PNAS NEXUS

PNAS Nexus, 2025, **4**, pgaf133
https://doi.org/10.1093/pnasnexus/pgaf133
Advance access publication 24 April 2025
Perspective

Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI

Doyeon Lee[a], Joseph Pruitt[a], Tianyu Zhou[b], Jing Du[b] and Brian Odegaard[a,c,*]

[a]Department of Psychology, University of Florida, 945 Center Dr., P.O. Box 112250, Gainesville, FL 32611, USA
[b]Department of Civil and Coastal Engineering, University of Florida, Weil Hall 360, 1949 Stadium Road, Gainesville, FL 32611, USA
[c]Present address: Department of Psychology, University of Florida, Gainesville, FL, USA
*To whom correspondence should be addressed: Email: bodegaard@ufl.edu
Edited By Stephen Fleming

Abstract

Knowing when to trust and incorporate the advice from artificially intelligent (AI) systems is of increasing importance in the modern world. Research indicates that when AI provides high confidence ratings, human users often correspondingly increase their trust in such judgments, but these increases in trust can occur even when AI fails to provide accurate information on a given task. In this piece, we argue that measures of metacognitive sensitivity provided by AI systems will likely play a critical role in (1) helping individuals to calibrate their level of trust in these systems and (2) optimally incorporating advice from AI into human-AI hybrid decision making. We draw upon a seminal finding in the perceptual decision-making literature that demonstrates the importance of metacognitive ratings for optimal joint decisions and outline a framework to test how different types of information provided by AI systems can guide decision making.

Keywords: artificial intelligence, metacognitive sensitivity, joint decision making, optimal decisions, trust calibration

Research on artificial intelligence (AI) has entered a new era. The last decade has seen remarkable breakthroughs in large language models (LLMs) (1–3), healthcare and medical-related AI (4–6), automated navigation systems (7–9), and many other fields. The ability of AI to perform specific tasks that meet or exceed human performance spans a range of domains (10, 11), but there are plenty in which performance is still suboptimal, stressing the possibility that, at least for now, collaboration with humans represents a reasonable strategy to make the best possible decisions, given noise and uncertainty in available information (12–14).

How might optimal decision making between humans and AI be achieved? In order to effectively utilize the information supplemented by AI, one issue that must be addressed is whether humans properly calibrate their degree of trust in the information coming from systems and agents (15–18). When the mechanisms and reliabilities of AI's decision-making processes are unknown, it can become difficult for individuals to decide whether or not they should trust the outputs of such systems, especially because many AI systems still err in specific judgments (19–21). Thus, a challenge remains: to optimize human decision making, it may need to be supplemented (if not ultimately supplanted) by AI as individuals incorporate advice from AI systems, but if humans fail to properly discern when to trust those systems, they are unlikely to make the best possible use of AI's advice.

In this piece, we propose that reports of metacognitive sensitivity are a key component for 2 related issues: (1) calibrating trust in AI systems (22) and (2) optimally incorporating information in decision making (23). As others have noted, to foster successful collaboration, what is needed is the capacity for AI to selectively share relevant states with humans to facilitate coordination and cooperation (24). Drawing upon insights from the field of perceptual metacognition, we propose that AI must report performance on not only "type 1" tasks (e.g. choosing or discriminating between choice alternatives), but also "type 2" tasks (e.g. reporting metrics of how correct and incorrect judgments are "correctly" endorsed with high and low confidence, respectively) (25–27) to facilitate trust and optimal decision making. Building on recent proposals in the AI literature (28), and literature exploring optimal decision making in multiple human observers (29–31), we outline a framework that can be exploited to test how to calibrate human trust in AI's decisions, and how to explore whether those decisions are reaching the best possible outcomes.

Collaboration with AI can be beneficial, but trust issues remain

Across several domains, hybrid decisions made by humans interacting with AI have proven to be an especially powerful approach to enhance decision making (32). For example, AI-based support for skin cancer diagnoses has been shown to be superior to AI or human decision making alone (33), as have diagnoses for radiologists' decisions supplemented by AI (34). In medicine and

Competing Interest: We have no conflicts of interest to disclose.
Received: November 19, 2024. **Accepted:** April 4, 2025

healthcare, the possibility of AI aiding in tasks such as improving the accuracy of diagnosis or defining new preventions or treatments has been identified by doctors as being especially promising (35). In K–12 education, teacher-AI partnerships have been shown to improve test performance (36), and AI can be successfully used to aid in deception detection (37). However, hybrid human-AI performance is not always better than human or AI performance alone (37–41), and appears to often be most beneficial when AI and human accuracy on a given task is comparable (42).

Incorporating and accepting AI input in decisions is contingent on trust (43). How AI is represented (e.g. robot, virtual, or embedded) and its level of machine intelligence (i.e. capabilities and capacities) are factors that influence humans' trust in such systems (44). Trust can increase when AI provides explicit information about its reasoning (45). However, situations can arise in which explanations from AI increase human trust in its advice but fail to improve accuracy, providing evidence that trust and accuracy can dissociate (46). Preliminary work shows that having AI rate confidence in its judgments and advice can also change human trust in its outputs (47) even if accuracy is unchanged (48), but importantly, confidence ratings alone are not necessarily enough to ensure proper usage of AI's advice, showing that trust calibration does not always translate into improvement in AI-assisted decision outcomes (49). We argue that what has been missing in most studies to date is for AI to report its metacognitive sensitivity, or the correspondence between its confidence judgments and accuracy on specific tasks (27, 50, 51), and that these reports will be key to knowing when to trust AI, and how to efficiently and accurately incorporate its advice.

Defining metacognition in humans and AI

Before discussing metacognitive sensitivity, it is first important to define what metacognition is. Researchers in cognitive psychology have long been interested in individuals' ability to effectively evaluate or oversee their own cognitive processes, an ability that is known as "metacognition" (52–54). This oversight has been identified as being critical for implementing effective student learning strategies (55–57) (such as evaluating one's state of knowledge to decide what to study), monitoring the accuracy of perceptual decisions (58, 59), and understanding social context (60). Modern conceptions of metacognition have been shaped by Flavell, who proposed that cognitive monitoring could be broadly distinguished by four classes of phenomena: metacognitive knowledge (knowledge of how variables and factors interact to influence cognitive outcomes), metacognitive experiences (conscious reflections about cognitive processes), metacognitive goals (objectives of a cognitive enterprise), and metacognitive actions (strategies employed to achieve cognitive goals) (61).

More recently, Fleming (62) specifically defined metacognition as the set of processes that enable us to develop beliefs about various mental functions and operations. According to this account, judgments that are "metacognitive" in nature have ways of taking into account uncertainty. Uncertainty could take the form of noisy estimates of sensory properties in the world (63–66), assessments of how well material has been learned (67–69), precision in controlling motor actions (70), and many other perceptual and cognitive domains. The key idea is that the representation of certainty is coded with respect to a self-centered frame of reference; uncertainty is transformed into some type of propositional assessment (such as a confidence judgment), and this judgment is globally broadcast in the system for widespread availability in guiding behaviors and updating current self-knowledge. To provide a more concrete example, according to this account, any estimate of a sensory property or attribute in the environment is encoded with some degree of noise, and we form beliefs about how likely it is that certain features of the world are present. Those beliefs can be converted into propositional confidence in our estimates (which are also subject to noise).

This description provides one way to link confidence judgments and real-world metacognition; humans have an ability to assess whether self-evaluative judgments match up with the reality of cognitive or physical performance, which is central to adaptive behavior (62). Similarly, AI also has the capacity to reflect on the degree to which its self-evaluative judgments line up with reality in the long run. Thus, drawing upon insights from human perceptual decision making, an opportunity arises to explore how metacognitive sensitivity can be exploited to calibrate human trust with artificial systems. In the following sections, we define metacognitive sensitivity and explain its importance for human-AI interactions.

Measuring and quantifying metacognitive sensitivity

Metacognition can be measured via different types of reports about the fidelity of cognitive processes. Historically, confidence judgments have provided one useful form of report about introspection (52, 53, 71–73). Confidence reports lead to two different metacognitive quantities: metacognitive bias and metacognitive sensitivity (50). While metacognitive bias is reflected in long-run averages of over- or underconfidence in judgments of task performance, metacognitive sensitivity is reflected in the confidence-accuracy correlation, or more specifically, how effectively confidence judgments distinguish between correct and incorrect answers (27, 74, 75). High metacognitive sensitivity reflects a strong correlation (responding with high confidence when correct and low confidence when incorrect), and low metacognitive sensitivity reflects more irregular or haphazard confidence ratings on a trial-by-trial basis.

In Figure 1, we outline the types of reports that AI can provide as it makes decisions. Essentially, there are four things that AI could report: (1) its decision about a given perceptual (or cognitive) process, or (relatedly) its long-run accuracy in a specific task; (2) a subjective estimate about a specific decision (such as how confident it is in a particular judgment); (3) a report about its long-run metacognitive sensitivity for making these types of judgments (i.e. the correspondence between confidence ratings and accuracy for a specific task or type of decision); (4) more complex and intricate introspection regarding the different stages of its decision-making process, to explicate on why it made specific type 1 and type 2 decisions (Figure 1). This framework could be tested empirically to determine how different types of information from AI systems influence (1) the calibration of trust for these systems and (2) the incorporation of information from AI into joint human-AI decisions.

Quantifying task performance and metacognitive sensitivity has been facilitated using tools from signal detection theory (SDT). In type 1 SDT, the measure d' quantifies the subject's ability to discriminate between signal and noise, with higher d' values indicating better performance (76). Response bias for selecting one option over another (in two-choice tasks) is captured by the perceptual criterion, c, which indicates the level of sensory evidence needed to make a decision (77). In contrast, type 2 SDT focuses on metacognitive processes, evaluating how well participants can distinguish between their own correct and incorrect responses. Metacognitive sensitivity, measured by $meta\text{-}d'$, estimates how

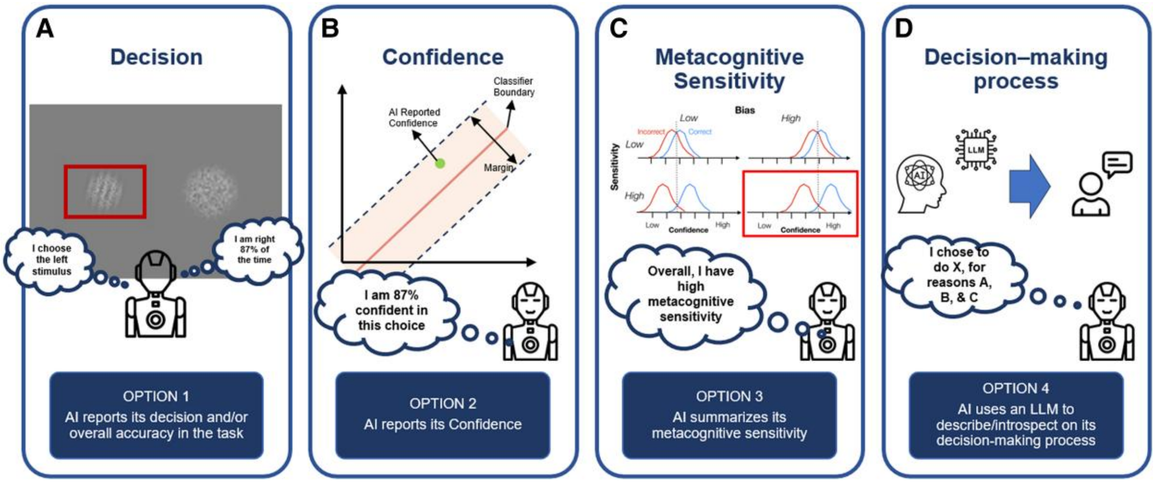
A
Decision
I choose the left stimulus
I am right 87% of the time
OPTION 1
AI reports its decision and/or overall accuracy in the task
B
Confidence
Classifier Boundary
AI Reported Confidence
Margin
I am 87% confident in this choice
OPTION 2
AI reports its Confidence
C
Metacognitive Sensitivity
Bias
Low
High
Incorrect
Correct
Sensitivity
Low
High
Low
Confidence
High
Low
Confidence
High
Overall, I have high metacognitive sensitivity
OPTION 3
AI summarizes its metacognitive sensitivity
D
Decision–making process
AI
LLM
I chose to do X, for reasons A, B, & C
OPTION 4
AI uses an LLM to describe/introspect on its decision-making process

Fig. 1. Possible types of reports from AI about its decisions. (A) Type 1 reports about behaviors. One possibility is that AI may only report the outcome of a specific decision it has chosen to make, or how frequently it is correct for a particular type of decision. In the example shown here, the AI agent determines which stimulus contains an oriented Gabor and reports its overall accuracy across trials on this task. (B) Subjective (e.g. confidence) reports. Potentially, AI could report how confident it is in a specific decision, or some other type 2 subjective rating. Depending on the algorithm used, this report could be based on something akin to distance from a decision or classification boundary using signal detection theory, support vector machines, or some other algorithm. (C) Reports of metacognitive sensitivity, schematic from Fleming and Lau (50). AI systems may also be able to summarize their metacognitive sensitivity, or how effectively their confidence judgments distinguish between correct and incorrect judgments over long-run averages. In this sense, observers could gain insights into the degree to which they could use/trust the metacognitive or subjective ratings offered by a system on a given trial. (D) Finally, AI models could also try to introspect on their decision-making process to identify why or how they reached specific type 1 or type 2 decisions. Critically, these insights could be about the mechanisms or functions that led to specific type 1 or type 2 decisions. Interestingly, these types of introspections can sometimes lead to increased trust but not to increased accuracy (49).

effectively confidence judgments distinguish between correct and incorrect judgments (27, 74). Participants who are more confident in correct responses and less confident in incorrect ones will have higher *meta-d'* scores. Because this measure of metacognitive sensitivity scales with task performance, an additional measure, *metacognitive efficiency*, given by the M-ratio (*meta-d'/d'*), adjusts *meta-d'* for task performance level, recognizing that participants with higher *d'* have a higher potential for *meta-d'* (50, 78). In this sense, measures of metacognitive sensitivity and metacognitive efficiency may both prove important for facilitating human trust in AI; quantifying metacognitive sensitivity gives information about the correspondence between metacognitive ratings and task accuracy, but quantifying metacognitive efficiency incorporates considerations of task performance, and because this value can be directly compared against an optimal value of 1, it may be more easily interpreted by users when evaluating whether to trust AI's judgments and recommendations.

Literature on perceptual decision making provides examples of paradigms that can be used to assess metacognitive sensitivity (50). Many tasks in which researchers measure *meta-d'* follow a specific format: on each trial, observers make a perceptual judgment (such as the presence or tilt of a Gabor) and rate their confidence in that judgment (51, 79, 80). Because animals cannot explicitly provide confidence ratings, adding the option of choosing "bets" of different values can be substituted for confidence ratings in perceptual and memory paradigms, facilitating measurement of *meta-d'* in monkeys (81, 82). Further, waiting times for rewards can also be taken as a proxy for confidence ratings, allowing estimates of meta-*d'* in rats (83). Thus, across a range of tasks and paradigms, different systems and organisms capable of metacognition can have their metacognitive sensitivity quantified.

One challenge that will need to be addressed moving forward involves extending measures of metacognitive sensitivity to more complex and naturalistic tasks. In the scientific literature, the current precedent for measuring metacognitive sensitivity (with measures such as *meta-d'*) comes from well-controlled designs in which type 1 decisions are discrete, rather than continuous. In more naturalistic environments (such as drone navigation, in which automation and user control might trade-off at different time points), multidimensional decisions such as which routes to select, the elevation height of travel, the speed at which to proceed, and other factors might make it more difficult to quantify metacognitive sensitivity (and efficiency). As noted in a recent review (84), while some measures of metacognition handle continuous confidence judgments effectively, there appear to be few (if any) options that handle continuous type 1 judgments well. Progress is being made on this topic in new experiments probing metacognitive sensitivity in sensorimotor domains (85, 86), but additional research is needed. With AI metacognition having recently been identified as being a critical component of cooperation and safety in human-AI interactions (87, 88), further refinement and testing of assessments in continuous domains becomes even more important.

Metacognition (and metacognitive sensitivity) is key for optimal joint decision making

Most previous tasks to assess metacognitive sensitivity have evaluated this trait in single observers. However, previous work on optimal decisions in multiple (n = 2) observers provides insights into how metacognitive ratings can aid joint decision making. In seminal work by Bahrami et al. (29), the authors noted that within a single mind, there are rules which govern how multiple senses are optimally combined into a coherent percept: when discrepant estimates of a quantity must be reconciled, observers weight each sense's estimate (e.g. visual and auditory estimates about spatial position) by their reliability (63, 89, 90). Similarly, when multiple (n = 2) observers must make decisions about ambiguous sensory information, Bahrami et al. found something interesting: for 2 observers with similar perceptual sensitivities, the optimal rule was to combine their 2 estimates in a manner similar to the optimal rules previously found in studies of multisensory integration, but when one observer was substantially better than another, combining estimates comes with a cost, rather than a benefit (91). In their perceptual task, on each trial, 2 participants each observed 2 intervals with Gabor patches, and had to judge whether the first or second stimulus contained the "oddball" Gabor with higher contrast (Figure 2). Across 4 experiments, the authors manipulated noise levels in the stimuli, whether or not communication was allowed, and whether feedback was provided.

How would one observer know whether to incorporate another mind's estimate, or keep it separate? The authors found that communication about confidence in the estimate was key. The authors compared 4 possible models about what could be communicated: a coin flip model (essentially resolving disagreements by flipping a coin), a behavior and feedback model (in which participants cannot communicate anything besides their choice during the trial), a weighted confidence sharing model (in which participants communicate their confidence in their judgment), and a direct signal sharing model (in which the mean and standard deviation of the sensory responses are shared). Results showed that the weighted confidence model best fit the data from the task and demonstrated the benefits of multiple minds sharing confidence judgments in joint decision making.

What are the insights that can be gleaned from Bahrami et al. (29) and how does this relate to trust and optimal decision making in AI? Several key ideas emerged: first, it was only optimal for one observer to incorporate advice from another when performance was at similar levels, something that has been shown previously in the AI literature (42). Thus, we posit that for humans to trust AI, an AI agent needs to report not only type 1 judgments about what it chooses, but also (1) how confident it is in that specific judgment and (2) its long-run accuracy in a given task that it performs (60). Second, trial-by-trial confidence judgments proved key for the collective benefit in decision making (92, 93), but feedback did not impact decision making. Thus, this finding supports the idea that when a second observer provides a type 2 report of confidence, it can help observers make an optimal decision. However, we argue that for confidence decisions to be meaningful, they must differentiate (at least to some degree) between correct and incorrect judgments. Without a reasonable degree of metacognitive sensitivity, confidence judgments would be of little help.

Indeed, Pescetelli et al. (60) clarified that the weighted confidence sharing model assumes that interacting individuals' metacognitive sensitivities are both good and similar to each other, and their empirical results showed a direct correlation between the mean metacognitive sensitivity in the dyad and the collective benefit of dyadic performance over individual performance. Thus, it takes more than communicating confidence for joint decisions to be beneficial; metacognitive sensitivity is needed, too. While the correlation between metacognitive sensitivity and benefits in joint decision making revealed in this work is important (60), future work will need to characterize how communicating metacognitive sensitivity impacts decision making not only when metacognitive sensitivity differs across human-AI dyads, but also when overall accuracy differs across the dyad, too.

Critically, we note that even if confidence judgments (with adequate metacognitive sensitivity) might help humans calibrate when to trust AI systems, it is not a guarantee that they will optimally incorporate that information (94). In this perspective, we posit that trust and optimal decision making are distinct. Drawing upon research in other fields, evidence indicates that trust can dissociate from related concepts. For example, in consumer adoption studies, "trust" and "intention to adopt" are distinct factors, understood as hierarchical stages of perception and behavioral intention (43, 95). Further, factors including AI personality, anthropomorphism, and behavior all influence trust in AI, apart from the effects of the quality of information that it provides (44, 96). Therefore, we hypothesize that human agents may calibrate their trust in AI to a given level, but we do not think this guarantees that they will combine their own judgments with that of AI in a way defined by a given "optimal" rule. However, our thesis is still that metacognitive sensitivity provides the key to both elements: it allows humans to calibrate the level of trust in AI agents and can inform the weights humans assign to decisions made by AI.

The rules governing optimal decision making will vary depending on the types of tasks that are involved. Further, as noted in previous studies of joint decision making (92), how task demands are distributed, how the information is exchanged and integrated, and whether feedback is provided are all factors that can influence outcomes. Despite these factors playing a role, having a meaningful degree of metacognitive sensitivity is key for collective benefits in joint decision making (60, 97).

Can humans use measures of metacognitive sensitivity to enhance joint decisions with AI?

One example of how metacognitive sensitivity could play a role in joint human-AI decision making is found in identifying perceptual signatures of cancer in medical images. For example, using magnetic resonance imaging, specific perceptual features define prostate cancer in 2 regions of the prostate: the transition zone and the peripheral zone. In the peripheral zone for diffusion-weighted images, hyperintense patches denote cancer, but in the transition zone, hypointense patches denote cancer (98, 99). Additionally, cancerous lesions are defined by shape (e.g. linear, wedge, round, lenticular), signal intensity (e.g. mild, moderate, markedly hypo/hyperintense), and/or boundary type (e.g. completely or mostly encapsulated, obscured margins). Current clinical training uses different combinations of these 3 groups for category assignment of detected lesions. If an AI system is trained to detect prostate cancer and performs at a similar overall level to human doctors, how would a doctor know whether to combine his or her diagnosis with that of AI in a specific instance? AI's confidence ratings in a diagnosis could provide further information, but only to the degree to which confidence ratings distinguish between correct and incorrect judgments. Further, identifying the unique types of errors made by the AI (such as whether it overgeneralizes learned perceptual features from one part of the prostate to another) is another critical issue, especially if those errors are made with high confidence. The lack of metacognition in AI has recently been noted as a hindrance for reliable medical reasoning (100), stressing the need for evaluating metacognitive sensitivity for automated medical diagnoses in the future.

One issue that inevitably arises is the following: if AI reports its metacognitive sensitivity, how should it display this metric, so that humans understand what the measure means and use it effectively? A range of measures for metacognitive sensitivity exist, with pros and cons for different methods of quantifying the correspondence between confidence and accuracy (78), but these measures are not necessarily readily interpretable. For example, telling a human observer that AI has a *meta-d'* value of "3.2" does not provide an immediate heuristic for most observers to

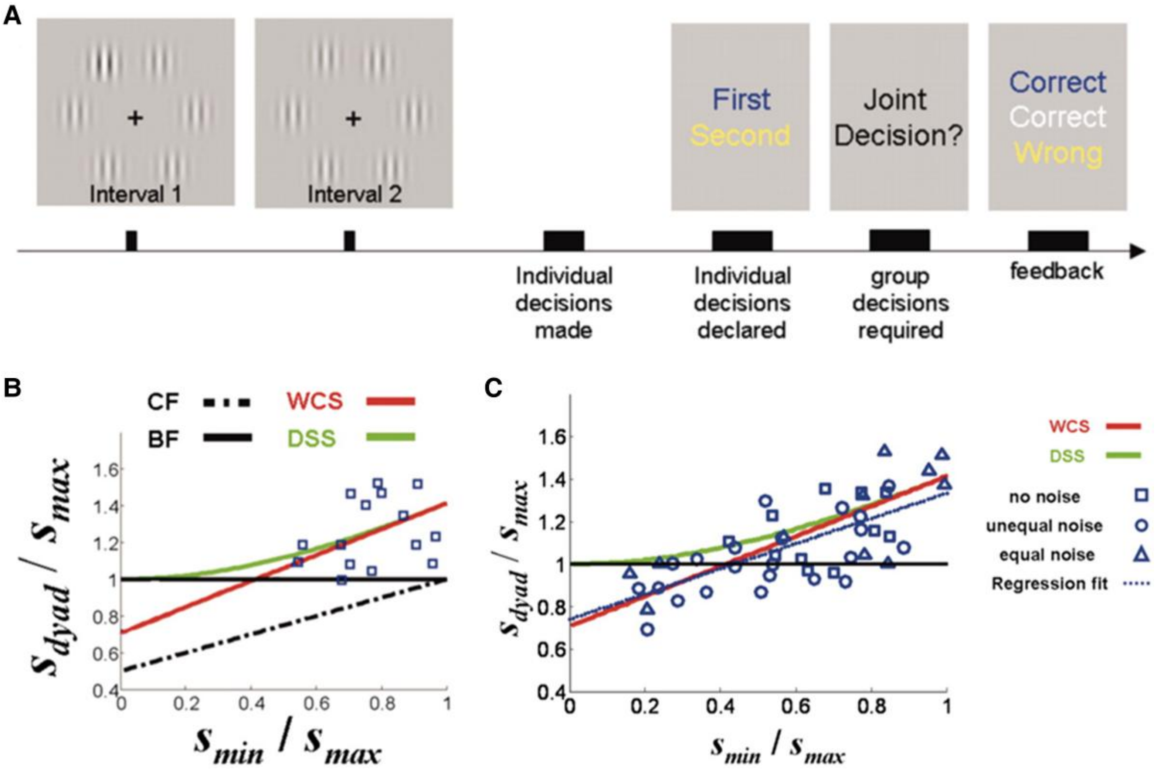

Fig. 2. (A) Task paradigm from Bahrami et al. (29). On each trial, 2 human observers viewed 2 displays and had to judge which contained a Gabor with increased contrast. In the first experiment, individual decisions were shown on the screen, and if the observers disagreed, they were allowed to communicate to reach a joint decision. Feedback was shown for each observer's judgment (blue, yellow) and the joint decision (white). (B) Four models were fit to the data from experiment 1 (squares; each square is one pair), and the collective benefit (s_{dyad}/s_{max}) was plotted on the y-axis. Results showed that a model that assumed confidence accurately reflected the probability of being correct (weighted confidence sharing [WCS] model) and a model that assumed that mean and standard deviations of sensory responses are shared (direct signal sharing model [DSS]) fit the data equally well. (C) In a second experiment, extra noise was added to the Gabor patches of one, both, or neither participant's displays. Results showed that a model that incorporated confidence judgments (WCS) provided the best account of the data. BF, behavior and feedback; CF, coin flip.

know whether or not to trust its judgment on a given trial. Moving forward, it will be critical to explicitly test whether metacognitive sensitivity provides benefits over and above reporting percent correct on a task, and how measures of metacognitive sensitivity should be presented.

We suggest 2 possible methods for presenting information about AI's metacognitive sensitivity. First, visualizations of metacognitive sensitivity on a sliding scale (e.g. a vertical, thermometer-like scape) could be presented, ranging from "no metacognitive sensitivity" at the lower end, to "high metacognitive sensitivity" at the upper end. These types of graphs could be preceded by a single instruction such as, "Metacognitive sensitivity measures how effectively confidence distinguishes between correct and incorrect judgments. High metacognitive sensitivity means the AI is often confident in judgments when it is right, and less confident in judgments when it is wrong. Low metacognitive sensitivity means its confidence ratings often fail to track whether it is right or wrong in a specific judgment." Second, it is possible that demonstrations or training may be needed for humans to further understand AI's metacognitive sensitivity, which could be shown by briefly presenting a demonstration of how 2 AI agents (one with high metacognitive sensitivity, one with low metacognitive sensitivity) perform on different trials. These training procedures could occur for 5 to 10 minutes before humans interact in joint decision-making tasks with the AI agents, and questions/feedback could occur and the end of the training to evaluate if humans understand the concept and can identify which AI agent's confidence judgments are more trustworthy.

It is also an open question of whether metacognitive sensitivity should be reported by the system that performs the task itself, or an additional program or system that can monitor long-run performance of a given system or agent. Because AI systems can show biases in their confidence judgments (48, 101), research is needed to determine whether one could also trust metacognitive sensitivity reports in such systems, or whether a more reliable protocol would be to outsource such monitoring to an external source.

Further, combining quantitative and qualitative metacognitive reports may be most effective in optimally combining information in real-world tasks. For example, AI may be particularly effective at optimizing routes of drones for real-world navigation (102–104), but in order to understand why AI is making a specific decision about its speed, route, altitude, or other factors during a task, a combination of quantitative and qualitative descriptions of how different factors were weighted in the decision-making process may be necessary for a human to trust a given decision for whether to override, supplement, or approve a given navigation plan. In this sense, we argue that joint decision making between AI and humans may need LLMs for real-world tasks, as the need to make real-world decisions extends beyond the simplified paradigms used for perceptual metacognition.

Metacognitive sensitivity and LLMs

Work has already begun to explore metacognitive capacities in LLMs. Preliminary research indicates that LLMs tend to exhibit overconfidence in some reasoning tests, in a manner similar to humans (105, 106). Yet, in easy tasks (such as identifying the "best" answer), LLMs show metacognitive sensitivity comparable to humans, with less tendency toward overconfidence (107). In more complex tasks, their sensitivity can surpass human performance (108). Metacognitive sensitivity may depend on the type of judgment that is made: while LLMs and humans are similar in prospective confidence judgments, LLMs show lower metacognitive accuracy in retrospective judgments (109). This difference is attributed to LLMs' reliance on training data for confidence ratings, which are not adjusted based on task experience (109). This lack of calibration contributes to "hallucinations" or "confabulations," often seen in LLMs exhibiting metacognitive myopia (110). To help calibrate human trust in AI systems, recent studies have proposed various strategies for aligning AI confidence and response accuracy through metacognitive learning models (110–112). Once model confidence and accuracy were made consistent, Steyvers et al. (113) found that further metacognitive alignment between the model and user may be facilitated via the LLM reporting its internal model confidence. Therefore, it appears that building in training to assist AI systems in aligning their confidence and accuracy may be key to providing benefits in human-AI collaboration. LLMs provide a tool to explain their decision-making processes to humans (114), but we posit that as LLMs are increasingly incorporated into joint decisions, it will be critical for them not only to explain their decision-making process (Figure 1, option 4), but also to incorporate assessments of metacognitive sensitivity (Figure 1, option 3) as the functions underlying decision making are explicated (114).

One critical issue that needs to be studied is how AI systems can improve metacognitive sensitivity through training. One recent example of successful training comes in the use of synthetic data to improve medical reasoning in LLMs (100). Specifically, via prompt engineering, the model or agent was trained to recognize knowledge limitations and fine-tune its confidence ratings from explicit feedback that it was given for specific medical problems it had to solve. A similar rationale could be applied to training models to diagnose cancer in medical images: if classification models are combined with LLMs and required to reflect upon what specific perceptual decision-making errors that they made and why, it seems possible that this form of iterative feedback may help them to refine both type 1 and type 2 decision making, improving metacognitive sensitivity along the way. However, this needs to be explicitly tested. As humans have many biases that are often resistant to feedback, it remains an open question as to what the best training methods are reduce biases in AI models, and this problem is amplified when AI systems might be trained on datasets (e.g. the Internet) in which misinformation and disinformation abound (115). Thus, for complex topics (e.g. metacognitive sensitivity in making decisions about climate change), rigorous testing must be conducted to determine which sources of evidence should be used for training, and what types of feedback might be most efficacious in improving metacognitive sensitivity.

Future directions and conclusion

Both type 1 and type 2 judgments can vary across human and AI agents, and the information present in each type of report can be beneficial when making joint decisions. Several questions emerge from this: in situations with differences in performance across humans and AI, do models similar to those from Bahrami et al. (29) still account for data, or do humans adopt more complex strategies when deciding to integrate or segregate their judgments with AI, similar to Bayesian causal inference (65)? Further, does metacognitive bias influence whether information is integrated or segregated across agents? And how does the type of domain (perceptual, medical, etc.) influence how the information is used, as judgments are made? It is clear that what is needed moving forward is a rigorous program of research in which performance levels, confidence levels, metacognitive sensitivities, and the domain of decision making are all varied systematically, to better understand how these factors influence human trust in AI, and whether they follow the rules of optimality in joint decisions. This program will likely need to include work on cultivating procedures to try to improve metacognitive sensitivity in AI models; recent work suggests that synthetic data and prompt engineering may be techniques that can improve metacognition (100), but more empirical research is needed.

To date, research provides evidence of possible benefits when observers interact with one another before making joint decisions. This benefit has been observed not only in joint human decision making (29), but also in human-AI interactions (34–37). Preliminary evidence shows that metacognition can be a critical component in the benefits of these interactions (29), but because evidence also indicates that metacognitive ratings can lead humans astray when incorporating advice from AI (49), how will we know when to trust and incorporate its advice? In this piece, we have argued that a critical component is likely to be found in reports from AI about its metacognitive sensitivity. When AI is able to faithfully report on the correspondence between its confidence and accuracy, this information can be utilized by humans to evaluate whether to incorporate its estimates and advice into joint human-AI decisions. Challenges remain in training humans to make use of information about metacognitive sensitivity, and the field needs to develop measures (27, 50, 74) that extend to more complex and naturalistic tasks. But as the world increasingly relies on the use of AI to make decisions in healthcare, finance, and education, we think research to explore the importance of measuring and reporting both type 1 and type 2 sensitivity in AI systems will be critical, as we look to make the best possible use of information that these systems provide.

Acknowledgments

This manuscript was posted on a preprint server: https://osf.io/preprints/psyarxiv/xy8tf.

Funding

This work was supported by an Office of Naval Research Young Investigator Award to B.O. (N00014-22-1-2534), and an Air Force Office of Scientific Research Grant FA9550-22-1-0492 to principal investigator J.D. and co-investigator B.O.

Data Availability

There are no data to be made available, as no data underlie this work beyond the cited articles that are discussed in this paper.

References

1 Zhao WX, *et al.* 2023. A survey of large language models. arXiv 18223. https://doi.org/10.48550/arXiv.2303.18223, preprint: not peer reviewed.

2 Wei J, *et al.* 2022. Emergent abilities of large language models. arXiv 07682. https://doi.org/10.48550/arXiv.2206.07682, preprint: not peer reviewed.

3 Kasneci E, *et al.* 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 103:102274.

4 Alowais SA, *et al.* 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 23(1): 689. https://doi.org/10.1186/s12909-023-04698-z

5 Sherani AMK, Khan M, Qayyum MU, Hussain HK. 2024. Synergizing AI and healthcare: pioneering advances in cancer medicine for personalized treatment. *Int J Multidiscip Sci Arts*. 3(2):270–277.

6 Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. 2020. Artificial intelligence in healthcare: review and prediction case studies. *Engineering*. 6(3):291–301.

7 Aizat M, Azmin A, Rahiman W. 2023. A survey on navigation approaches for automated guided vehicle robots in dynamic surrounding. *IEEE Access*. 11:33934–33955.

8 Thombre S, *et al.* 2022. Sensors and AI techniques for situational awareness in autonomous ships: a review. *IEEE Trans Intell Transp Syst*. 23(1):64–83.

9 Onyekpe U, Palade V, Kanarachos S. 2020. Learning to localise automated vehicles in challenging environments using Inertial Navigation Systems (INS). *Appl Sci*. 11(3):1270. https://doi.org/10.3390/APP11031270

10 Silver D, *et al.* 2017. Mastering the game of go without human knowledge. *Nature*. 550(7676):354–359.

11 Jones N. 2024. AI now beats humans at basic tasks—new benchmarks are needed, says major report. *Nature*. 628(8009):700–701.

12 Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. 2019. 'Hello AI': uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proc ACM Hum Comput Interact*. 3(CSCW):1–24.

13 Jain R, Garg N, Khera SN. 2023. Effective human–AI work design for collaborative decision-making. *Kybernetes*. 52(11):5017–5040.

14 Puranam P. 2021. Human–AI collaborative decision-making as an organization design problem. *J Organ Des*. 10(2):75–80.

15 Ahn D, Almaatouq A, Gulabani M, Hosanagar K. 2021. Will we trust what we don't understand? Impact of model interpretability and outcome feedback on trust in AI. arXiv 08222. https://doi.org/10.48550/arXiv.2111.08222, preprint: not peer reviewed.

16 Troshani I, Rao Hill S, Sherman C, Arthur D. 2021. Do we trust in AI? Role of anthropomorphism and intelligence. *J Comput Inf Syst*. 61(5):481–491.

17 Alvarado R. 2022. What kind of trust does AI deserve, if any? *AI Ethics*. 3(4): 1169–1183.

18 Bedué P, Fritzsche A. 2022. Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *J Enterp Inf Manage*. 35(2):530–549.

19 Agudo U, Liberal KG, Arrese M, Matute H. 2024. The impact of AI errors in a human-in-the-loop process. *Cogn Res Princ Implic*. 9(1): 1. https://doi.org/10.1186/s41235-023-00529-3

20 Chanda SS, Banerjee DN. 2022. Omission and commission errors underlying AI failures. *AI Soc*. 39(3):1–24.

21 Russell S, Moskowitz IS, Raglin A. Human information interaction, artificial intelligence, and errors. In: Lawless WF, Mittu R, Sofge D, Russell S, editors. *Autonomy and artificial intelligence: a threat or savior?* Springer International Publishing, Cham, 2017. p. 71–101.

22 Bacciu D, *et al.* TEACHING—trustworthy autonomous cyber-physical applications through human-centred intelligence. In: 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS), August 23–25, 2021, Barcelona, Spain.

23 Rahnev D, Denison RN. 2018. Suboptimality in perceptual decision making. *Behav Brain Sci*. 41: e223.

24 Deroy O, Bacciu D, Bahrami B, Della Santina C, Hauert S. 2024. Shared awareness across domain-specific artificial intelligence: an alternative to domain-general intelligence and artificial consciousness. *Adv Intell Syst*. 6(10):2300740.

25 Clarke FR, Birdsall TG, Tanner WP. 1959. Two types of ROC curves and definitions of parameters. *J Acoust Soc Am*. 31(5): 629–630.

26 Galvin SJ, Podd JV, Drga V, Whitmore J. 2003. type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon Bull Rev*. 10(4):843–876.

27 Maniscalco B, Lau H. 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn*. 21(1):422–430.

28 Fleming SM. 2023. Metacognitive psychophysics in humans, animals, and AI: a research agenda for mapping introspective systems. *J Conscious Stud*. 30(9–10):113–128.

29 Bahrami B, *et al.* 2010. Optimally interacting minds. *Science*. 329(5995):1081–1085.

30 Fusaroli R, *et al.* 2012. Coming to terms: quantifying the benefits of linguistic coordination: quantifying the benefits of linguistic coordination. *Psychol Sci*. 23(8):931–939.

31 Barrera-Lemarchand F, Balenzuela P, Bahrami B, Deroy O, Navajas J. The Wisdom of Extremized Crowds: Promoting Erroneous Divergent Opinions Increases Collective Accuracy. 2023 [accessed 2024 Nov]. https://repositorio.utdt.edu/server/api/core/bitstreams/c3326f09-b161-4f6a-9354-76612e1741a6/content. https://scholar.google.com/citations?view_op=view_citation&hl=en&citation_for_view=eIghKWMAAAAJ:S16KYo8Pm5AC.

32 Rastogi C, Leqi L, Holstein K, Heidari H. 2022. A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity. arXiv 10806. https://doi.org/10.48550/arXiv.2204.10806, preprint: not peer reviewed.

33 Tschandl P, *et al.* 2020. Human-computer collaboration for skin cancer recognition. *Nat Med*. 26(8):1229–1234.

34 Patel BN, *et al.* 2019. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med*. 2(1): 111.

35 Göndöcs D, Dörfler V. 2024. AI in medical diagnosis: AI prediction & human judgment. *Artif Intell Med*. 149:102769.

36 Holstein K, Aleven V. 2022. Designing for human–AI complementarity in K-12 education. *AI Mag*. 43(2):239–248.

37 Lai V, Tan C. 2019. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: ACM.

38 Lai V, Liu H, Tan C. 2020. 'Why is "Chicago" deceptive?' Towards building model-driven tutorials for humans. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York (NY): ACM.

39 Green B, Chen Y. 2019. The principles and limits of algorithm-in-the-loop decision making. In: Proceedings of the ACM on Human-Computer Interaction, Volume 3, Issue CSCW. New York (NY): ACM.

40 Lundberg SM, *et al.* 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2(10):749–760.

41 Buçinca Z, Lin P, Gajos KZ, Glassman EL. 2020. Proxy tasks and subjective measures can be misleading in evaluating

explainable AI systems. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. New York (NY): ACM.

42 Bansal G, *et al.* 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. New York (NY): ACM.

43 Choung H, David P, Ross A. 2022. Trust in AI and its role in the acceptance of AI technologies. *Int J Hum-Comput Interact*. 40(7): 1532–1544.

44 Glikson E, Woolley AW. 2020. Human trust in artificial intelligence: review of empirical research. *Acad Manag Ann*. 14(2): 627–660.

45 Vössing M, Kühl N, Lind M, Satzger G. 2022. Designing transparency for effective human-AI collaboration. *Inf Syst Front*. 24(3): 877–895.

46 Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. 11 August 2024. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. arXiv 15266. https://doi.org/10.48550/arXiv.2408.15266, preprint: not peer reviewed.

47 Bruzzese T, Gao I, Dietz G, Ding C, Romanos A. 2020. Effect of confidence indicators on trust in AI-generated profiles. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. New York (NY): ACM.

48 Rechkemmer A, Yin M. 2022. When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models. In: CHI Conference on Human Factors in Computing Systems. New York (NY): ACM.

49 Zhang Y, Vera Liao Q, Bellamy RKE. 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. New York (NY): ACM.

50 Fleming SM, Lau HC. 2014. How to measure metacognition. *Front Hum Neurosci*. 8:443.

51 Maniscalco B, Peters MAK, Lau H. 2016. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Atten Percept Psychophys*. 78(3): 923–937.

52 Peirce CS, Jastrow J. On Small Differences in Sensation. 1884 [accessed 2024 Oct 22]. https://philarchive.org/archive/PEIOSD.

53 Henmon VAC. 1911. The relation of the time of a judgment to its accuracy. *Psychol Rev*. 18(3):186–201.

54 Gigerenzer G. 1991. How to make cognitive illusions disappear: beyond 'heuristics and biases.' *Eur Rev Soc Psychol*. 2(1):83–115.

55 Serra MJ, Metcalfe J. 2009. Effective implementation of metacognition. In: Hacker DJ, Dunlosky J, Graesser AC, editors. *Handbook of metacognition in education*. Boca Raton (FL): Routledge/Taylor & Francis Group. p. 278–298.

56 Lumpkin A. 2020. Metacognition and its contribution to student learning Introduction. *Coll Stud J*. 54:1–7.

57 Stanton JD, Sebesta AJ, Dunlosky J. 2021. Fostering metacognition to support student learning and performance. *CBE Life Sci Educ*. 20(2):fe3.

58 Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. 2010. Relating introspective accuracy to individual differences in brain structure. *Science*. 329(5998):1541–1543.

59 Fleming SM, Dolan RJ. 2012. The neural basis of metacognitive ability. *Philos Trans R Soc Lond B Biol Sci*. 367(1594):1338–1349.

60 Pescetelli N, Rees G, Bahrami B. 2016. The perceptual and social components of metacognition. *J Exp Psychol Gen*. 145(8):949–965.

61 Flavell JH. 1979. Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. *Am Psychol*. 34(10): 906–911.

62 Fleming SM. 2024. Metacognition and confidence: a review and synthesis. *Annu Rev Psychol*. 75(1):241–268.

63 Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*. 415(6870): 429–433.

64 Rohde M, van Dam LCJ, Ernst M. 2016. Statistically optimal multisensory cue integration: a practical tutorial. *Multisens Res*. 29(4–5):279–317.

65 Körding KP, *et al.* 2007. Causal inference in multisensory perception. *PLoS One*. 2(9):e943.

66 Shams L, Beierholm U. 2022. Bayesian causal inference: a unifying neuroscience theory. *Neurosci Biobehav Rev*. 137:104619.

67 Rhodes MG. 2016. Judgments of learning: methods, data, and theory. In: Dunlosky J, Tauber SK, editors. *The Oxford handbook of metamemory*. Oxford, United Kingdom: Oxford University Press. p. 65–80.

68 Rhodes MG, Tauber SK. 2011. The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychol Bull*. 137(1):131–148.

69 Metcalfe J, Finn B. 2008. Evidence that judgments of learning are causally related to study choice. *Psychon Bull Rev*. 15(1):174–179.

70 van Beers RJ, Baraduc P, Wolpert DM. 2002. Role of uncertainty in sensorimotor control. *Philos Trans R Soc Lond B Biol Sci*. 357(1424):1137–1145.

71 Schraw G. 2009. Measuring metacognitive judgments . In: Hacker DJ, Dunlosky J, Graesser AC, editors. *Handbook of metacognition in Education*. Boca Raton (FL): Routledge/Taylor & Francis Group. p. 415–429.

72 Schwarz N. 2015. Metacognition. In: Mikulincer M, Shaver PR, Borgida E, Bargh JA, editors. *APA Handbook of personality and social psychology, Vol. 1. attitudes and social cognition*. Washington (DC): American Psychological Association. p. 203–229.

73 Nelson TO. 1999. Cognition versus metacognition. In: *The nature of cognition*. Cambridge (MA): MIT Press. p. 625–642.

74 Maniscalco B, Lau H. 2014. Signal detection theory analysis of type 1 and type 2 data: meta-D′, response-specific meta-D′, and the unequal variance SDT model. *The cognitive neuroscience of metacognition*. Berlin, Germany: Springer. p. 25–66.

75 Mazancieux A, Fleming SM, Souchay C, Moulin CJA. 2020. Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *J Exp Psychol Gen*. 149(9):1788–1799.

76 Macmillan NA, Douglas Creelman C. 2005. *Detection theory: a user's guide. 2nd ed*. Mahwah (NJ): Erlbaum.

77 Green DM, Swets JA. 1966. *Signal detection theory and psychophysics*. New York (NY): Wiley.

78 Rahnev D. 12 May 2023. Measuring metacognition: A comprehensive assessment of current methods. PsyArXiv waz9h. https://doi.org/10.31234/osf.io/waz9h, preprint: not peer reviewed.

79 Maniscalco B, McCurdy LY, Odegaard B, Lau H. 2017. Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. *J Neurosci*. 37(5):1213–1224.

80 Maniscalco B, Lau H. 2015. Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neurosci Conscious*. 2015(1): niv002.

81 Miyamoto K, Setsuie R, Osada T, Miyashita Y. 2018. Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates. *Neuron*. 97(4): 980–89.e6.

82 Cai Y, *et al.* 2022. Time-sensitive prefrontal involvement in associating confidence with task performance illustrates metacognitive introspection in monkeys. *Commun Biol*. 5(1):799.

83 Stolyarova A, *et al.* 2019. Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nat Commun*. 10(1):4704.

84 Rahnev D. 2025. A comprehensive assessment of current methods for measuring metacognition. *Nat Commun*. 16(1):701.

85 Locke SM, Mamassian P, Landy MS. 2020. Performance monitoring for sensorimotor confidence: a visuomotor tracking study. *Cognition*. 205:104396.

86 Fassold ME, Locke SM, Landy MS. 2023. Feeling lucky? Prospective and retrospective cues for sensorimotor confidence. *PLoS Comput Biol*. 19(6):e1010740.

87 Johnson SGB, *et al.* 4 November 2024. Imagining and building wise machines: The centrality of AI metacognition. arXiv 02478. https://doi.org/10.48550/arXiv.2411.02478, preprint: not peer reviewed.

88 Tankelevitch L, *et al.* 2024. The metacognitive demands and opportunities of generative AI. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Vol. 57. New York (NY): ACM.

89 Ernst MO, Bülthoff HH. 2004. Merging the senses into a robust percept. *Trends Cogn Sci*. 8(4):162–169.

90 Alais D, Burr D. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol*. 14(3):257–262.

91 Ernst MO. 2010. Behavior. Decisions made better. *Science*. 329(5995):1022–1023.

92 Wahn B, Kingstone A, König P. 2018. Group benefits in joint perceptual tasks-a review: group benefits in joint perceptual tasks. *Ann N Y Acad Sci*. 1426(1):166–178.

93 Bang D, *et al.* 2017. Confidence matching in group decision-making. *Nat Hum Behav*. 1(6):1–7.

94 Bahrami B, *et al.* 2012. What failure in collective decision-making tells us about metacognition. *Philos Trans R Soc Lond B Biol Sci*. 367(1594):1350–1365.

95 Chi OH, Chi CG, Gursoy D, Nunkoo R. 2023. Customers' acceptance of artificially intelligent service robots: the influence of trust and culture. *Int J Inf Manage*. 70:102623.

96 Kaplan AD, Kessler TT, Brill JC, Hancock PA. 2023. Trust in artificial intelligence: meta-analytic findings. *Hum Factors*. 65(2): 337–359.

97 Mahmoodi A, *et al.* 2015. Equality bias impairs collective decision-making across cultures. *Proc Natl Acad Sci U S A*. 112(12):3835–3840.

98 Gupta RT, Mehta KA, Turkbey B, Verma S. 2020. PI-RADS: past, present, and future. *J Magn Reson Imaging*. 52(1):33–53.

99 Hyndman ME, Pavlovich CP, Eure G, Fradet V, Ghai S. 2018. PD37-06 prospective validation of Pri-Mus^{TM}, the prostate risk identification using micro-ultrasound protocol for real-time detection of prostate cancer using high-resolution micro-ultrasound imaging. *J Urol*. 199(4S):e733.

100 Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nat Commun*. 16(1):642.

101 Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M. 2023. Explainable artificial intelligence: evaluating the objective and subjective impacts of xAI on human-agent interaction. *Int J Hum-Comp Interact*. 39(7):1390–1404.

102 Hodge VJ, Hawkins R, Alexander R. 2020. Deep reinforcement learning for drone navigation using sensor data. *Neural Comput Appl*. 33(6):2015–2033. https://doi.org/10.1007/s00521-020-05097-x

103 Pokhrel N. Drone Obstacle Avoidance and Navigation Using Artificial Intelligence. 2018 [accessed 2024 Oct 22]. https://aaltodoc.aalto.fi/items/e4f158ad-a089-42a2-8067-0168070c573d.

104 Joshi A, Spilbergs A, Miķelsone E. 2024. AI-enabled drone autonomous navigation and decision making for defence security. *Proc Int Sci Pract Conf*. 4:138–143.

105 Xiong M, *et al.* 2023. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv 13063. https://doi.org/10.48550/arXiv.2306.13063, preprint: not peer reviewed.

106 Singh AK, Devkota S, Lamichhane B, Dhakal U, Dhakal C. 2023. The confidence-competence gap in large language models: A cognitive study. arXiv 16145. https://doi.org/10.48550/arXiv.2309.16145, preprint: not peer reviewed.

107 Zhou K, Hwang JD, Ren X, Sap M. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. arXiv 06730. https://doi.org/10.48550/arXiv.2401.06730, preprint: not peer reviewed.

108 Pavlovic J, *et al.* 2024. Generative AI as a metacognitive agent: A comparative mixed-method study with human participants on ICF-mimicking exam performance. arXiv 05285. https://doi.org/10.48550/arXiv.2405.05285, preprint: not peer reviewed.

109 Cash TN, Oppenheimer DM, Christie S, Devgan M. 2024. Quantifying UncertAInty: Testing the accuracy of LLMs' confidence judgments. PsyArXiv 47df5_v3. https://doi.org/10.31234/osf.io/47df5_v3, preprint: not peer reviewed.

110 Scholten F, Rebholz TR, Hütter M. 2024. Metacognitive myopia in large language models. arXiv 05568. https://doi.org/10.48550/arXiv.2408.05568, preprint: not peer reviewed.

111 Tao S, *et al.* 2024. When to trust LLMs: Aligning confidence with response quality. arXiv 17287. https://doi.org/10.48550/arXiv.2404.17287, preprint: not peer reviewed.

112 Yang H, Wang Y, Xu X, Zhang H, Bian Y. 2024. Can we trust LLMs? Mitigate overconfidence bias in LLMs through knowledge transfer. arXiv 16856. https://doi.org/10.48550/arXiv.2405.16856, preprint: not peer reviewed.

113 Steyvers M, *et al.* 2024. The calibration gap between model and human confidence in large language models. arXiv 13835. https://doi.org/10.48550/arXiv.2401.13835, preprint: not peer reviewed.

114 Zytek A, Pidò S, Veeramachaneni K. 2024. LLMs for XAI: Future directions for explaining explanations. arXiv 06064. https://doi.org/10.48550/arXiv.2405.06064, preprint: not peer reviewed.

115 Fischer H, Fleming S. 2024. Why metacognition matters in politically contested domains. *Trends Cogn Sci*. 28(9):783–785.