Vision Research 244 (2026) 108831

Contents lists available at ScienceDirect

Vision Research

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

Confidence-accuracy dissociations in perceptual decision making

Dobromir Rahnev

School of Psychology, Georgia Institute of Technology, Atlanta, GA, United States

ARTICLE INFO

Keywords:
Confidence
Perceptual decision making
Visual metacognition
Positive evidence bias
Variance neglect

ABSTRACT

Confidence in perceptual tasks typically follows accuracy, making it difficult to study the mechanisms specific to confidence computation. However, the last decade and a half has seen a proliferation of experimental manipulations that lead to confidence-accuracy dissociations in which one condition leads to higher confidence, but equal (or lower) accuracy compared to another condition. Such dissociations promise to reveal the underlying principles behind confidence ratings. Nevertheless, there is no consensus surrounding how to create confidence-accuracy dissociations or how to interpret their existence. This review first categorizes existing confidence-accuracy dissociations, highlighting that they can be produced by stimulus, task, or brain manipulations. It then examines the most common mechanistic explanations for why such dissociations occur. Finally, it explores several confusions and open questions regarding the use of the term "positive evidence," the relationship between confidence-accuracy dissociations and metacognitive noise, and the role of reaction time in such dissociations. Overall, this review highlights the fact that confidence-accuracy dissociations can be produced by multiple mechanisms and suggests that computational models of confidence should be evaluated based on their ability to predict new dissociations.

1. Introduction

Most experimental manipulations affect accuracy and confidence in a similar fashion. For example, increasing stimulus contrast or motion coherence leads to increasing both accuracy and confidence (Kiani et al., 2014; Shekhar & Rahnev, 2021b). Conversely, increasing noise decreases both accuracy and confidence (Rafiei et al., 2024). The strong empirical correspondence between accuracy and confidence is mirrored in models of confidence, which typically predict that accuracy and confidence change in tandem under most circumstances (Pleskac & Busemeyer, 2010; Rausch et al., 2021; Shekhar & Rahnev, 2024a). However, the tight relationship between accuracy and confidence makes it difficult to reveal the mechanisms that underlie confidence ratings.

To uncover the principles behind the confidence computation, a major focus of experimental research on confidence is finding ways to dissociate confidence from accuracy by discovering confidence-accuracy dissociations (Fleming, 2024; Kepecs & Mainen, 2012; Mamassian, 2016). A confidence-accuracy dissociation occurs when one condition leads to higher confidence, but equal (or lower) accuracy compared to another condition. For example, imagine that one is equally good at making apple-orange and apple-peach discriminations when faced with noisy images of fruits. If that person is consistently more confident for one of these judgments than the other (e.g., apple-orange over apple-peach), that would be an example of a confidence-accuracy dissociation. In practice, confidence-accuracy dissociations are rare and often produce only moderate effect sizes, making their systematic exploration especially important.

Early research on confidence focused on confidence calibration, that is, how well average confidence ratings in each condition reflect the average accuracy in that condition. This line of research has revealed many examples of both overconfidence (Adams, 1957; Baranski & Petrusic, 1994; Dawes, 1980; Keren, 1988; Koriat, 2011) and underconfidence (Baranski & Petrusic, 1994; Björkman et al., 1993; Dawes, 1980). However, while such findings are informative, it is difficult to use them to reveal the mechanisms that underlie confidence because they can often be explained by something as simple as assuming an overall bias towards giving high or low confidence ratings.

Perhaps the first example of a confidence-accuracy dissociation was reported by Tulving (1981) in the context of recognition memory. Tulving asked observers to study a series of images. At test, participants saw two images and had to pick which one they studied. In one condition, the distractor image was very similar to the target image. In another condition, the distractor image was very similar to another studied item (which was not presented during test). Tulving showed that participants were more accurate but less confident in the first condition, a finding that has been conceptually replicated numerous times since

E-mail address: rahnev@psych.gatech.edu.

https://doi.org/10.1016/j.visres.2026.108831
Received 20 November 2025; Received in revised form 31 March 2026; Accepted 8 April 2026
Available online 16 April 2026
0042-6989/© 2026 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

then (Heathcote et al., 2009). However, this finding was of "secondary interest" (p.494) for Tulving, reflecting the prevailing attitude at the time that treated confidence as a "Cinderella" variable – useful for understanding other topics instead of being a worthy subject in its own right (Vickers, 2001).

Subsequent research on confidence-accuracy dissociations continued to focus on memory, cognitive and social judgments (Busey et al., 2000; Wells & Olson, 2003; Zarnoth & Sniezek, 1997). A major focus of this research was the concept of fluency – typically defined as the subjective experience of ease associated with cognitive processing – with many studies suggesting that fluency can increase confidence even if it doesn't improve accuracy. For example, participants report higher confidence that they will remember words displayed in larger font sizes even though font size has virtually no effect on recall accuracy, an effect theorized to be driven by fluency (Rhodes & Castel, 2008). In a different study, Gill et al. (1998) demonstrated that richer mental representations of other people (thought to increase the sense of fluency) increase participants' confidence in their prediction as to how these people would respond to a sexual history and self-concept questionnaire but did not increase the accuracy of the predictions.

Modern research on confidence-accuracy dissociations in perception began to emerge in the 2000s and exploded in the early 2010s. In the last 15 years, there have been well over 50 studies reporting confidence-accuracy dissociations in perceptual decision-making tasks. Various studies have used different paradigms, provided different explanations for their effects, and have sometimes contradicted each other. Further, despite many prominent reviews on confidence in perceptual tasks (Fleming, 2024; Mamassian, 2016; Miyoshi et al., 2024; Peters, 2022; Rahnev, 2021b), no review has thus far focused on summarizing the literature on confidence-accuracy dissociations. This paper seeks to review, organize, and make sense of this burgeoning literature. We start by organizing the many reports of confidence-accuracy dissociations (Section 2), then critically examine the most prominent theories advanced to explain these dissociations (Section 3) and then discuss several confusions and open questions (Section 4). Finally, we conclude by highlighting the most important lessons and critical steps for future research (Section 5).

2. Types of confidence-accuracy dissociations

One of the difficulties of studying confidence-accuracy dissociations is that virtually every paper has a different way of inducing these dissociations. Yet, we lack a clear organization of the different manipulations, which makes it hard for each new paper to appropriately discuss how it relates to the previous literature. Below, we organize confidence-accuracy dissociations into three main categories: (1) stimulus manipulations, (2) task manipulations, and (3) brain manipulations. Within these broad categories, we distinguish six types of dissociations: (1) overall-strength manipulations, (2) attention manipulation and peripheral presentation, (3) multiple choice alternatives, (4) context effects, (5) cognitive manipulations, and (6) brain stimulation (Fig. 1).

To be clear, these categories should be seen as provisional. Subsequent research may produce new categories or motivate novel ways of organizing the existing confidence-accuracy dissociations. In fact, there are already numerous findings that do not fit neatly in the categories above. For example, confidence-accuracy dissociations have been reported to arise from factors such as the timing of stimulus presentation relative to a physiological process (e.g., saccades, Navajas et al., 2014, or the pre-stimulus alpha power in visual cortex, Samaha et al., 2017, 2022), unconscious presentation of relevant information (Vlassova et al., 2014), subthreshold motor activations (Gajdos et al., 2019), the physical effort necessary to indicate one's decision (Turner et al., 2021), whether participants were judging Gabor patches or dot motion (Xue et al., 2026), or whether trials included a "reverse pulse" in a dot motion task (Kiani et al., 2014). Currently, most of these findings are based on just one or a few papers and have not always received independent replication. However, additional work that follows up on these studies can certainly motivate the addition of new categories in Fig. 1.

2.1. Overall-strength manipulations

By far the largest set of studies that report confidence-accuracy dissociations use manipulations that can be broadly classified as "overall-strength manipulations." An overall-strength manipulation increases the overall stimulus strength in a choice-independent manner. The typical finding for this manipulation is that participants exhibit higher confidence for stimuli with high overall strength even when accuracy is matched for high- vs. low-strength stimuli. Note, however, that the literature on these manipulations rarely uses the term "overall-strength manipulations." Instead, different studies refer to their manipulations as "positive evidence" (Koizumi et al., 2015), "volatility" (Zylberberg et al., 2016), "mean vs. variance" (de Gardelle & Mamassian, 2015), or "energy" (Gao et al., 2025; Shekhar & Rahnev, 2024b). To add to the confusion, some studies in this group report seemingly opposite findings from others. Here, we review this literature and briefly discuss possible reasons for discrepancies in previous findings; we return to the issue of how this manipulation should be named in Section 4.1.

The most common type of overall-strength manipulations involves

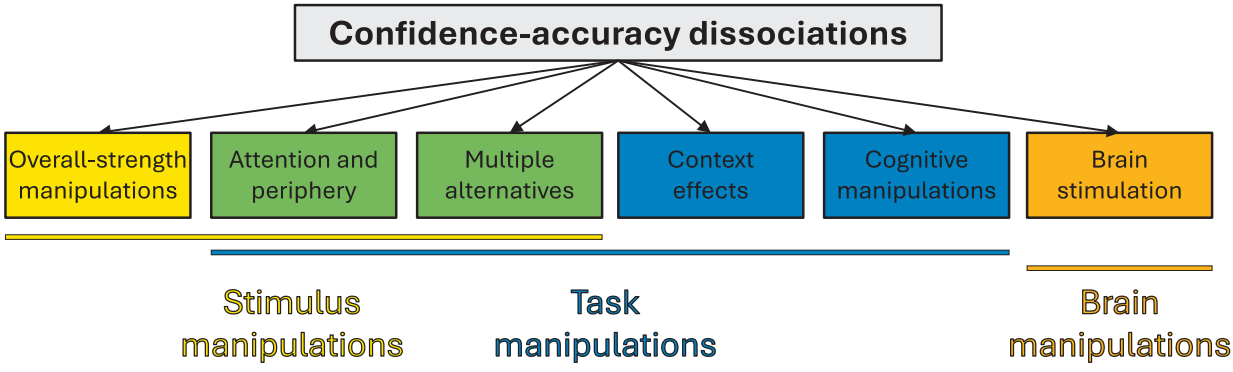

Fig. 1. Types of confidence-accuracy dissociations. Confidence-accuracy dissociations can be organized in three main categories: (1) stimulus manipulations, (2) task manipulations, and (3) brain manipulations. Within these three broad categories, one can further identify six sub-categories. Specifically, manipulations related to overall stimulus strength, attention, peripheral presentation, and multiple choice alternatives all typically affect the stimulus in some way. Similarly, manipulations related to attention, peripheral presentation, multiple choice alternatives, task context, and cognitive factors typically affect the task in some way. Note that manipulations of attention and peripheral presentation and manipulations of the number of choice alternatives can be categorized as both stimulus and task manipulations (hence, they are colored in green – a combination of yellow and blue – in the figure). Finally, brain stimulation is the only current example of direct brain manipulations that can induce confidence-accuracy dissociations.

increasing both the mean signal and the variance across an ensemble of items. For example, Zylberberg et al. (2016) manipulated overall stimulus strength by increasing the moment-by-moment volatility in the direction of motion for a random dot kinematogram (RDK). The researchers found that increased volatility increased confidence and lowered reaction time (RT) in the absence of accuracy change. A conceptually similar manipulation was introduced by Herce Castañón et al. (2019) who manipulated overall stimulus strength by increasing both the orientation variance and the contrast for an array of Gabor patches. Participants, who had to judge the average orientation of the Gabor patches, again exhibited higher confidence in the high-strength condition even though accuracy was matched between the high- and low-strength conditions. Multiple other papers have used the same conceptual trick – increasing overall stimulus strength by increasing the mean and variance of the stimulus – and have typically found higher confidence even in the absence of accuracy changes (Lee, Yabuki, et al., 2023; Samaha et al., 2016, 2019; Samaha & Denison, 2022; Xue et al., 2026; Zylberberg et al., 2014; but see de Gardelle & Mamassian, 2015). Further, the same manipulation has been demonstrated to produce confidence-accuracy dissociations not only in standard two-choice discrimination tasks but also in estimation tasks (Samaha et al., 2016, 2019).

A second way of increasing overall stimulus strength is to directly increase the strength associated with each alternative. For example, Koizumi et al. (2015) overlaid left- and right-tilted gratings and asked participants to judge which orientation was associated with higher contrast. In the high-strength condition, both gratings had high contrast, while in the low-strength condition both gratings had low contrast. While accuracy was matched across the two conditions, confidence was higher for the high-strength stimuli. A conceptually similar manipulation was employed by Odegaard, Chang, et al. (2018) who designed RDKs with coherent motion to both left and right. Participants had higher confidence in the high-strength condition where both left and right motion had high coherence compared to an accuracy-matched low-strength condition defined by low coherences for both motion directions. These results have been replicated multiple times (Gao et al., 2025; Rollwage et al., 2020; Song et al., 2015) and have been extended to multi-alternative tasks in the context of numerosity judgments (Xue et al., 2024).

Interestingly, several studies have found confidence-accuracy dissociations that seemingly exhibit a reversed overall-strength effect such that the high-strength stimuli were judged with low confidence. For example, Boldt et al. (2017, 2019) and Desender et al. (2018) used an array of colored patches with participants' task being to judge whether the overall color was more blue or more red. They increased the overall stimulus strength by increasing both the variance and the mean of the display but found that the high-strength stimuli were judged with lower confidence. One possible explanation for this effect is that it arises from "robust averaging" – a phenomenon where humans down-weight highly atypical evidence samples (de Gardelle & Summerfield, 2011). However, recent modeling work has suggested that robust averaging may not fully explain the observed confidence-accuracy dissociation; to explain the dissociation, the authors postulated the presence of metacognitive noise that differs between the low- and high-strength conditions (Boldt et al., 2025).

Another violation of the overall-strength effect was reported by Spence et al. (2016) who used RDK displays where all dots were moving coherently in a generally upwards direction. Participants judged whether the mean direction of motion was tilted slightly to the left or right of vertical. Spence et al. found that increasing both the range of motion and the mean angle of motion direction (a type of overall-strength manipulation) led to lower confidence but matched accuracy. It remains unclear why this particular manipulation produces a reversed overall-strength effect. One speculative possibility is that the stimuli used by Spence et al. produce a nonlinear mapping to internal evidence, such that the evidence for each alternative is actually lower in their high-strength condition. Future research is needed to replicate Spence et al.'s results and explore exact mechanistic explanations.

Overall, despite several notable exceptions, the great majority of studies have found support for the notion that overall-strength manipulations, which increase the overall stimulus strength in a choice-independent manner, increase confidence even for matched accuracy. Critically, this effect replicates across many types of stimuli (RDKs, arrays of Gabor patches, composite gratings, etc.) and types of manipulations (increasing both the mean and the variance vs. increasing the strength of each alternative).

2.2. Manipulations of attention and peripheral stimulus presentation

Confidence-accuracy dissociations have been extensively studied in the context of attention and peripheral stimulus presentation. Manipulations of attention and peripheral presentation can be seen as hybrid stimulus and task manipulations. Specifically, these manipulations (i) alter the sensory evidence associated with the stimuli and (ii) allow participants to adjust their decision strategies. Both manipulations have also led to mixed results.

Early research on attention seemed to suggest that attention increases accuracy more than confidence. For example, Wilimzig et al. (2008) had participants discriminate the orientation of briefly presented Gabor patches. Critically, each Gabor patch was either attended or unattended due to the presentation of a pre-stimulus cue that indicated the stimulus that participants would most likely be asked to respond to. The authors found that attention significantly increased accuracy but had no effect on confidence, thus resulting in a confidence-accuracy dissociation. A similar effect was subsequently reported using visibility instead of confidence judgments (Rahnev et al., 2011) and by using pre-stimulus activity in the dorsal attention network as a proxy of attention (Rahnev, Bahdo, et al., 2012). Consistent with the notion that attention increases accuracy more than confidence, other research found that prior information was down-weighted for attended compared to unattended stimuli (Morales et al., 2015). However, two other studies reported the opposite effect such that attention increased confidence more than accuracy (Kurtz et al., 2017; Zizlsperger et al., 2012). An opposite effect was also found in the context of temporal attention using the attentional blink paradigm: lag-1 targets had much higher accuracy, but matched confidence compared to lags-2 and lag-3 targets (Recht et al., 2019). Finally, other research has suggested that attention may have a similar effect on accuracy and confidence (Denison et al., 2018). A large recent adversarial collaboration concluded that attention increases the visibility of a stimulus more than accuracy but did not examine confidence ratings (Tian et al., 2025).

The findings on attention are largely paralleled when considering manipulations of peripheral stimulus presentation. Early research demonstrated that people have more liberal detection biases in the periphery, which may explain why people have an impression of a rich percept across the visual field (Li et al., 2018; Solovey et al., 2014). Follow-up research confirmed these findings and further demonstrated that participants exhibit overconfidence for crowded vs. uncrowded stimuli in the periphery, a finding that was seen to imply peripheral overconfidence (Odegaard, Chang, et al., 2018). However, later studies found no dissociation between accuracy and confidence across different peripheral eccentricities between 10° and 40° (Pruitt et al., 2024) and one study even found underconfidence in the periphery (Toscani et al., 2021). Research on this topic is ongoing with the most recent study showing peripheral overconfidence in a scene categorization task that was selected to reflect the type of processing that peripheral vision has evolved for (Sharvashidze et al., 2025).

Overall, it is clear that both attention and peripheral stimulus presentation strongly influence both accuracy and confidence. Further, converging evidence suggests that both inattention and peripheral stimulus presentation are associated with more liberal detection biases (Li et al., 2018; Rahnev et al., 2011; Wyart et al., 2012). However,

whether attention and peripheral stimulus presentation lead to over- or under-confidence remains hotly debated, with the answer possibly depending on the specifics of the stimuli and task. It also remains an open question whether these two effects are linked as often hypothesized (Odegaard, Chang, et al., 2018; Solovey et al., 2014; Winter & Peters, 2021).

2.3. Multiple alternatives

Most research on perceptual confidence focuses on two-choice tasks. However, naturalistic decisions typically involve multiple alternatives (Rahnev, 2020) and there is consensus about the need to investigate confidence in multi-alternative perceptual decision making (Rahnev et al., 2022). Correspondingly, a growing number of studies have compared how confidence and accuracy are affected by the number and identity of the available alternatives, with several studies reporting confidence-accuracy dissociations.

Baldassi et al. (2006) conducted the first study that reported a confidence-accuracy dissociation driven by the number of available alternatives. Participants saw between 1 and 16 Gabor patches with one of them being the target and the remaining being distractors. The target was tilted either left or right, while all distractors were vertical. Participants had to indicate the direction of tilt of the target and provide a confidence rating. Baldassi et al. demonstrated that including a larger number of distractors reduced the accuracy of the decision but increased confidence on error trials. The results were explained by a simple model that assumes variability in the orientation perception of each Gabor patch leading to more errors and more biased percepts for larger displays, which in turn lead to higher confidence for error trials.

More recent research has focused on tasks where participants indicate which of several alternatives is the largest in terms of area or number of dots. In one study, participants judged the shape with the largest area among either two or three alternatives (Comay et al., 2023). Critically, in the 3-choice conditions, one alternative was obviously wrong (i.e., its area was much smaller than the other two) and was almost never selected. Yet, despite matched accuracy between the 2- and 3-choice conditions, participants exhibited higher confidence in the 3-choice condition. This effect only occurred when the first two shapes were of similar area (i.e., the task was difficult) and did not extend to an alternative design where participants judged which cloud of dots a target belongs to (Li & Ma, 2020). Another study reported a similar effect in a dot numerosity task, such that adding a third option that was almost never selected increased confidence without affecting accuracy (Miyoshi & Sakamoto, 2025).

While both studies above suggest that the presence of more alternatives leads to higher confidence, another study showed that this is not always the case (Xue et al., 2024). The researchers included 3- and 5-choice tasks where participants judged which color had the most dots. They found that 3-choice displays (e.g., [100,80,60] or [100,70,70]) produced lower accuracy but higher confidence than 5-choice conditions where the third option was substituted with three unlikely options (e.g., [100,80,30,30,30] or [100,70,35,35,35]). Together, this literature suggests that the effect of the number of alternatives may depend on the specifics of the experimental design and demonstrates that a comprehensive picture of how confidence is computed requires moving beyond two-choice tasks.

Finally, while most confidence-accuracy dissociations for multi-alternative tasks have been elicited via manipulations of the number of alternatives, at least one study found a dissociation based on the relative evidence for each alternative. Specifically, the study showed that conditions where the two incorrect options had similar numerosity (e.g., [98,72,72]) were associated with higher accuracy but lower confidence compared to conditions where the two incorrect options had dissimilar numerosity (e.g., [98,84,48]) (Xue et al., 2024). This can be categorized as a pure stimulus manipulation (as opposed to manipulations of the number of choices that constitute joint stimulus and task manipulations), making it particularly exciting as it cannot be explained by high-level cognitive strategies related to the task.

2.4. Context effects: Confidence leak and criterion attraction

The manipulations thus far always affected the stimulus in some way. We now turn to two manipulations that only affect the task. The first class of such manipulations consists of manipulations that vary the context in which confidence judgments arise. The most common approach is to consider the context provided by the confidence judgment on a previous trial. Rahnev et al. (2015) demonstrated that the confidence on the previous trial strongly influences confidence on the current trial, a phenomenon they called "confidence leak." This effect occurred even when the task on the previous trial was completely different from the task on the current trial. Critically, Rahnev et al. showed that making the previous trial easier led to higher confidence on the current trial in the absence of a change in accuracy on the current trial—a type of confidence-accuracy dissociation. Follow-up research demonstrated that this effect occurs even if the prior task does not require participants to report a confidence rating (Aguilar-Lleyda et al., 2021). Other work on confidence leak has shown that the phenomenon occurs even across perceptual and memory tasks (Kantner et al., 2019), that it can be used to predict confidence across cognitive domains (Mei et al., 2023), and that it interacts with other factors such as choice serial dependence and the motor action required to give the confidence response (Bocheva et al., 2025; Bocheva & Rahnev, 2025; Law et al., 2025; Locke et al., 2018). Further, Spence et al. (2018) demonstrated that a leakage of confidence can occur even within the same trial. In their study, participants had higher confidence when judging one feature of a stimulus (its brightness) when another feature (direction of motion) was easier to discriminate. This body of work demonstrates the robustness of the confidence leak phenomenon and its ability to produce confidence-accuracy dissociations by influencing confidence independent of accuracy.

Another context effect that induces robust confidence-accuracy dissociations is based on the phenomenon of "criterion attraction" – an effect where interleaving trials from two different conditions leads to the decision criteria across the two conditions becoming attracted to each other (Gorea & Sagi, 2000, 2001; Zak et al., 2012). In the context of confidence, a researcher can interleave two tasks that optimally require confidence criteria located either close to or far from the decision criterion. Doing so does not change the accuracy of the judgments in either task but shifts the confidence criteria toward each other, thus resulting in lower confidence in the first task and higher confidence in the second task. Empirical work demonstrates that this manipulation results in a large confidence difference even when both accuracy and RT are matched (Rahnev, 2021a). Together, these results suggest that context effects are a robust way to induce confidence-accuracy dissociations.

2.5. Cognitive manipulations

Another class of confidence-accuracy dissociations based purely on task manipulations is a broad category we refer to as "cognitive manipulations." Studies from this class do not manipulate the stimulus or the trial-to-trial context but instead investigate whether instructions and manipulating participants' beliefs can affect confidence but not accuracy (Boldt et al., 2019; Desender et al., 2025; Van Marcke et al., 2024). For example, Van Marcke et al. (2024) randomly assigned participants to receive feedback informing them that they performed either better or worse than other participants. This feedback was received during a training phase, and its effects causally affected confidence (but not accuracy) in a subsequent testing phase. Further, Van Marcke et al. obtained equivalent results if instead of false feedback, they made the training phase harder for some participants and easier for others. Similar results were found in an experiment where people learned to associate different cues with task difficulty even though this association was never

made explicit by the researchers (Boldt et al., 2019). Critically, after an unexpected task difficulty reversal, participants' confidence was still influenced by the cues even though their accuracy was not.

Beyond feedback or cue manipulations, Sepulveda et al. (2020) used goal framing to shift how perceptual evidence is used to make perceptual decisions and confidence judgments. In a two-choice dot discrimination task, participants were instructed to choose the circle with more vs. fewer dots. These different task framings led participants to weight evidence differently, altering their confidence in ways not fully predictable from stimulus strength alone. A similar finding was obtained by Sakamoto & Miyoshi (2024). In addition, confidence-accuracy dissociations have been reported as arising from instructions to emphasize speed vs. accuracy (Baranski & Petrusic, 1994). Overall, this line of research demonstrates that human confidence is ultimately determined by cognitive processing and therefore confidence-accuracy dissociations can be produced by not only affecting the stimuli or context, but also by purely cognitive manipulations.

2.6. Brain stimulation

Beyond using various stimulus and task manipulations, confidence-accuracy dissociations can be induced by direct brain changes to brain function. An early study demonstrated that single-pulse transcranial magnetic stimulation (TMS) delivered to the occipital cortex reduced performance (d') but increased confidence (Rahnev, Maniscalco, et al., 2012). A conceptually similar effect was obtained by another study that used microstimulation of the motion-sensitive area MT/MST in monkeys, finding that brain stimulation of sensory cortex reduces accuracy but increases confidence (Fetsch et al., 2014). Both papers argued that the effect occurs by inducing noise in the visual representations (see Section 3.2). Instead of targeting the sensory cortex, another paper used TMS to target premotor cortex and found that stimulation of the motor representation associated with the unchosen response reduced confidence in correct responses (Fleming et al., 2015), thus suggesting that confidence may incorporate action-specific information.

Other studies have targeted the parts of the prefrontal cortex, where confidence computation is presumed to take place (Fleming et al., 2012; Morales et al., 2018; Yeon et al., 2020). Shekhar & Rahnev (2018) targeted the dorsolateral prefrontal cortex (DLPFC) using an inhibitory train of three pulses and observed that the stimulation did not affect accuracy but decreased confidence. Interestingly, targeting the same site with single-pulse TMS – presumed to have an excitatory effect on neural circuits – led to the opposite dissociation with no change of accuracy but increased confidence (Xue et al., 2023). Further, this latter effect

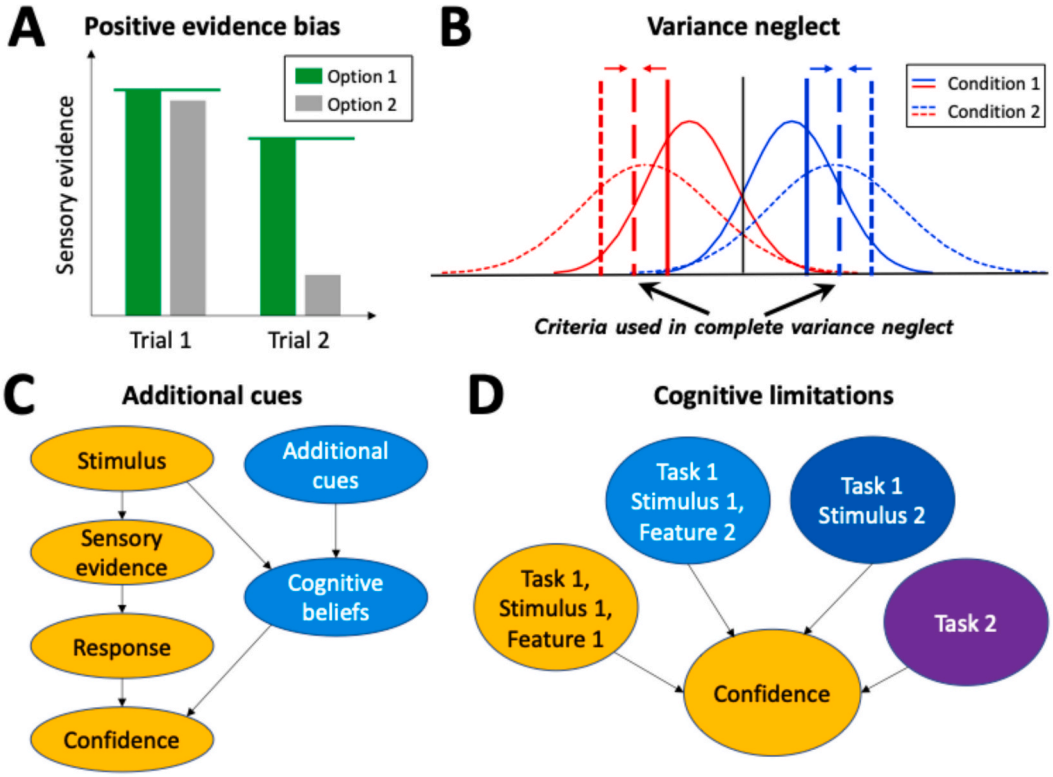

Fig. 2. Explanations for confidence-accuracy dissociations. The figure illustrates four of the most prominent explanations for confidence-accuracy dissociations. (A) The positive evidence computation. Internal sensory evidence for each of two options is depicted for two separate trials. Option 1 is associated with higher evidence and is therefore chosen in both cases. According to the positive evidence computation, confidence is based exclusively on the evidence for the chosen option. Therefore, Trial 1 is given higher confidence even though it is less likely to be accurate. (B) Variance neglect. Condition 1 gives rise to distributions of internal evidence with relatively small distance between the means and standard deviations. In Condition 2, both of these quantities are scaled by the same number, thus resulting in matched d' between the two conditions. These two conditions would optimally be judged using separate sets of criteria. However, in the presence of complete variance neglect, a single set of criteria is used for both conditions, resulting in large confidence differences. Note that even partial variance neglect would still produce a confidence-accuracy dissociation as long as the two sets of criteria move toward each other. (C) Additional cues. Most models of visual metacognition postulate that a stimulus gives rise to internal sensory evidence, which is used to make a perceptual response and give a confidence rating (orange ovals). However, additional cues (e.g., font size or amount of effort expended in giving a response) can lead to cognitive beliefs about task performance that influence the confidence rating (blue ovals). Note that stimulus characteristics (e.g., subjective impression of stimulus visibility) could also influence these cognitive beliefs. (D) Cognitive limitations. Confidence for a given decision should be based on the relevant task, stimulus, and feature (orange ovals). However, another feature of the same stimulus (light blue oval), another stimulus (dark blue oval), or another task (purple oval) can influence this confidence because of participants' inability to inhibit the influence of these extraneous factors.

occurred for a range of TMS timings (between 200 and 500 ms after stimulus onset), suggesting that confidence computation in the prefrontal cortex starts very soon after stimulus presentation. In contrast to these papers where TMS was delivered "online" (i.e., during each trial), other studies have used "offline" TMS protocols where an area is temporarily suppressed by repetitive TMS. When targeting the prefrontal cortex, these studies typically did not observe effects on either accuracy or confidence, though the trial-by-trial confidence-accuracy correlation was often affected (Lapate et al., 2020; Rahnev et al., 2016; Rounis et al., 2010). Finally, a recent study extended the work on TMS by demonstrating a confidence-accuracy dissociation using transcranial direct current stimulation (Saccenti et al., 2025).

3. The mechanisms behind confidence-accuracy dissociations

Having reviewed the empirical findings of confidence-accuracy dissociations, we now turn to the question of why these dissociations occur. There have been numerous explanations provided for different dissociations (Fig. 2). Some of these explanations apply to several of the six manipulation categories discussed in Section 2, whereas other explanations apply to just one of the manipulations. Critically, many of these explanations remain somewhat speculative and controversial. Equally important, different confidence-accuracy dissociations may be produced by different underlying mechanisms.

3.1. Positive evidence (i.e., choice-congruent) computation

Perhaps the most widely known explanation for confidence-accuracy dissociations is the notion of a "positive evidence" computation (Maniscalco et al., 2016, 2021; Peters et al., 2017; Zylberberg et al., 2012). This explanation, sometimes referred to as choice-congruent or decision-congruent computation, casts confidence as based solely on the strength of evidence supporting the choice made by the participant while ignoring the strength of evidence for other alternatives (Fig. 2**A**). The positive evidence computation was first proposed by Zylberberg et al. (2012) who used reverse correlation to examine how time-resolved sensory evidence for each of two options influences both choice and confidence. They found that perceptual choices were influenced by evidence for both alternatives, but confidence was influenced only by the strength of choice-congruent evidence. This finding has been recently replicated for discrimination tasks, though an opposite "negative evidence bias" was observed when reverse correlation analyses were applied to detection tasks (Mazor et al., 2023). However, one difficulty with this line of research is that the reverse correlation analyses function at the level of the stimulus. However, it is generally unknown how external sensory information is transformed into internal decision evidence (Green et al., 2026; Mazor et al., 2023). Further, reverse correlation has been shown to reflect both sensory and decision-making processes (Okazawa et al., 2018), which means that it is difficult to infer the computations performed based on internal sensory evidence by only considering stimulus fluctuations.

In fact, the studies using reverse correlation are not the main reason why the positive evidence computation is so prominent as an explanation for confidence-accuracy dissociations. The main reason is that the idea that confidence is exclusively based on evidence for the chosen alternative motivated the most successful and widespread method for inducing confidence-accuracy dissociation – the overall-strength manipulations. The idea is simple: if confidence only reflects the choice-congruent evidence then a confidence-accuracy dissociation should be achieved by increasing the sensory evidence for both alternatives in a two-choice task. Indeed, since both alternatives will be associated with higher evidence, then no matter which alternative is chosen, a confidence computation that only considers the evidence for the chosen alternative would predict that confidence should increase for high-strength stimuli. And indeed, as reviewed in Section 2.1, dozens of studies have confirmed this prediction.

The link between the stimulus manipulation and the presumed internal computation is considered so strong that the same name is often used for both: "positive evidence." In fact, when that phrase is used in the literature, it is sometimes impossible to say whether the authors are referring to the stimulus manipulation (which we call "overall-strength manipulation" in this paper) or to internal confidence computation (which we call "positive evidence" or "choice-congruent" computation). (We return to this issue in Section 4.1 and argue that the stimulus manipulation and confidence computation should have different names to avoid the current confusion.)

Importantly, the success of the overall-strength manipulations in inducing confidence-accuracy dissociations does not directly imply the existence of a choice-congruent bias in the confidence computation. It could very well be that overall-strength manipulations induce confidence-accuracy dissociations via a completely different mechanism. And, in fact, numerous papers have now demonstrated that qualitative effects initially thought to support the positive evidence computation can be explained via alternative mechanisms (Khalvati et al., 2021; Mazor et al., 2023; Miyoshi & Lau, 2020; Shekhar & Rahnev, 2024b; Webb et al., 2023; Xue et al., 2024).

Ideally, support for the positive evidence computation would be provided in the form of model comparison. When this is done, most studies have found that the positive evidence computation provides a worse explanation of the confidence data compared to alternative models. For example, out of 13 models examined by Shekhar & Rahnev, 2024a, the positive evidence model ranked 12th, 11th, and 2nd across three experiments. The large difference in ranking in these three experiments likely stems from the fact that the first two featured either no difficulty manipulation or a difficulty manipulation that induces the standard "folded-X pattern" where easier conditions lead to increased confidence for correct trials but decreased confidence for error trials (Sanders et al., 2016). However, the third experiment featured a double-increase pattern where easier conditions lead to increased confidence for both correct and error trials. Most existing models always predict a folded-X pattern, while the positive evidence model always predicts the double-increase pattern, thus explaining the large discrepancy in the model's rankings across the three experiments. Other work that only examined manipulations inducing a double-increase pattern also found relatively favorable rankings for the positive evidence model (Rausch et al., 2020, 2021). However, since many manipulations induce the folded-X pattern (Fung et al., 2025; Miyoshi & Sakamoto, 2025; Sanders et al., 2016; Xue et al., 2026), the fact that the positive evidence computation always predicts a double-increase pattern is problematic.

Further, the positive evidence computation makes counter-intuitive predictions that have been disconfirmed empirically. For example, consider an experiment where participants decide which of two circles has more dots inside of it. In one condition, one circle always has 100 dots, and the other has 80 dots. In a second condition, the first circle again has 100 dots, but the second one has 99 dots. The [100, 99] condition is much harder, and one may intuitively expect substantially lower accuracy and confidence on that task, and this type of effect is indeed observed empirically (Xue et al., 2024). Yet, according to the positive evidence computation, the [100, 99] condition should exhibit higher confidence because, on average, the larger activation in the [100, 99] condition is higher than in the [100, 80] condition. To make this concrete, we ran simple simulations, which assume that internal evidence is sampled independently for each circle from a Gaussian distribution with a mean equal to the true number of dots and a standard deviation fixed at 15. These values result in a very large difference in accuracy between the [100, 80] and [100, 99] conditions: 83% vs. 52%. If high confidence is given whenever the chosen alternative has activation above 110, then the average confidence is reversed compared to the accuracy: high confidence in the [100, 80] and [100, 99] conditions is given for 27% vs. 43% of trials. Note that the same qualitative result would be obtained for different values of the standard deviation or the high-confidence threshold. The reason why the positive evidence

computation makes the wrong prediction here is that it ignores the fact that the losing activation is closer to the winning one in the [100, 99] compared to the [100, 80] condition. Indeed, a computation based on the difference between the winning and losing activation produces the correct qualitative result: high confidence (produced using a cutoff of 15 for the difference between the two activations) in the [100, 80] and [100, 99] conditions is given for 64% vs. 48% of trials.

Overall, despite its popularity, the positive evidence computation is not supported by the majority of the empirical evidence. It remains possible that confidence is based exclusively on choice-congruent evidence in some specific circumstances, though no clear example has emerged yet. Another possibility is that confidence overweighs the choice-congruent evidence while still considering choice-incongruent evidence, though at least one paper did not support this possibility (Shekhar & Rahnev, 2024a). Therefore, whether any confidence-accuracy dissociations are due to a positive evidence computation that only considers choice-congruent information remains an open question.

3.2. Variance neglect

An alternative explanation for many confidence-accuracy dissociations is variance neglect: the idea that some manipulations increase the variability of internal activations and the decision system does not sufficiently account for this increase. The concept of variance neglect is represented graphically within a signal detection theoretic perspective in Fig. 2**B**. The figure depicts two conditions that are matched on accuracy but differ on confidence. In Condition 1, the two internal distributions corresponding to the two stimulus categories have a given standard deviation and distance between their means. In Condition 2, both the standard deviation and the distance between the means are multiplied by the same factor, resulting in the same sensitivity. If a participant in an experiment is unaware of this difference in variability, they would use the same confidence criteria in both conditions. Critically, such a strategy leads to higher confidence in the condition with higher internal variability (Fig. 2**B**).

The variance neglect explanation was first proposed by Rahnev et al. (2011) in the context of a study that examined the effects of spatial attention on detection biases and visibility ratings. Essentially the same idea was then used to explain confidence-accuracy dissociations stemming from transcranial magnetic stimulation applied to the visual cortex (Rahnev, Maniscalco, et al., 2012), pre-stimulus activity in dorsal attention network on confidence (Rahnev, Bahdo, et al., 2012), microstimulation in monkeys (Fetsch et al., 2014), variance in the orientations of an ensemble of tilted lines (Zylberberg et al., 2014), and frame-to-frame volatility in an RDK (Zylberberg et al., 2016). A similar conceptual explanation of variance neglect has more recently been applied to variability at the decision instead of the sensory level (Herce Castañón et al., 2019).

Much of this early work advocated for complete variance neglect such that the confidence criteria are identical for conditions with different internal variability. This strong position was later criticized by researchers who demonstrated that human participants have the capacity to at least partially take into account their internal variability (Denison et al., 2018; Lee, Denison, et al., 2023). However, complete insensitivity to internal variability is not needed for variance neglect to lead to confidence-accuracy dissociations. Later work used external noise to better characterize how well people account for differences in variance across conditions (Rahnev, 2021a). This work demonstrated that people are indeed sensitive to large differences between conditions but do not fully adjust for these differences. This work argued that the ideal confidence criteria in the two conditions experience "criterion attraction" even if they do not converge onto the same unified criterion. Critically, the "variance neglect" explanation for confidence-accuracy dissociations remains the same whether people use the same criterion across conditions or different criteria that are attracted to each other.

The positive evidence computation and variance neglect explanations have sometimes been invoked to explain the same confidence-accuracy dissociations (Gao et al., 2025; Odegaard, Grimaldi, et al., 2018). However, directly distinguishing between them is difficult because we do not have direct access to the internal sensory evidence used by humans – in most cases we can only compare models that make different assumptions about this evidence. Recently, Shekhar & Rahnev (2024b) tackled this problem by testing whether artificial neural networks (ANNs) spontaneously exhibit similar confidence-accuracy dissociations as humans. The researchers tested the performance of three different ANN architectures for three distinct overall-strength manipulations. They found that all ANNs exhibited human-like confidence-accuracy dissociations for all three overall-strength manipulations. Critically, Shekhar & Rahnev examined the internal sensory evidence upon which the ANN decisions were made and demonstrated that the confidence-accuracy dissociations emerged from variance neglect. Specifically, the high-strength stimuli produced internal distributions with higher variance, which then translated to higher confidence even when accuracy was matched. This study strongly suggests that confidence-accuracy dissociations produced by overall-strength manipulations in humans are also likely to be produced by a variance neglect mechanism. In fact, variance neglect could explain many of the existing categories of dissociations, including overall-strength manipulations (Shekhar & Rahnev, 2024b), attention (Rahnev, Bahdo, et al., 2012), peripheral stimulus presentation (Odegaard, Chang, et al., 2018), and brain stimulation to the sensory cortex (Fetsch et al., 2014; Rahnev, Maniscalco, et al., 2012). We consider variance neglect to be the most promising explanation for the majority of existing confidence-accuracy dissociations.

3.3. Using additional cues for confidence

Both the positive evidence computation and variance neglect explain confidence-accuracy dissociations as arising from differences in how the same internal sensory evidence is used to make choices vs. provide confidence. A qualitatively different explanation for confidence-accuracy dissociations is that they may arise from a process where confidence judgments are at least partially based on additional cues not derived from the internal sensory evidence (Fig. 2**C**). In fact, some researchers have argued that confidence-accuracy dissociations may automatically imply that confidence is based on additional cues instead of being based on the sensory evidence (Schwartz, 2024).

This explanation is quite common in domains outside of perception. A paradigmatic example is Rhodes & Castel (2008)'s demonstration that participants exhibit higher confidence in their ability to memorize words displayed in larger font sizes. Here, the font size is used as a cue for confidence even though it has minimal or no effect on the internal encoding of the words.

Within the field of perception, several studies have suggested that confidence is at least partially affected by additional cues. For example, the weighted evidence and visibility (WEV) model postulates that confidence is based on a weighted average of the internal sensory evidence and a "visibility" signal that reflects observers' subjective estimate of the difficulty of each trial (Hellmann et al., 2023; Rausch et al., 2018, 2021). The model has been used to explain how task difficulty influences confidence separately for correct vs. error trials, but it has so far not been used to explain any known confidence-accuracy dissociation.

Nevertheless, there are at least three examples where the use of additional cues has been used to explain confidence-accuracy dissociations. First, two studies have found confidence-accuracy dissociations that emerge from confidence being influenced by the amount of physical effort that participants exerted while indicating their choice (Gajdos et al., 2019; Turner et al., 2021). The results are consistent with many other studies that demonstrate that action information can be used to inform confidence (Bocheva & Rahnev, 2025; Filevich et al., 2020; Fleming et al., 2015; Pereira et al., 2023; Sanchez et al., 2024; Siedlecka et al., 2021). Second, one study has argued that, under certain

circumstances, confidence may be informed by decision time, resulting in lower confidence for conditions where RT is longer despite no loss of accuracy (Kiani et al., 2014). Finally, several studies have demonstrated that confidence-accuracy dissociations can be obtained by manipulating people's beliefs about their own abilities (e.g., by giving false feedback) (Boldt et al., 2019; Desender et al., 2025; Van Marcke et al., 2024).

These studies make it clear that confidence in perceptual tasks is likely to be affected by several cues not directly related to the internal sensory evidence. This makes sense – humans use available information to make better judgments. In fact, real-world judgments are likely to be influenced even more strongly by a host of cues unrelated to the internal sensory evidence (Shekhar et al., 2025). That said, it is also true that non-sensory cues for confidence likely have a relatively small influence in most current experimental work on perceptual confidence, which tends to use well controlled stimuli that minimize relevant environmental cues.

3.4. Cognitive limitations

Another explanation for some confidence-accuracy dissociations is that they are due to cognitive limitations (Fig. 2**D**). Unlike the cues discussed in Section 3.3, here participants' confidence ratings are influenced by factors that should not affect confidence, and which participants try to ignore rather than actively use.

An elegant example of how confidence limitations can lead to confidence-accuracy dissociations comes from the work of Spence et al. (2018). The researchers used pairs of dot motion displays in which the dots varied on two dimensions: brightness and range of motion directions. After each pair, a post-cue indicated whether participants should report which display was brighter or the relative motion direction of the second display. Crucially, when participants performed the brightness discrimination, the range of motion directions should have had no bearing on their judgment because motion was irrelevant to the judged feature. Nevertheless, increasing the motion-direction range reliably reduced confidence in brightness judgments while leaving accuracy unchanged. These results suggest that when judging confidence, humans are unable to completely disregard superfluous information even if they know that the information is irrelevant.

Another type of cognitive limitation that has been used to explain confidence-accuracy dissociations is the inability to keep two independent sets of confidence criteria. This phenomenon was first described in the context of perceptual decisions (Gorea & Sagi, 2000, 2001, 2002). More recently, it was used to produce a confidence-accuracy dissociation by interleaving two conditions that optimally should have a distinct set of confidence criteria (Rahnev, 2021a). Participants judged the identity of stimuli generated with either low or high external noise. The results showed that the two conditions were matched on accuracy and RT but produced a large difference in confidence that stemmed from attraction between the confidence criteria used in each condition. A similar explanation can be used to explain the phenomenon of "confidence leak" where confidence on the previous trial influences the confidence on the current trial even when participants are performing two independent tasks (Rahnev et al., 2015). Nevertheless, the conclusion in Section 3.3 applies here too: while cognitive limitations can certainly induce confidence-accuracy dissociations for specific designs, they are unlikely to meaningfully contribute to the majority of confidence-accuracy dissociations reported in the literature.

3.5. Other explanations

Many other explanations for confidence-accuracy dissociations are specific to the paradigm used to induce them. For example, several studies have used brain stimulation to the prefrontal cortex to alter confidence without affecting accuracy (Rounis et al., 2010; Saccenti et al., 2025; Shekhar & Rahnev, 2018; Xue et al., 2023), which can be explained by postulating a bigger role for these brain areas in confidence compared to perceptual decisions. Similarly, studies have shown that the pre-stimulus alpha power in visual cortex affects confidence but not accuracy (Samaha et al., 2017, 2022), which can be explained by assuming that pre-stimulus alpha power affects the strength but not the quality of the sensory signal. Somewhat relatedly, one study demonstrated the possibility of confidence being less affected by unconscious information compared to the perceptual decision (Vlassova et al., 2014), though this effect awaits replication and likely can't explain other confidence-accuracy dissociations. Finally, still another explanation for confidence-accuracy dissociations is that they arise due to post-decisional processes (Navajas et al., 2016). However, follow-up research has shown that this is not true for confidence-accuracy dissociations based on overall-strength manipulations (Samaha & Denison, 2022), while other research has questioned whether confidence is routinely based on post-decisional information (Chen & Rahnev, 2023; Matthews et al., 2025; Xue et al., 2023). Overall, these effects can explain specific confidence-accuracy dissociations, but none of them applies more broadly to a large swath of the dissociations in the literature.

A final class of explanations draws on normative or Bayesian considerations. Indeed, normative considerations have been used to explain several confidence-accuracy dissociations (Baldassi et al., 2006; Khalvati et al., 2021; Rahnev et al., 2015; Zylberberg et al., 2014). Several specific confidence-accuracy dissociations are particularly well explained by normative considerations: for example, the finding that participants who come to believe that they are bad at a task use low confidence (Boldt et al., 2019; Desender et al., 2025; Van Marcke et al., 2024) reflects a normative approach to confidence. However, normative explanations typically work at Marr's computational level (Marr, 1982), meaning that they are concerned with how a system should operate. Conversely, the explanations we focus on here fall within Marr's algorithmic level. In fact, to fit actual data, normative models need to make assumptions about the algorithmic level, and those assumptions at the algorithmic level can be seen as the more direct explanations of an effect (Rahnev & Denison, 2018a). For example, a Bayesian model based on partially observable Markov decision processes (POMDPs) was able to explain overall-strength-based confidence-accuracy dissociations by using the variance neglect mechanism covered in Section 3.2 (Khalvati et al., 2021). Thus, while normative considerations are important for understanding confidence-accuracy dissociations, a normative model ultimately needs to make assumptions at the algorithmic level that can then be seen as the direct explanation of the observed dissociation.

4. Confusions, controversies, and open questions

Having reviewed the various findings and explanations related to confidence-accuracy dissociations, here we discuss in more detail three issues that require special attention by researchers.

4.1. Should the term "positive evidence" be used to describe stimulus manipulations?

As discussed in Section 2.1 and 3.1, the term "positive evidence" has a dual meaning in the literature. In some cases, it is used to refer to stimulus manipulations (which we call "overall-strength manipulations" here). In other cases, it is used to refer to the computation underlying confidence judgments. Unfortunately, it is often unclear which of these two meanings is intended. Further, as reviewed in Section 3.1, it is common for researchers to think that confidence-accuracy dissociations produced by a "positive evidence" manipulation (i.e., an overall-strength manipulation) provide direct evidence for a "positive evidence" computation even though variance neglect may be a more plausible mechanistic explanation (see Section 3.2).

We therefore suggest that the term "positive evidence" should be reserved for either the manipulation or the computation but not be used for both. Because the plain meaning of the term (as well as the

predominant use in the literature) is to convey a principle of confidence computation, we argue that the term "positive evidence" should be reserved for the implied confidence computation. In that case, the stimulus manipulations should not be called by the same name. It is possible to flip this reasoning and argue that "positive evidence" should be used only for the stimulus manipulation (and the resulting behavioral effect) but not for the confidence computation. However, historically, the reason for adopting the name "positive evidence" for the stimulus manipulation was the assumption that the manipulation works because the underlying computation only considers the decision-congruent evidence. Further, if one is to drop this computational assumption, then there would seem to be no reason to adopt this specific name for the stimulus manipulation. For example, under the interpretation that the manipulation increases both positive and negative internal evidence, then the label "negative evidence" manipulation should apply equally well. The reason this term sounds odd is the fact that we judge the manipulation name based on whether it conforms with the assumed computation. Thus, to avoid confusion, we argue that it is best to reserve "positive evidence" for the computation, but use theory-neutral language for the stimulus manipulation.

Here we propose that the term "overall-strength manipulation" is the most appropriate as it is general enough to capture the variety of manipulations in this class while remaining theory-neutral. In this language, the "overall-strength manipulation" produces an "overall-strength bias" where high-strength stimuli are judged with higher confidence compared to low-strength stimuli even for matched accuracy. We argue that even if it ultimately turns out that overall-strength manipulations produce confidence-accuracy dissociations via choice-congruent computations, the clear separation of terminology for stimulus manipulation vs. confidence computation would still be an asset. And this separation is even more valuable if the choice-congruent computation is ultimately found not to describe confidence judgments.

4.2. Confidence-accuracy dissociations and metacognitive inefficiency

A separate confusion sometimes arises when considering the relationship between confidence-accuracy dissociations and metacognitive inefficiency. Metacognitive efficiency refers to measures that quantify how well a participant can distinguish between their correct and incorrect judgments given a certain level of task performance (Fleming & Lau, 2014). Metacognitive inefficiency typically refers to the observation that confidence ratings are usually less predictive of accuracy than the theoretical maximum (Maniscalco & Lau, 2012). Consequently, metacognitive inefficiency is virtually synonymous with the idea that confidence judgments are subject to metacognitive noise (Bang et al., 2019; Boundy-Singer et al., 2023; Guggenmos, 2022; Mamassian & de Gardelle, 2022; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2021b). In contrast, as can be seen in Section 3, confidence-accuracy dissociations are explained by many mechanisms, but metacognitive noise is not usually one of them. The reason is that metacognitive noise leads to noisier trial-by-trial confidence ratings, but this noise is averaged out over trials and does not reliably lead to one condition having higher confidence than another. Therefore, metacognitive inefficiency, understood as synonymous with metacognitive noise, is a phenomenon that is mostly separate from confidence-accuracy dissociations. Relatedly, models of metacognition successfully explain confidence-accuracy dissociations if they include the mechanisms discussed in Section 3; the inclusion of metacognitive noise in such models can improve the fits and explain other effects but typically not the dissociation itself (one exception here is Boldt et al., 2025 who used condition-dependent metacognitive noise).

That said, metacognitive inefficiency can be considered in a broader perspective as reflecting any suboptimality in confidence judgment. Under this more expansive definition, some confidence-accuracy dissociations that do not reflect normative computations (e.g., see Section 3.4) can be seen as examples of metacognitive inefficiency. It is in this sense that confidence-accuracy dissociations have been said to constitute "model-free evidence for metacognitive inefficiency" (Shekhar & Rahnev, 2021a). To avoid confusion, we suggest that future studies should reserve the term "metacognitive inefficiency" to refer specifically to confidence corruptions that happen within a single condition (which are typically modeled as some type of metacognitive noise). Instead, the broader sense of any suboptimality in confidence judgment implied by Shekhar & Rahnev (2021a) should be conveyed more directly as "metacognitive suboptimality."

4.3. The role of RT in confidence-accuracy dissociations

Shorter RTs are generally associated with both higher accuracy and higher confidence (Pleskac & Busemeyer, 2010), making it important to understand whether RT would be coupled to confidence or accuracy in cases where these are dissociated. Kiani et al. (2014) argued that confidence is causally affected by RT, which would predict that RT would mostly follow confidence in cases of confidence-accuracy dissociations. However, while most papers examining confidence-accuracy dissociations do not explicitly examine the role of RT, the papers that do examine RT paint a complex picture.

Some early studies suggested that RT indeed follows confidence in conditions of confidence-accuracy dissociations. This was true for overall-strength manipulations that used Gabor patches (Samaha et al., 2016), RDKs (Zylberberg et al., 2016), and ensembles of colored circles (Boldt et al., 2019). However, other studies have failed to replicate this pattern. For example, Rollwage et al. (2020) employed an overall-strength manipulation with an RDK and showed that RTs followed accuracy instead of confidence. RTs also followed accuracy when comparing the effects of different manipulations on accuracy vs. confidence (Fung et al., 2025) and in a study that obtained a confidence-accuracy dissociation via criterion attraction (Rahnev, 2021a). A more recent study that reported on three separate confidence-accuracy dissociations in the context of multi-alternative perceptual judgments also found a complex pattern of results such that RT followed confidence in some cases, accuracy in other cases, and sometimes reflected a compromise between these two variables (Xue et al., 2024).

Overall, these effects are inconsistent with any simple story of how RT behaves in cases of confidence-accuracy dissociation. Given the interest of building joint models of choice, confidence, and RT (Hellmann et al., 2023; Herregods et al., 2023; Pleskac & Busemeyer, 2010), as well as using RT as a proxy for confidence (Miyoshi et al., 2025), it is important for future studies presenting confidence-accuracy dissociations to report on how RT behaves in these studies. This practice would ultimately allow us to better understand the complex interrelationship between accuracy, confidence, and RT.

5. Conclusions

This paper examined the confidence-accuracy dissociations reported in the context of perceptual decision making. Section 2 organized the many dissociations into several categories, Section 3 explored the different explanations for these dissociations, and Section 4 delved into several issues that the field should pay more attention to. Overall, the review demonstrates the diversity of mechanisms behind confidence-accuracy dissociations and should help researchers position any new findings within the existing literature. Beyond this, the current paper highlights five conclusions of particular importance.

First, confidence-accuracy dissociations are invaluable for understanding the nature of the confidence computation. Virtually any model of confidence predicts that confidence and accuracy go together under normal circumstances (because they almost always do). Therefore, the ability to correctly explain circumstances where confidence and accuracy dissociate is critical for demonstrating the generalizability of models. Going forward, researchers should focus on building models that not only explain existing confidence-accuracy dissociations but that

predict new ones (Shmueli, 2010).

Second, researchers should reserve the term "positive evidence" for confidence computations and not use it to refer to stimulus manipulations (see Section 4.1). The term currently can refer to both an experimental paradigm and to a computational mechanism. This encourages the notion that the success of the experimental paradigm provides direct evidence for the computational mechanism. Because "positive evidence" was originally intended as a computational mechanism, we advocate for calling the experimental paradigm "overall-strength manipulation" and have done so in this paper.

Third, confidence-accuracy dissociations are not, in and of themselves, evidence for non-normative computation. While such dissociations are by definition suboptimal for an agent with full information about the world, humans rarely possess such complete information. Virtually all existing explanations of confidence-accuracy dissociations can be seen from a normative lens as either reflecting incomplete information about the environment (Khalvati et al., 2021), reasonable assumptions about the environment (Rahnev et al., 2015), or cognitive limitations (Rahnev, 2021a). These explanations are generally consistent with the resource rationality research program (Lieder & Griffiths, 2019). Conversely, thinking about confidence-accuracy dissociations from a purely normative perspective may also be counterproductive as it tends to de-emphasize the critical role of the algorithmic considerations that give rise to these dissociations. Overall, while confidence-accuracy dissociations certainly have implications about the extent of human optimality or suboptimality (Rahnev & Denison, 2018b), they do not directly reveal whether the underlying computation is Bayesian or not.

Fourth, confidence-accuracy dissociations are not, in and of themselves, evidence for metacognitive noise (see Section 4.2). Indeed, metacognitive noise leads to noisier trial-by-trial confidence ratings, but this noise is averaged out over trials and does not reliably lead to one condition having higher confidence than another. Therefore, metacognitive noise itself does not typically lead to confidence-accuracy dissociations, and existing dissociations are rarely explained by metacognitive noise. Confidence-accuracy dissociations should thus not be seen as evidence for metacognitive inefficiency as that term is typically understood to be synonymous with metacognitive noise.

Finally, the role of RT in confidence-accuracy dissociations is a critical topic for future research. Current findings about the role of RT in confidence-accuracy dissociations are inconsistent: RT sometimes follows confidence, sometimes follows accuracy, and sometimes reflects a compromise between these two variables. How RT behaves in cases where confidence and accuracy dissociate has enormous implications for understanding the computations that jointly determine choice, confidence, and RT. Therefore, future studies should routinely report RT in studies that demonstrate confidence-accuracy dissociations.

Confidence-accuracy dissociations are one of the best tools we have for understanding not only the mechanisms of confidence but also how sensory information is represented internally. Discovering new dissociations and finding better explanations of existing ones should remain a central goal for the field.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Institutes of Health (award: R01MH119189). The author thanks the Vision Science Society (VSS) and Elsevier for sponsoring the Elsevier/VSS Young Investigator Award.

Data availability

No data was used for the research described in the article.

References

Adams, J. K. (1957). A confidence scale defined in terms of expected percentages. *The American Journal of Psychology, 70*(3), 432–436.

Aguilar-Lleyda, D., Konishi, M., Sackur, J., & de Gardelle, V. (2021). Confidence can be automatically integrated across two visual decisions. *Journal of Experimental Psychology: Human Perception and Performance, 47*(2), 161–171. https://doi.org/10.1037/xhp0000884

Baldassi, S., Megna, N., & Burr, D. C. (2006). Visual clutter causes high-magnitude errors. *PLoS Biology, 4*(3), e56.

Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology: General, 148*(3), 437–452. https://doi.org/10.1037/xge0000511

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55*(4), 412–428.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics, 54*(1), 75–81.

Bocheva, M., & Rahnev, D. (2025). Switching the motor response weakens confidence serial dependence. *Psychonomic Bulletin & Review, 32*(6), 3134–3142. https://doi.org/10.3758/s13423-025-02748-9

Bocheva, M., Xue, K., & Rahnev, D. (2025). *Serial effects in choice and confidence modulate each other: Evidence from 26 experiments*. PsyArXiv. https://doi.org/10.31219/osf.io/rzus6_v1.

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance, 43*(8), 1520–1531. https://doi.org/10.1037/xhp0000404

Boldt, A., Schiffer, A.-M., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific Reports, 9*(1), 4031. https://doi.org/10.1038/s41598-019-40681-9

Boldt, A., Sun, Y., & Desender, K. (2025). How disconfirmatory evidence shapes confidence in decision-making. *Communications Psychology, 3*(1), 150. https://doi.org/10.1038/s44271-025-00325-3

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour, 7*(1), 142–154. https://doi.org/10.1038/s41562-022-01464-x

Busey, T. A, Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26–48.

Chen, S., & Rahnev, D. (2023). Confidence response times: Challenging postdecisional models of confidence. *Journal of Vision, 23*(7), 11. https://doi.org/10.1167/jov.23.7.11

Comay, N. A., Della Bella, G., Lamberti, P., Sigman, M., Solovey, G., & Barttfeld, P. (2023). The presence of irrelevant alternatives paradoxically increases confidence in perceptual decisions. *Cognition, 234*, Article 105377. https://doi.org/10.1016/j.cognition.2023.105377

Dawes, R. M. (1980). Confidence in intellectual vs. confidence in perceptual judgments. In E. D. Lantermann, & H. Feger (Eds.), *Similarity and choice: Papers in honor of Clyde Coombs* (pp. 327–345). Han Huber.

de Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence judgments. *PLoS One, 10*(3), Article e0120870. https://doi.org/10.1371/journal.pone.0120870

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences, 108*(32), 13341–13346. https://doi.org/10.1073/pnas.1104517108

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences, 115*(43), 11090–11095. https://doi.org/10.1073/pnas.1717720115

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science, 29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Desender, K., Smet, A., Erdil, D., Gunay, E. M., Visser, Y. F., Le Denmat, P., & Van Marcke, H. (2025). Common neural choice signals reflect accumulated evidence, not confidence. *Cerebral Cortex, 35*(8), Article bhaf257. https://doi.org/10.1093/cercor/bhaf257

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron, 83*(4), 797–804. https://doi.org/10.1016/j.neuron.2014.07.011

Filevich, E., Koß, C., & Faivre, N. (2020). Response-related signals increase confidence but not metacognitive performances. *eNeuro*, ENEURO.0326-19.2020. https://doi.org/10.1523/ENEURO.0326-19.2020.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology, 75*(1), 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00443
Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. *Psychological Science, 26*(1), 89–98. https://doi.org/10.1177/0956797614557697
Fung, H., Shekhar, M., Xue, K., Rausch, M., & Rahnev, D. (2025). Similarities and differences in the effects of different stimulus manipulations on accuracy and confidence. *Consciousness and Cognition, 136*, Article 103942. https://doi.org/10.1016/j.concog.2025.103942
Gajdos, T., Fleming, S., Garcia, M. S., Weindel, G., & Davranche, K. (2019). Revealing subthreshold motor contributions to perceptual confidence. *Neuroscience of Consciousness, 5*(1), Article niz001. https://doi.org/10.1101/330605
Gao, Y., Xue, K., Odegaard, B., & Rahnev, D. (2025). Automatic multisensory integration follows subjective confidence rather than objective performance. *Communications Psychology, 3*(1), 1–11. https://doi.org/10.1038/s44271-025-00221-w
Gill, M. J., Swann, W. B., Jr., & Silvera, D. H. (1998). On the genesis of confidence. *Journal of Personality and Social Psychology, 75*(5), 1101–1114. https://doi.org/10.1037/0022-3514.75.5.1101
Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences of the United States of America, 97*(22), 12380–12384. https://doi.org/10.1073/pnas.97.22.12380
Gorea, A., & Sagi, D. (2001). Disentangling signal from noise in visual contrast discrimination. *Nature Neuroscience, 4*(11), 1146–1150. https://doi.org/10.1038/nn741
Gorea, A., & Sagi, D. (2002). Natural extinction: A criterion shift phenomenon. *Visual Cognition, 9*(8), 913–936. https://doi.org/10.1080/13506280143000638
Green, M. L., Hu, M., Denison, R. N., & Rahnev, D. (2026). Using artificial neural networks to relate external sensory features to internal decisional evidence. *Open Mind, 10*, 29–46. https://doi.org/10.1162/OPMI.a.317
Guggenmos, M. (2022). Reverse engineering of metacognition. *eLife*, 11. https://doi.org/10.7554/eLife.75420
Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review, 16*(5), 824–831. https://doi.org/10.3758/PBR.16.5.824
Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review, 130*(6), 1521–1543. https://doi.org/10.1037/rev0000411
Herce Castañón, S., Moran, R., Ding, J., Egner, T., Bang, D., & Summerfield, C. (2019). Human noise blindness drives suboptimal cognitive inference. *Nature Communications, 10*(1), 1719. https://doi.org/10.1038/s41467-019-09330-7
Herregods, S., Denmat, P. L., & Desender, K. (2023). *Modelling Speed-Accuracy Tradeoffs in the Stopping Rule for Confidence Judgments* (p. 2023.02.27.530208). bioRxiv. https://doi.org/10.1101/2023.02.27.530208.
Kantner, J., Solinger, L. A., Grybinas, D., & Dobbins, I. G. (2019). Confidence carryover during interleaved memory and perception judgments. *Memory & Cognition, 47*(2), 195–211. https://doi.org/10.3758/s13421-018-0859-8
Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037
Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica, 67*(2), 95–119. https://doi.org/10.1016/0001-6918(88)90007-8
Khalvati, K., Kiani, R., & Rao, R. P. N. (2021). Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature Communications, 12*(1), 5704. https://doi.org/10.1038/s41467-021-25419-4
Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015
Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics, 77*(4), 1295–1306. https://doi.org/10.3758/s13414-015-0843-3
Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General, 140*(1), 117–139. https://doi.org/10.1037/a0022171
Kurtz, P., Shapcott, K. A., Kaiser, J., Schmiedt, J. T., & Schmid, M. C. (2017). The influence of endogenous and exogenous spatial attention on decision confidence. *Scientific Reports, 7*(1), 6431. https://doi.org/10.1038/s41598-017-06715-w
Lapate, R. C., Samaha, J., Rokers, B., Postle, B. R., & Davidson, R. J. (2020). Perceptual metacognition of human faces is causally supported by function of the lateral prefrontal cortex. *Communications Biology, 3*(1), 360. https://doi.org/10.1038/s42003-020-1049-3
Law, F., Ng, L., Lam, A., Or, C., & Lee, A. (2025). *Serial dependence of confidence judgments reveals metacognitive adaptation*. PsyArXiv. https://doi.org/10.31234/osf.io/f35br_v1.
Lee, A. L. F., Yabuki, H., Lee, I. C. L., & Or, C.-C.-F. (2023). Metacognitive bias resulting from trade-off between local and global motion signals. *Journal of Vision, 23*(10), 7. https://doi.org/10.1167/jov.23.10.7
Lee, J. L., Denison, R., & Ma, W. J. (2023). Challenging the fixed-criterion model of perceptual decision-making. *Neuroscience of Consciousness, 2023*(1), Article niad010. https://doi.org/10.1093/nc/niad010
Li, H.-H., & Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nature Communications, 11*(1), 2004. https://doi.org/10.1038/s41467-020-15581-6
Li, M. K., Lau, H., & Odegaard, B. (2018). An investigation of detection biases in the unattended periphery during simulated driving. *Attention, Perception, & Psychophysics, 80*(6), 1325–1332. https://doi.org/10.3758/s13414-018-1554-3
Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 1–85*. https://doi.org/10.4319/lo.2007.52.2.0517
Locke, S., Mamassian, P., & Landy, M. (2018). Sequential effects in confidence. *Journal of Vision, 18*(10), 658. https://doi.org/10.1167/18.10.658
Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science, 2*(1). https://doi.org/10.1146/annurev-vision-111815-114630. annurev-vision-111815-114630.
Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review, 129*(5), 976–998. https://doi.org/10.1037/rev0000312
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021
Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness, 2016*(1), 1–17. https://doi.org/10.1093/nc/niw002
Maniscalco, B., Odegaard, B., Grimaldi, P., Cho, S. H., Basso, M. A., Lau, H., & Peters, M. A. K. (2021). Tuned inhibition in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior. *PLoS Computational Biology, 17*(3), Article e1008779. https://doi.org/10.1371/journal.pcbi.1008779
Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics, 78*(3), 923–937. https://doi.org/10.3758/s13414-016-1059-x
Marr, D. (1982). *Vision*. Freeman.
Matthews, J. R., Sugihara, N., Nagisa, S., Ohashi, H., & Shibata, K. (2025). The order of task decisions and confidence ratings has little effect on metacognition. *Communications Psychology, 3*(1), 140. https://doi.org/10.1038/s44271-025-00321-7
Mazor, M., Maimon-Mor, R. O., Charles, L., & Fleming, S. M. (2023). Paradoxical evidence weighting in confidence judgments for detection and discrimination. *Attention, Perception, & Psychophysics, 85*(7), 2356–2385. https://doi.org/10.3758/s13414-023-02710-8
Mei, N., Rahnev, D., & Soto, D. (2023). Using serial dependence to predict confidence across observers and cognitive domains. *Psychonomic Bulletin & Review, 30*(4), 1596–1608. https://doi.org/10.3758/s13423-023-02261-x
Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review, 127*(5), 655–671. https://doi.org/10.1037/rev0000184
Miyoshi, K., Rahnev, D., & Lau, H. (2025). *Response Time as Decision Confidence: Insights from Type-2 ROC Analysis* (6gyjf_v1). PsyArXiv. https://doi.org/10.31234/osf.io/6gyjf_v1.
Miyoshi, K., & Sakamoto, Y. (2025). Early sensory evidence shapes multichoice confidence bias: A registered report. *Journal of Experimental Psychology: General, 154*(11), 2944–2962. https://doi.org/10.1037/xge0001809
Miyoshi, K., Webb, T., Rahnev, D., & Lau, H. (2024). Confidence and metacognition. In *Encyclopedia of the Human Brain* (2nd edition, pp. 252–268).
Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *The Journal of Neuroscience, 38*(14), 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018
Morales, J., Solovey, G., Maniscalco, B., Rahnev, D., de Lange, F. P., & Lau, H. (2015). Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. *Attention, Perception & Psychophysics, 77*(6), 2021–2036. https://doi.org/10.3758/s13414-015-0897-2
Navajas, J., Bahrami, B., & Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences, 11*, 55–60. https://doi.org/10.1016/j.cobeha.2016.05.005
Navajas, J., Sigman, M., & Kamienkowski, J. E. (2014). Dynamics of visibility, confidence, and choice during eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 1213–1227. https://doi.org/10.1037/a0036321
Odegaard, B., Chang, M. Y., Lau, H., & Cheung, S.-H. (2018). Inflation versus filling-in: Why we feel we see more than we actually do in peripheral vision. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1755), Article 20170345. https://doi.org/10.1098/rstb.2017.0345
Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A. K., Lau, H., & Basso, M. A. (2018). Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences of the United States of America, 115*(7), E1588–E1597. https://doi.org/10.1073/pnas.1711628115
Okazawa, G., Sha, L., Purcell, B. A., & Kiani, R. (2018). Psychophysical reverse correlation reflects both sensory and decision-making processes. *Nature Communications, 9*(1), 3479. https://doi.org/10.1038/s41467-018-05797-y
Pereira, M., Skiba, R., Cojan, Y., Vuilleumier, P., & Bègue, I. (2023). Preserved metacognition for undetected visuomotor deviations. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0133-23.2023
Peters, M. A. K. (2022). Confidence in Decision-Making. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.371.
Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour, 1*(7), 0139. https://doi.org/10.1038/s41562-017-0139

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864–901. https://doi.org/10.1037/a0019737

Pruitt, J., Knotts, J. D., & Odegaard, B. (2024). Consistent metacognitive efficiency and variable response biases in peripheral vision. *Journal of Vision, 24*(8), 4. https://doi.org/10.1167/jov.24.8.4

Rafiei, F., Shekhar, M., & Rahnev, D. (2024). The neural network RTNet exhibits the signatures of human perceptual decision-making. *Nature Human Behaviour, 8*(9), 1752–1770. https://doi.org/10.1038/s41562-024-01914-8

Rahnev, D. (2020). Confidence in the real world. *Trends in Cognitive Sciences, 24*(8), 590–591. https://doi.org/10.1016/J.TICS.2020.05.005

Rahnev, D. (2021a). A robust confidence–accuracy dissociation via criterion attraction. *Neuroscience of Consciousness, 2021*(1), 1–11. https://doi.org/10.1093/NC/NIAB039

Rahnev, D. (2021b). Visual metacognition: Measures, models, and neural correlates. *American Psychologist, 76*(9), 1445–1453. https://doi.org/10.1037/amp0000937

Rahnev, D., Bahdo, L., de Lange, F. P., & Lau, H. (2012). Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *Journal of Neurophysiology, 108*(5), 1529–1536. https://doi.org/10.1152/jn.00184.2012

Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M., Reyes, G., Rouault, M., Sackur, J., & Zylberberg, A. (2022). Consensus goals in the field of visual metacognition. *Perspectives on Psychological Science, 17*(6), 1746–1765. https://doi.org/10.1177/17456916221075615

Rahnev, D., & Denison, R. N. (2018a). Behavior is sensible but not globally optimal: Seeking common ground in the optimality debate. *Behavioral and Brain Sciences, 41*, e251.

Rahnev, D., & Denison, R. N. (2018b). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences, 41*(e223), 1–66. https://doi.org/10.1017/S0140525X18000936

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological Science, 26*(11), 1664–1680. https://doi.org/10.1177/0956797615595037

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., De Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience, 14*(12), 1513–1515. https://doi.org/10.1038/nn.2948

Rahnev, D., Maniscalco, B., Luber, B., Lau, H., & Lisanby, S. H. (2012). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of Neurophysiology, 107*(6), 1556–1563. https://doi.org/10.1152/jn.00985.2011

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences, 113*(21), 6059–6064. https://doi.org/10.1073/pnas.1522551113

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, and Psychophysics, 80*, 134–154. https://doi.org/10.3758/s13414-017-1431-5

Rausch, M., Hellmann, S., & Zehetleitner, M. (2021). Modelling visibility judgments using models of decision confidence. *Attention, Perception, & Psychophysics, 83*(8), 3311–3336. https://doi.org/10.3758/s13414-021-02284-3

Rausch, M., Zehetleitner, M., Steinhauser, M., & Maier, M. E. (2020). Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage, 218*(January), Article 116963. https://doi.org/10.1016/j.neuroimage.2020.116963

Recht, S., Mamassian, P., & de Gardelle, V. (2019). Temporal attention causes systematic biases in visual confidence. *Scientific Reports, 9*(1), 11622. https://doi.org/10.1038/s41598-019-48063-x

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–625. https://doi.org/10.1037/a0013684

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications, 11*(1), 2634. https://doi.org/10.1038/s41467-020-16278-6

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience, 1*(3), 165–175. https://doi.org/10.1080/17588921003632529

Saccenti, D., Moro, A. S., Salvetti, G., Sassaroli, S., Malgaroli, A., Lamanna, J., & Ferro, M. (2025). Transcranial direct current stimulation over the orbitofrontal cortex enhances self-reported confidence but reduces metacognitive sensitivity in a perceptual decision-making task. *Biomedicines, 13*(7), Article 7. https://doi.org/10.3390/biomedicines13071522

Sakamoto, Y., & Miyoshi, K. (2024). A confidence framing effect: Flexible use of evidence in metacognitive monitoring. *Consciousness and Cognition, 118*, Article 103636. https://doi.org/10.1016/j.concog.2024.103636

Samaha, J., Barrett, J. J., Sheldon, A. D., LaRocque, J. J., & Postle, B. R. (2016). Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Frontiers in Psychology, 7*, 851. https://doi.org/10.3389/fpsyg.2016.00851

Samaha, J., & Denison, R. (2022). The positive evidence bias in perceptual confidence is unlikely post-decisional. *Neuroscience of Consciousness, 2022*(1). https://doi.org/10.1093/nc/niac010

Samaha, J., Iemi, L., & Postle, B. R. (2017). Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Consciousness and Cognition, 54*, 47–55. https://doi.org/10.1016/j.concog.2017.02.005

Samaha, J., LaRocque, J. J., & Postle, B. R. (2022). Spontaneous alpha-band amplitude predicts subjective visibility but not discrimination accuracy during high-level perception. *Consciousness and Cognition, 102*, Article 103337. https://doi.org/10.1016/J.CONCOG.2022.103337

Samaha, J., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation estimation. *Journal of Vision, 19*(4), 25. https://doi.org/10.1167/19.4.25

Sanchez, R., Courant, A., Desantis, A., & Gajdos, T. (2024). Making precise movements increases confidence in perceptual decisions. *Cognition, 249*, Article 105832. https://doi.org/10.1016/j.cognition.2024.105832

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron, 90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Schwartz, B. L. (2024). Inferential theories of retrospective confidence. *Metacognition and Learning, 19*(3), 907–938. https://doi.org/10.1007/s11409-024-09396-9

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *eLife, 9*, 1–58. https://doi.org/10.7554/eLife.60705

Sharvashidze, N., Toscani, M., & Valsecchi, M. (2025). Peripheral overconfidence in a scene categorization task. *Journal of Vision, 25*(10), 2. https://doi.org/10.1167/jov.25.10.2

Shekhar, M., Cleeremans, A., & Rahnev, D. (2025). *Confidence in real-world decision making* (Qam49_v1). PsyArXiv. https://doi.org/10.31234/osf.io/qam49_v1.

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *Journal of Neuroscience, 38*(22), 5078–5087. https://doi.org/10.1523/JNEUROSCI.3484-17.2018

Shekhar, M., & Rahnev, D. (2021a). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences, 25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Shekhar, M., & Rahnev, D. (2021b). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review, 128*(1), 45–70. https://doi.org/10.1037/rev0000249

Shekhar, M., & Rahnev, D. (2024a). How do humans give confidence? A comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General, 153*(3), 656–688. https://doi.org/10.1037/xge0001524

Shekhar, M., & Rahnev, D. (2024b). Human-like dissociations between confidence and accuracy in convolutional neural networks. *PLOS Computational Biology, 20*(11), Article e1012578. https://doi.org/10.1371/journal.pcbi.1012578

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Siedlecka, M., Koculak, M., & Paulewicz, B. (2021). Confidence in action: Differences between perceived accuracy of decision and motor response. *Psychonomic Bulletin & Review, 28*(5), 1698–1706. https://doi.org/10.3758/s13423-021-01913-0

Solovey, G., Graney, G. G., & Lau, H. (2014). A decisional account of subjective inflation of visual perception at the periphery. *Attention, Perception, & Psychophysics, 77*(1), 258–271. https://doi.org/10.3758/s13414-014-0769-1

Song, A., Koizumi, A., & Lau, H. (2015). A behavioral method to manipulate metacognitive awareness independent of stimulus awareness. In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (pp. 77–85). Oxford University Press.

Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology. Human Perception and Performance, 42*(5), 671–682. https://doi.org/10.1037/xhp0000179

Spence, M. L., Mattingley, J. B., & Dux, P. E. (2018). Uncertainty information that is irrelevant for report impacts confidence judgments. *Journal of Experimental Psychology: Human Perception and Performance, 44*(12), 1981–1994. https://doi.org/10.1037/xhp0000584

Tian, K. J., Maniscalco, B., Epstein, M. L., Shen, A., Castaneda, O. G., Kurosawa, T., Motzer, J. A., Olsson, E., Russell, E. E., Walsh, M. E., Wang, J., Zeb, T. B. A., Brown, R., Lamme, V. A. F., Lau, H., He, B. J., Brascamp, J. W., Block, N., Chalmers, D., … Denison, R. N. (2025). *Attention robustly dissociates objective performance and subjective visibility reports* (p. 2025.07.03.661972). bioRxiv. https://doi.org/10.1101/2025.07.03.661972.

Toscani, M., Mamassian, P., & Valsecchi, M. (2021). Underconfidence in peripheral vision. *Journal of Vision, 21*(6), 2. https://doi.org/10.1167/jov.21.6.2

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 479–496. https://doi.org/10.1016/S0022-5371(81)90129-8

Turner, W., Angdias, R., Feuerriegel, D., Chong, T.-T.-J., Hester, R., & Bode, S. (2021). Perceptual decision confidence is sensitive to forgone physical effort expenditure. *Cognition, 207*, Article 104525. https://doi.org/10.1016/j.cognition.2020.104525

Van Marcke, H., Denmat, P. L., Verguts, T., & Desender, K. (2024). Manipulating prior beliefs causally induces under- and overconfidence. *Psychological Science, 35*(4), 358–375. https://doi.org/10.1177/09567976241231572

Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence?. In *Proceedings of the Seventeenth Annual Meeting of the International Society for Psychophysics* (pp. 148–153).

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academy of Sciences, 111*(45), 16214–16218. https://doi.org/10.1073/pnas.1403619111

Webb, T. W., Miyoshi, K., So, T. Y., Rajananda, S., & Lau, H. (2023). Natural statistics support a rational account of confidence biases. *Nature Communications, 14*(1), Article 1. https://doi.org/10.1038/s41467-023-39737-2

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*, 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., & Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision, 8*(5), 1–10. https://doi.org/10.1167/8.5.7

Winter, C. J., & Peters, M. A. K. (2021). Variance misperception under skewed empirical noise statistics explains overconfidence in the visual periphery. *Attention, Perception, & Psychophysics, 84*(1), 161–178. https://doi.org/10.3758/s13414-021-02358-2

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences, 109*(16), 6354. https://doi.org/10.1073/pnas.1204601109

Xue, K., Fung, H., & Rahnev, D. (2026). Stimulus reliability but not boundary distance manipulations violate the folded-X pattern of confidence. *Cognition, 272*, Article 106490. https://doi.org/10.1016/j.cognition.2026.106490

Xue, K., Shekhar, M., & Rahnev, D. (2024). *A novel behavioral paradigm reveals the nature of confidence computation in multi-alternative perceptual decision making*. PsyArXiv. https://doi.org/10.31234/osf.io/67ufv.

Xue, K., Zheng, Y., Rafiei, F., & Rahnev, D. (2023). The timing of confidence computations in human prefrontal cortex. *Cortex, 168*, 167–175. https://doi.org/10.1016/j.cortex.2023.08.009

Yeon, J., Shekhar, M., & Rahnev, D. (2020). Overlapping and unique neural circuits are activated during perceptual decision making and confidence. *Scientific Reports, 10*(1), 20761. https://doi.org/10.1038/s41598-020-77820-6

Zak, I., Katkov, M., Gorea, A., & Sagi, D. (2012). Decision criteria in dual discrimination tasks estimated using external-noise methods. *Attention, Perception & Psychophysics, 74*(5), 1042–1055. https://doi.org/10.3758/s13414-012-0269-0

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology, 33*(4), 345–366. https://doi.org/10.1006/jesp.1997.1326

Zizlsperger, L., Sauvigny, T., & Haarmeier, T. (2012). Selective attention increases choice certainty in human decision making. *PLoS One, 7*(7), Article e41136. https://doi.org/10.1371/journal.pone.0041136

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience, 6*(September), 79. https://doi.org/10.3389/fnint.2012.00079

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife, 5*(5), Article e17688. https://doi.org/10.7554/eLife.17688

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition, 27*, 246–253. https://doi.org/10.1016/j.concog.2014.05.012