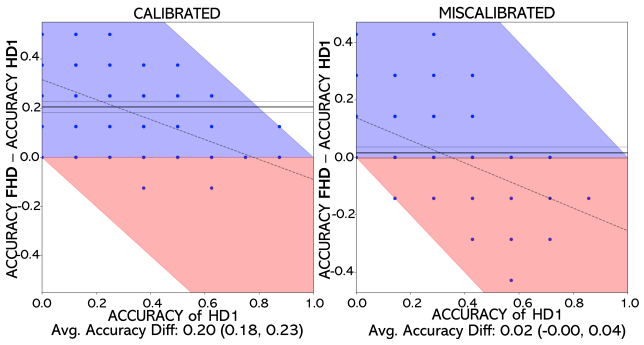

Figure 2: Benefit diagrams comparing well-calibrated (left) and miscalibrated (right) confidence scores. Each blue dot represents an individual user (n=184). The blue regions indicate areas of improvement, while the red regions denote areas of accuracy decline. The mean accuracy improvement is significantly greater in the well-calibrated condition (+0.20) compared to the miscalibrated condition (+0.02).

Interestingly, an intermediate expression of confidence, still leads to an improvement in accuracy, nearly reaching a saturation point. This suggests that even an ambiguous confidence signal provides valuable guidance, helping users make better decisions compared to when confidence is entirely misaligned. However, the most notable performance gains occur only when confidence is both polarized and calibrated. This implies that polarization alone–where the system expresses extreme confidence levels–is not necessarily beneficial and can even be harmful, in terms of accuracy, if miscalibrated. Instead, the key factor driving optimal performance is calibration, that is the proper alignment between the AI's confidence and the actual correctness of its advice.

Appropriate Reliance

Our findings reveal significant differences in appropriate reliance depending on both the calibration of AI recommendations and the confidence level displayed by the system. Specifically, participants demonstrated a higher appropriate

[5]In this metric the patterns in which both humans and AI agree are weighted in terms of the extent human confidence decreases, when both are wrong, and human confidence increases, when both are correct.

	Calibrated	Miscalibrated	High Confidence	Intermediate Confidence	Low Confidence
Appropriate Reliance	0.47	0.25	0.49	0.43	0.27
Automation Bias	0.14 [0.09, 0.21]	0.46 [0.34, 0.62]	0.46 [0.34, 0.62]	0.07 [0.05, 0.12]	0.14 [0.09, 0.21]
Conservatism Bias	0.54 [0.45, 0.66]	6.29 [4.98 7.95]	0.53 [0.44, 0.65]	1.44 [1.11,1.86]	5.96 [4.73, 7.51]
Avg. Perceived Utility	2.21	2.09	2.48	2.24	1.87

Table 1: Results of the statistical analyses.

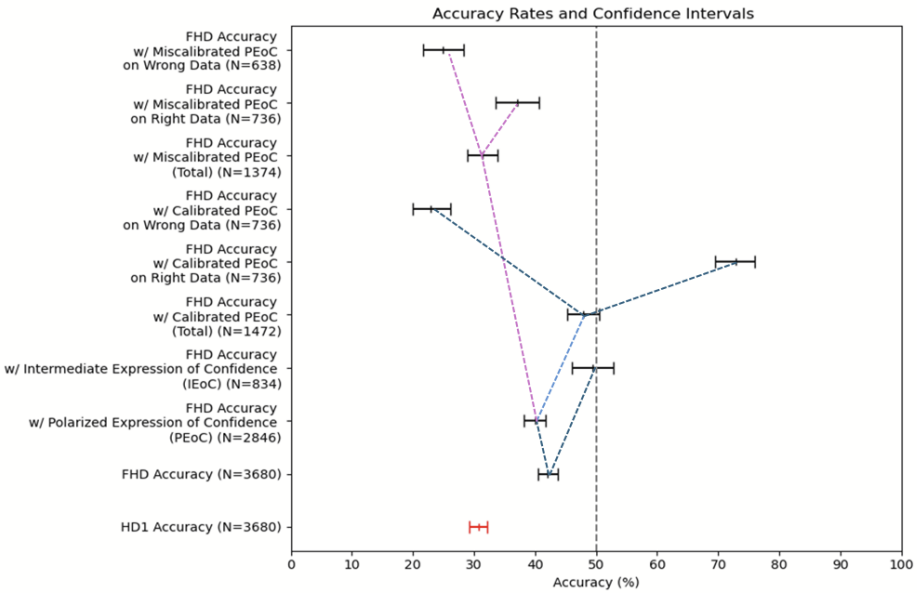

Figure 3: A visualization of the participants' accuracy, stratified by AI predictions' confidence level (polarized, meaning either high or low, or intermediate), calibration (calibrated or miscalibrated) and support correctness (wrong or right). For each estimate of accuracy, we also report the corresponding 95% confidence interval. The baseline AI support accuracy level (50%) is depicted as a vertical dashed line.

reliance on AI-generated recommendations when these were calibrated (0.47) compared to when they were miscalibrated (0.25), with a statistically significant difference ($Z = 11.96$, $p < .001$). Similarly, appropriate reliance was greater when the AI exhibited high confidence (0.49) rather than low (0.27) or intermediate (0.43) confidence , again with a significant difference (high vs low: $Z = 11.93$, $p < .001$; high vs intermediate: $Z = 2.79$, $p = 0.005$). Studies suggest that explicit and clear communication can facilitate the development and application of Theory of Mind by providing clear cues about an individual's mental states (Sodian, Kristen-Antonow, and Kloo 2020), thereby reducing ambiguity and aiding in more accurate mental state inferences. This could partly explain why appropriate reliance is higher when AI attaches high confidence in its bits of advice: while high confidence could persuade (Kruger and Dunning 1999) users also in case the bits are wrong, the users actually can discern these cases from those where AI is right in virtue of a better Theory of Mind (Tenney et al. 2007), fostered by clearer cues, as also shown by the fact that appropriate reliance was higher for calibrated than for miscalibrated advice. As expected, appropriate reliance in cases where the AI expressed an intermediate confidence level (0.43) fell between the values observed for low and high confidence.

The concept of appropriate reliance quantifies the proportion of cases in which humans effectively integrate AI recommendations into their decision-making, trusting and following correct advice while rejecting incorrect suggestions. A lack of this beneficial integration can be interpreted through the lens of two well-documented cognitive biases: automation bias and conservatism bias. Our results indicate that miscalibrated AI confidence amplifies both biases, though in distinct ways.

Automation bias When AI recommendations were miscalibrated, automation bias was significantly higher, leading users to uncritically accept AI suggestions, even when they were incorrect (see Figure 4). Similarly, high levels of automation bias were observed also when the AI expressed high confidence in its recommendations, suggesting that strong confidence signals can override users' natural skepticism, increasing the likelihood of accepting incorrect advice. In contrast, low-, and especially intermediate-confidence recommendations mitigated automation bias, as users might remain more cautious and engage in more critical evaluation before incorporating AI recommendation into

their decisions.

Conservatism bias At the same time, miscalibrated confidence also exacerbated conservatism bias, suggesting that users were less likely to adjust their initial decision in response to the AI-provided recommendation (see Figure 4). This effect indicates that inconsistencies between confidence expression and actual recommendation accuracy may lead users to dismiss AI suggestions, potentially because they recognize subtle misalignments in the system's reliability, even without explicit knowledge of its calibration. Conversely, when AI confidence was well-calibrated, conservatism bias was substantially reduced, demonstrating that users were more willing to incorporate AI support into their decision-making process when confidence levels accurately reflected correctness. Notably, while participants were not explicitly informed about the calibration of the AI system, their response indicates that they might have recognized discrepancies between the recommendations and confidence levels, thereby reinforcing their initial positions.

Furthermore, when the system provided recommendations with low confidence, users exhibited the highest levels of conservatism bias, suggesting that expressed uncertainty led them to devalue AI insights and rely more heavily on their initial decision. A similar results, although less extreme, was also observed for recommendations associated with intermediate confidence. In contrast, high-confidence recommendations corresponded to the lowest levels of conservatism bias, indicating that strong confidence signals can encourage users to reassess their initial decision.

Taken together, these findings highlight the dual importance of confidence expression and calibration: while high confidence may help mitigate conservatism bias, its benefits are maximized only when it is also well-calibrated, which allows to simultaneously mitigate both conservatism and automation bias.

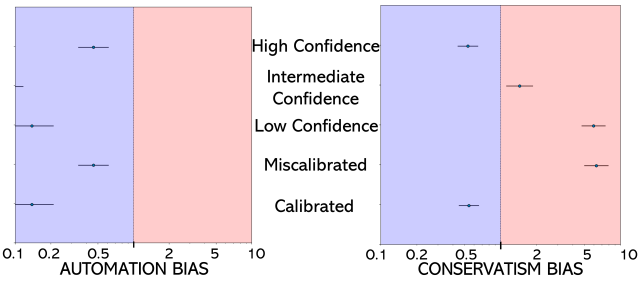

Figure 4: Comparison of automation bias (left) and conservatism bias (right) under different confidence conditions, with the corresponding 95% confidence intervals. Automation bias is markedly higher when AI recommendations are miscalibrated (0.46 [0.34, 0.62]) compared to calibrated ones (0.14 [0.09, 0.21]). High-confidence recommendations produce the strongest bias (0.46 [0.34, 0.62]), while intermediate and low confidence lead to much lower effects (0.07 [0.05, 0.12]; 0.14 [0.09, 0.21]). Conservatism bias also increases with miscalibrated recommendations (6.29 [4.98, 7.95] vs. 0.54 [0.45, 0.66]). It is highest for low-confidence inputs (5.96 [4.73, 7.51]), decreases with intermediate confidence (1.44 [1.11, 1.86]), and is lowest when recommendations are presented with high confidence (0.53 [0.44, 0.65]).

Perceived Utility

The analysis of perceived utility, as rated by participants, was conducted by stratifying cases based on whether the AI's recommendation was presented with calibrated or miscalibrated confidence, and with high or low confidence levels. A Mann-Whitney U test was performed to compare the perceived utility of advice given in calibrated versus miscalibrated confidence conditions. The results indicate a statistically significant difference between these two groups ($p = 0.002$), with utility ratings higher for calibrated cases.

This finding is particularly interesting because participants were not explicitly informed whether the AI's confidence was aligned with the correctness of its answers. Yet, the fact that they rated calibrated responses as more useful suggests they were, at some level, able to discern the alignment between confidence and accuracy. One possible explanation is that miscalibrated confidence introduces subtle inconsistencies in the AI's recommendations that users intuitively detect, even if they cannot explicitly articulate them. When the AI expresses unwarranted confidence in incorrect answers or hesitancy in correct ones, users may experience cognitive dissonance, leading to lower perceived utility. In contrast, a well-calibrated system offers a more coherent and predictable interaction experience, reinforcing users' trust in the AI's advice. However, the small effect size observed suggests that while calibration plays a role in shaping perceived utility, confidence level has a more substantial impact. This implies that, in practical applications, users may be more influenced by the apparent certainty of the AI rather than by its actual reliability. When comparing cases where the recommendation was given with high versus low confidence, the Mann-Whitney U test revealed a highly significant difference ($p < 0.001$) with a small effect size (=0.3). Utility ratings were higher when the AI expressed high confidence than when it expressed low confidence, indicating that users tend to value recommendations more when the system expresses strong confidence, even without explicit knowledge of its actual correctness.

The ethical implications of miscalibrated AI confidence are particularly critical in high-stakes decision settings, where user accountability is paramount. Our findings demonstrate that users are significantly influenced by extreme confidence representations–over-relying on highly confident but incorrect AI recommendations (automation bias) while disregarding low-confidence but correct advice (conservatism bias). In domains such as healthcare, miscalibrated confidence could shift responsibility away from human decision-makers, leading them to follow AI suggestions uncritically and without sufficient deliberation. Conversely, when AI confidence is systematically underestimated, professionals may hesitate to act on correct AI insights, potentially resulting in missed opportunities or harmful delays. This underscores the need for AI systems that not only calibrate their confidence properly, but also communicate their

level of calibration clearly and in a way that fosters human appropriate reliance.

This perspective stands in partial contrast to recent work suggesting that confidence distortion (deliberately altering model confidence) can, under certain conditions, improve joint human-AI performance. For instance, Vodrahalli, Gerstenberg, and Zou (2022) showed that making an AI system appear more confident than it actually is can sometimes enhance users' accuracy and confidence, owing to well-documented limitations in how people interpret probabilistic cues. However, our results highlight a potential risk in this strategy: in high-stakes or cognitively demanding contexts, miscalibrated confidence, whether intentional or not, can significantly exacerbate automation and conservatism biases. This suggests that while confidence distortion may serve as a useful "nudge" in simplified settings, it may also compromise decision quality when the misalignment between confidence and correctness is not apparent to users.

To address this challenge, it becomes essential to investigate whether users can meaningfully interpret recently proposed calibration metrics, such as the Estimated Calibration Index (ECI) (Famiglini, Campagner, and Cabitza 2023), which aim to improve transparency and interpretability compared to more widely used alternatives like the Brier Score or Expected Calibration Error (ECE).

Limitations and Future Work

While our study provides valuable insights into the role of confidence calibration in AI-assisted decision-making, we must acknowledge certain limitations. These limitations, however, do not undermine the credibility of our findings but rather highlight promising avenues for future research.

One potential limitation concerns the real-world applicability of our study. The decision-making tasks were based on logic puzzles, which, while cognitively demanding, may not fully capture the complexities of AI-assisted decision-making in high-stakes domains such as medicine, finance, or law. However, this methodological choice allowed us to control for confounding variables and isolate the effects of confidence calibration without the influence of domain-specific knowledge or external factors. By demonstrating robust effects in a simplified, yet rigorous setting, our findings establish a strong foundation for future research that can explore these mechanisms in more applied contexts.

A related limitation involves the representation of AI confidence, which in our study was conveyed through categorical labels (high, medium, and low) rather than numerical probability values. While numerical confidence scores could offer a more granular representation of uncertainty, prior research suggests that lay people often struggle to interpret probability values correctly (Fischhoff and Bruine De Bruin 1999). Our choice to use categorical labels aligns with existing human-centered design principles that prioritize interpretability over precision, especially for non-expert users (Zhang, Liao, and Bellamy 2020). Nevertheless, future work should investigate whether alternative representations–such as mixed confidence displays (e.g., combining numerical values with verbal descriptions)–might influence reliance behavior differently.

Another limitation concerns individual differences among participants, which were not explicitly accounted for in our study. Although the sample was demographically diverse (mean age = 29.46, SD = 14.37), we did not control for variations in logical reasoning ability, prior experience with AI systems, or cognitive styles such as risk aversion and risk-seeking tendencies (Kahneman 2003). These factors may influence how users interpret and respond to AI-generated confidence cues. For instance, risk-averse individuals may exhibit conservatism bias, while risk-seeking participants might be more prone to automation bias. While our within-subjects design helps mitigate some of this variability by exposing each participant to both calibrated and miscalibrated conditions, it does not fully account for such psychological and experiential differences. Future research should investigate whether expertise in decision-making domains, familiarity with AI systems, or stable personality traits modulate users' reliance on AI.

In addition, using a fixed question order and the same items across conditions may have introduced confounding factors related to learning or question difficulty. Future studies should address these design limitations through randomized question ordering and probabilistic calibration.

Finally, our study assumes a relatively universal interpretation of confidence cues, but cultural and individual differences in confidence perception were not explicitly addressed. Prior research has demonstrated that trust in automated systems varies across cultural contexts, with users from different backgrounds exhibiting distinct preferences for transparency and uncertainty expression (Okamura and Yamada 2020). While our findings are broadly applicable, future work should examine whether confidence calibration effects differ across cultural groups and whether AI interfaces should adapt confidence expressions to align with culturally specific expectations of reliability.

Conclusion

This study investigated how AI confidence calibration affects human reliance, decision accuracy, and perceived utility. Although participants were unaware of the calibration quality, well-calibrated confidence scores improved decision-making and reduced automation and conservatism bias. When AI confidence aligned with correctness, participants showed greater accuracy and appropriate reliance. In contrast, miscalibrated scores led to over-reliance on incorrect advice or undue skepticism, thus impairing decisions.

Despite some limitations, the findings remain significant and suggest directions for future research, particularly on how confidence calibration operates across different contexts, populations, and confidence representation schemes.

These results underscore the need for AI systems to express confidence in ways that accurately reflect the reliability of their recommendations. Poor calibration can influence user behavior and reinforce inappropriate reliance.

Ultimately, effective communication of confidence is critical for fostering appropriate human reliance. As AI continues to be integrated into decision-support systems, striking the right balance between transparency and trust will be key to maximizing its benefits while minimizing risks.

Ethical Statement

The Ethical Review Board of the University of Milano-Bicocca reviewed and approved the methodology described in this article. The study was conducted according to the approved protocol. Informed consent was obtained from all volunteers prior to participation, and no personal or sensitive information was collected.

Acknowledgments

CF, FC and LV acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision-making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

References

Banerji, C. R.; Chakraborti, T.; Harbron, C.; and MacArthur, B. D. 2023. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature medicine*, 29(12): 2996–2998.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and 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*, 1–16.

Bussone, A.; Stumpf, S.; and O'Sullivan, D. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, 160–169. IEEE.

Cabitza, F.; Campagner, A.; Angius, R.; Natali, C.; and Reverberi, C. 2023a. AI shall have no dominion: on how to measure technology dominance in AI-supported human decision-making. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–20.

Cabitza, F.; Campagner, A.; Fregosi, C.; Cameli, M.; Gallazzi, E.; Sconfienza, L. M.; and Tontini, G. E. 2025. Five Degrees of Separation: Investigating the Unexpected Potential of Displaced Human-AI Collaboration Protocols for Apter AI Support. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–28.

Cabitza, F.; Campagner, A.; Ronzio, L.; Cameli, M.; Mandoli, G. E.; Pastore, M. C.; Sconfienza, L. M.; Folgado, D.; Barandas, M.; and Gamboa, H. 2023b. Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine*, 138: 102506.

Cabitza, F.; Fregosi, C.; Campagner, A.; and Natali, C. 2024. Explanations considered harmful: The Impact of misleading Explanations on Accuracy in hybrid human-AI decision making. In *World Conference on Explainable Artificial Intelligence*, 255–269. Springer.

Cao, S.; Liu, A.; and Huang, C.-M. 2024. Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–32.

Dahlbäck, N.; Jönsson, A.; and Ahrenberg, L. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, 193–200.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1): 114.

Eysenck, H. J. 1976. *Q.I. Nuovi test d'intelligenza*. Milano: Feltrinelli.

Famiglini, L.; Campagner, A.; and Cabitza, F. 2023. Towards a Rigorous Calibration Assessment Framework: Advancements in Metrics, Methods, and Use. In *ECAI 2023*, 645–652. IOS Press.

Fischhoff, B.; and Bruine De Bruin, W. 1999. Fifty–fifty= 50%? *Journal of Behavioral Decision Making*, 12(2): 149–163.

Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1): 121–127.

Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics*, 83(5): 368–375.

Huang, H.-Y.; and Bashir, M. 2017. Personal influences on dynamic trust formation in human-agent interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction*, 233–243.

Jiang, J.; Kahai, S.; and Yang, M. 2022. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies*, 165: 102839.

Kahneman, D. 2003. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9): 697–720.

Kocielnik, R.; Amershi, S.; and Bennett, P. N. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kruger, J.; and 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.

Li, Z.; and Steyvers, M. 2025. The Importance of Metacognitive Sensitivity in Human-AI Decision-Making. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Marx, C.; Park, Y.; Hasson, H.; Wang, Y.; Ermon, S.; and Huan, L. 2023. But are you sure? an uncertainty-aware perspective on explainable ai. In *International Conference on Artificial Intelligence and Statistics*, 7375–7391. PMLR.

Mortier, T.; Bengs, V.; Hüllermeier, E.; Luca, S.; and Waegeman, W. 2023. On the Calibration of Probabilistic Classifier Sets. In *International Conference on Artificial Intelligence and Statistics*, 8857–8870. PMLR.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. In *CVPR workshops*, volume 2.

Okamura, K.; and Yamada, S. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one*, 15(2): e0229132.

Price, P. C.; and Stone, E. R. 2004. Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, 17(1): 39–57.

Pulford, B. D.; Colman, A. M.; Buabang, E. K.; and Krockow, E. M. 2018. The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General*, 147(10): 1431.

Sah, S.; Moore, D. A.; and MacCoun, R. J. 2013. Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, 121(2): 246–255.

Schemmer, M.; Kuehl, N.; Benz, C.; Bartos, A.; and Satzger, G. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422.

Schoeffer, J.; De-Arteaga, M.; and Kuehl, N. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18.

Seuß, D. 2021. Bridging the gap between explainable AI and uncertainty quantification to enhance trustability. *arXiv preprint arXiv:2105.11828*.

Silva Filho, T.; Song, H.; Perello-Nieto, M.; Santos-Rodriguez, R.; Kull, M.; and Flach, P. 2023. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9): 3211–3260.

Sodian, B.; Kristen-Antonow, S.; and Kloo, D. 2020. How does children's theory of mind become explicit? A review of longitudinal findings. *Child Development Perspectives*, 14(3): 171–177.

Tejeda, H.; Kumar, A.; Smyth, P.; and Steyvers, M. 2022. AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, 5(4): 491–508.

Tenney, E. R.; MacCoun, R. J.; Spellman, B. A.; and Hastie, R. 2007. Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, 18(1): 46–50.

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and 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.

Vodrahalli, K.; Gerstenberg, T.; and Zou, J. 2022. Uncalibrated Models Can Improve Human-AI Collaboration. *Advances in Neural Information Processing Systems*, 35(NeurIPS).

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.