Evidence weighting in confidence judgments for detection and discrimination

Matan Mazor1, Lucie Charles4, Roni Maimon-Mor4,5, & Stephen M. Fleming1,2,3

1 Wellcome Centre for Human Neuroimaging, UCL

2 Max Planck UCL Centre for Computational Psychiatry and Ageing Research

3 Department of Experimental Psychology, UCL

4 Institute of Cognitive Neuroscience, UCL

5 FMRIB Centre, Nuffield Department of Clinical Neuroscience, University of Oxford

Author note

The authors have no conflicting interests to declare.

Correspondence concerning this article should be addressed to Matan Mazor, 12 Queen Square, London WC1N 3BG. E-mail: [mtnmzor@gmail.com](mailto:mtnmzor@gmail.com)

Abstract

Confidence in perceptual decisions is more sensitive to evidence in support of the decision than to evidence against it. This positive evidence bias (PEB) has been demonstrated in confidence ratings in binary discrimination decisions between two stimulus categories. Recent theoretical proposals suggest that a PEB is due to observers adopting a detection-like strategy when rating their confidence, one that has functional benefits for metacognition in real-world settings where detectability and discriminability often go hand in hand. However, it is unknown whether, or how, a PEB is also in play for detection decisions about the presence or absence of a stimulus. In three experiments (one lab-based and two online) we first successfully replicate a PEB in discrimination confidence. We then show that a PEB is observed in detection decisions, where participants report the presence or absence of a stimulus, regardless of its identity. We discuss our findings in relation to models that account for a positive evidence bias as emerging from a confidence-specific heuristic, and alternative models where decision and confidence are generated by the same, Bayes-rational process.

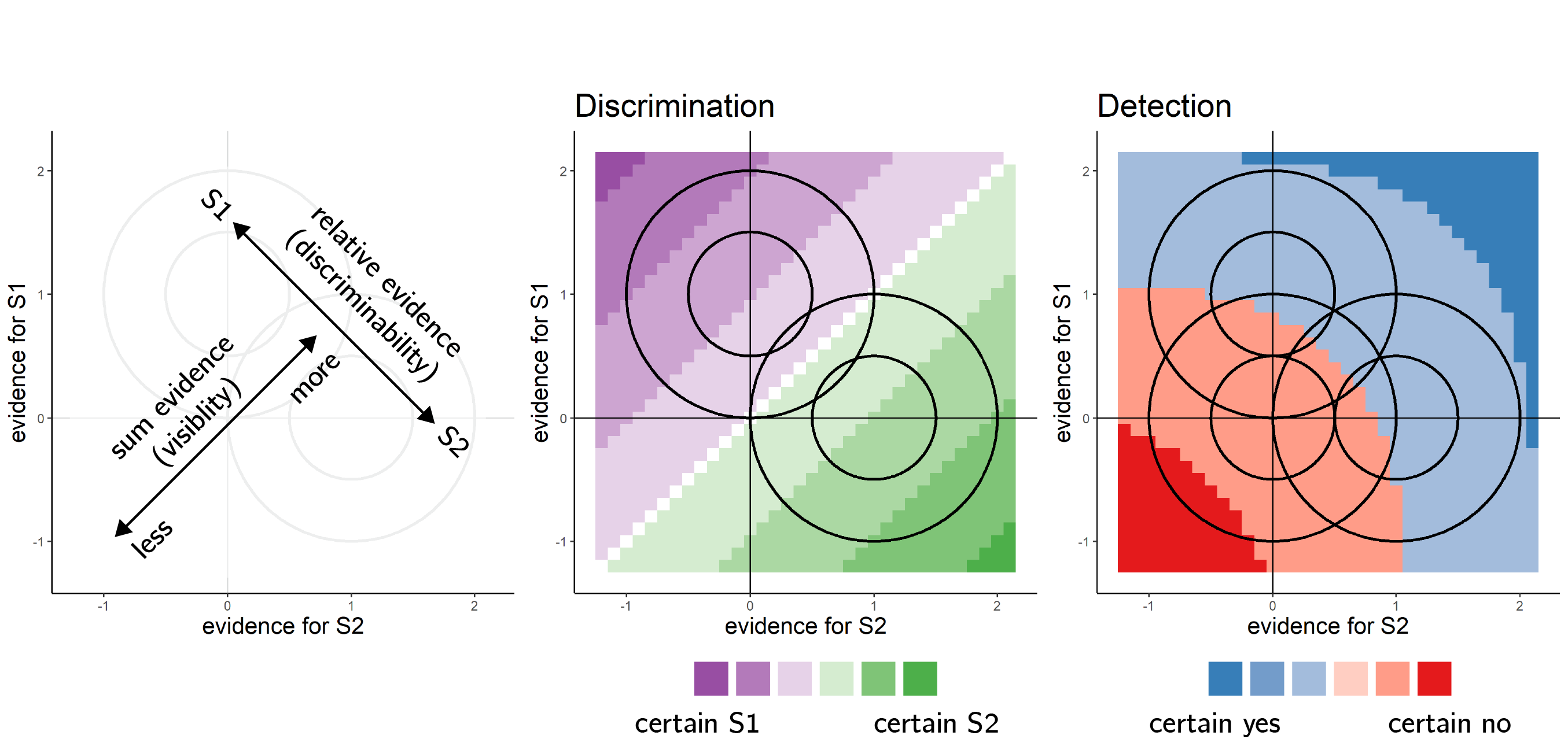
*Keywords:* confidence, detection, metacognition

*Word count:* 12,295

Evidence weighting in confidence judgments for detection and discrimination

# 1 Introduction

When considering two alternative hypotheses, the probability of a chosen hypothesis to be correct is not only a function of the likelihood of observations under the chosen hypothesis, but also under the unchosen one. For example, when deciding that a random dot motion display was drifting to the right and not to the left, confidence should not only positively weigh motion energy to the right (*positive evidence*), but also negatively weigh motion energy to the left (*negative evidence*). However, when rating subjective confidence, subjects place disproportional weight on evidence in favour of the choice, giving rise to a *positive evidence bias* (Koizumi, Maniscalco, & Lau, 2015; Peters et al., 2017; Rollwage et al., 2020; Samaha & Denison, 2020; Sepulveda et al., 2020; Zylberberg, Barttfeld, & Sigman, 2012). Equivalently, confidence ratings in discrimination are sensitive not only to the *relative evidence* of the chosen hypothesis compared with the unchosen one (also termed *balance of evidence*; see Fig. 1, left panel), but also to the *sum evidence* for the two hypotheses (which for perceptual decisions is often related to *visibility*, Rausch, Hellmann, & Zehetleitner, 2018).



*Figure* *1.*  Discrimination and detection in a two-dimensional Signal Detection Theory model. Left: in a two-dimensional SDT model, evidence is sampled from one of two Gaussian distributions (here centered at (0,1) and (1,0)). We define relative evidence as and sum evidence as . Circles represent contours of two-dimensional distributions. Center and Left: response and confidence accuracy are maximized when based on a log-likelihood ratio for the two stimulus categories. Center: in discrimination, this yields optimal decision and confidence criteria that are based on relative evidence (distance from the main diagonal), irrespective of sum evidence. Right: in detection, this yields optimal decision and confidence that are based on a non-linear interaction between relative and sum evidence. The third circle centred at (0,0) represents the two-dimensional distribution of percepts in the absence of stimuli.

To account for this apparently irrational discounting of incongruent evidence in confidence formation, Maniscalco, Peters, and Lau (2016) point out that outside of a lab setting, representational spaces are so high-dimensional that keeping track of evidence for every possible stimulus category is not feasible. For example, to be confident that an object is an apple, one would have to negatively weigh evidence for this object being an orange, a banana, a book and a ferret, among infinitely many other unsupported hypotheses. To resolve this engineering challenge, metacognitive systems may have evolved to positively weigh evidence for the chosen hypothesis, while ignoring conflicting evidence. Such a strategy is reasonable, as in Signal Detection space, samples that are farther away from the origin (high visibility) are on average farther away from the discrimination criterion (high discriminability). This strategy is then carried over to the lab, where decisions are made in low-dimensional representational spaces, and where keeping track of evidence for the two alternative stimulus categories is in fact feasible.

A more recent model identified the origin of this response-congruent heuristic not in this curse of dimensionality, but in the variance structure of perceptual evidence (Miyoshi & Lau, 2020). In a series of simulations, the authors augmented a two-dimensional Signal Detection model with realistic assumptions about the sensory encoding of signal and noise, most importantly that the variance of signal tends to be higher than that of noise. In these settings, a Response Congruent Evidence (RCE) heuristic provided more accurate confidence judgments, meaning ones that are more aligned with objective accuracy, than a Balance of Evidence (BE) heuristic. Again, this model implies that adopting a detection-like strategy when rating one’s confidence might have functional benefits for metacognition.

Notably, both models imply a link between confidence in discrimination, and detection judgments about the presence or absence of a stimulus. In a detection setting where multiple possible target stimuli can appear, the likelihood ratio between stimulus presence and absence is more sensitive to evidence for the detected stimulus (positive evidence) compared to evidence for the absence of other, undetected stimuli (negative evidence; see Fig. 1, right panel). Accordingly, recent studies have found that discrimination confidence is detection-like (Rausch, Hellmann, & Zehetleitner, 2018). Perhaps surprisingly, however, there has been limited focus on the complementary question: do detection decisions share features of discrimination confidence, such as a positive evidence bias? In other words, when faced with a detection task where targets are drawn from two stimulus classes, would detection decisions be sensitive to stimulus visibility (like discrimination confidence is), to the stimulus discriminability, or to both? Moreover, little is known about the properties of *detection* (rather than discrimination) confidence: would confidence in the presence of a target stimulus be susceptible to the same positive evidence bias as confidence in stimulus category? Finally, would detection confidence be sensitive to some form of positive evidence bias not only in decisions about target presence, but also in decisions about target absence?

To examine these questions, we conducted three experiments: one lab-based (N=10, 1800 trials per participant) and two online (N=102 and N=100, 112 and 168 trials per participant, respectively). In all experiments participants made discrimination and detection decisions about noisy stimuli, and rated their confidence in these decisions. Using reverse correlation analysis, we measured the influence of random fluctuations in stimulus energy on both responses and confidence ratings, and tested for the existence of processing asymmetries between ‘present’ and ‘absent’ responses in response time, general confidence, and metacognitive sensitivity (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2021; Mazor, Friston, & Fleming, 2020; Mazor, Moran, & Fleming, 2021; Meuwese, Loon, Lamme, & Fahrenfort, 2014). In all three experiments, we replicated previous findings of a positive evidence bias in confidence in discrimination decisions (although only marginally in Exp. 3; Zylberberg, Barttfeld, and Sigman, 2012). In contrast, our understanding of decision and confidence formation in detection evolved and changed following each experiment, as evident in our pre-registration documents. When considering the results of all three experiments together, we conclude that, similar to discrimination confidence, detection decisions are also sensitive to a positive evidence bias (we use the word bias here to mean a deviation from equal weighting of evidence for the two stimulus categories, and not in the sense of a deviation from rationality). Detection confidence presented a more complex pattern of results, with a positive evidence bias in one out of three experiments. We discuss our findings with respect to recent theoretical proposals regarding the origin of a positive evidence bias in discrimination confidence.

# 2 Experiment 1

## 2.1 Methods

### 2.1.1 Participants.

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 10 participants were recruited via the UCL’s psychology subject pool, and gave their informed consent prior to their participation. Each participant performed four sessions of 600 trials each, in blocks of 100 trials. Sessions took place on different days and consisted of 3 discrimination blocks interleaved with 3 detection blocks.

### 2.1.2 Experimental procedure.

The experimental procedure for Exp. 1 largely followed the procedure described in Zylberberg, Barttfeld, and Sigman (2012), Exp. 1. Participants observed a random-dot kinematogram for a fixed duration of 700 ms. In discrimination trials, the direction of motion was one of two opposite directions with equal probability, and participants reported the observed direction by pressing one of two arrow keys on a standard keyboard. In detection blocks participants reported whether there was coherent motion by pressing one of two arrow keys on a standard keyboard. In half of the detection trials dots moved coherently to one of two opposite directions, and in the other half they moved randomly.

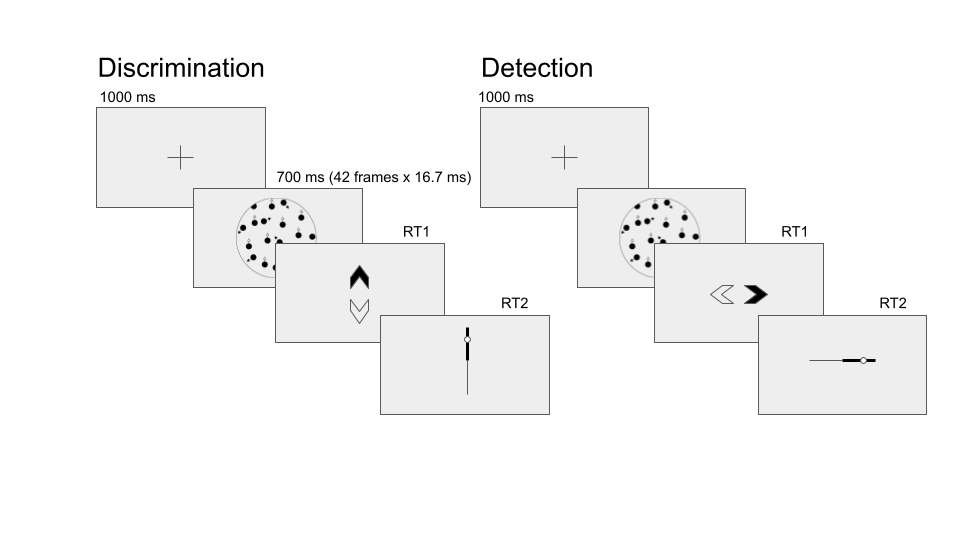
In both detection and discrimination blocks, participants indicated their confidence following each decision. Confidence was reported on a continuous scale ranging from chance to complete certainty. To avoid systematic response biases affecting confidence reports, the orientation (vertical or horizontal) and polarity (e.g., right or left) of the scale was set to agree with the type 1 response. For example, following an up arrow press, a vertical confidence bar was presented where ‘guess’ is at the center of the screen and ‘certain’ appeared at the upper end of the scale (see Fig. 2).

To control for response requirements, for five subjects the dots moved to the right or to the left, and for the other five subjects they moved upward or downward. The first group made discrimination judgments with the right and left keys and detection judgments with the up and down keys, and this mapping was reversed for the second group. The number of coherently moving dots (‘motion coherence’) was adjusted to maintain performance at around 70% accuracy for detection and discrimination tasks independently. This was achieved by measuring mean accuracy after every 20 trials, and adjusting coherence by a step of 3% if accuracy fell below 60% or went above 80%.

Stimuli for discrimination blocks were generated using the exact same procedure reported in Zylberberg, Barttfeld, and Sigman (2012)[[1]](#footnote-1). Trials started with a presentation of a fixation cross for one second, immediately followed by stimulus presentation. The stimulus consisted of 152 white dots (diameter = ), presented within a circular aperture centered on the fixation point for 700 milliseconds (42 frames, frame rate = 60 Hz). Dots were grouped in two sets of 76 dots each. Every other frame, the dots of one set were replaced with a new set of randomly positioned dots. For each coherence value of , a proportion of of the dots from the second set moved coherently in one direction by a fixed distance of , while the remaining dots in the set moved in random directions by a fixed distance of . On the next update, the sets were switched, to prevent participants from tracing the position of specific dots. Frame-specific coherence values were sampled for each screen update from a normal distribution centred around the coherence value with a standard deviation of 0.07, with the constraint that must be a number between 0 and 1.

Stimuli for detection blocks were generated using a similar procedure, with the only difference being that on a random half of the trials coherence was set to 0%, without random sampling of coherence values for different frames (see Fig. 1).

To probe global metacognitive estimates of task performance, at the end of each experimental block (100 trials) participants estimated the number of correct responses they have made. Analysis of these global metacognitive estimates is provided in Appendix A.4

*Figure* *2.*  Task design for Experiment 1. In both discrimination and detection blocks, participants viewed 700 milliseconds of a random dot motion array, after which they made a keyboard response to indicate their decision (motion direction in discrimination, signal absence or presence in detection), followed by a continuous confidence report using the mouse. 5 participants viewed vertically moving dots and indicated their detection responses on a horizontal scale, and 5 participants viewed horizontally moving dots and indicated their detection responses on a vertical scale.

### 2.1.3 Randomization.

The order and timing of experimental events was determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized in a way that ensures registration time-locking (Mazor, Mazor, & Mukamel, 2019).

### 2.1.4 Analysis.

Experiment 1 was pre-registered (pre-registration document is available here: <https://osf.io/z2s93/>). Our full pre-registered analysis is available in the Appendix.

#### Reverse correlation analysis.

For the reverse correlation analysis, we followed a procedure similar to the one described in Zylberberg, Barttfeld, and Sigman (2012). For each of the four directions (right, left, up and down), we applied two spatiotemporal filters to the frames of the dot motion stimuli as described in previous studies (Adelson & Bergen, 1985; Zylberberg, Barttfeld, & Sigman, 2012). The outputs of the two filters were squared and summed, resulting in a three-dimensional matrix with motion energy in a specific direction as a function of x, y, and time. We then took the mean of this matrix across the x and y dimensions to obtain an estimate of the overall temporal fluctuations in motion energy in the selected direction. Additionally, for every time point we extracted the variance along the x and y dimensions, to obtain a measure of temporal fluctuations in spatial variance. Using this filter, we obtained estimates of temporal fluctuations in the mean and variance of motion energy for upward, downward, leftward and rightward motion within each trial. Given a high correlation between our mean and variance estimates, we focused our analysis on the mean motion energy.

In order to distil random fluctuations in motion energy from mean differences between stimulus categories, we subtracted the mean motion energy from trial-specific motion energy vectors. The mean motion energy vectors were extracted at the group level, separately for each motion coherence level and as a function of motion direction. We chose this approach instead of the linear regression approach used by Zylberberg, Barttfeld, and Sigman (2012) in order to control for nonlinear effects of coherence on motion energy.

#### Statistical inference.

Statistics were extracted separately for each participant, and group-level inference was then performed on the first-order statistics. T-test Bayes factors were used to quantify the evidence for the null when appropriate, using a Jeffrey-Zellner-Siow Prior for the null distribution, with a unit prior scale (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

## 2.2 Results

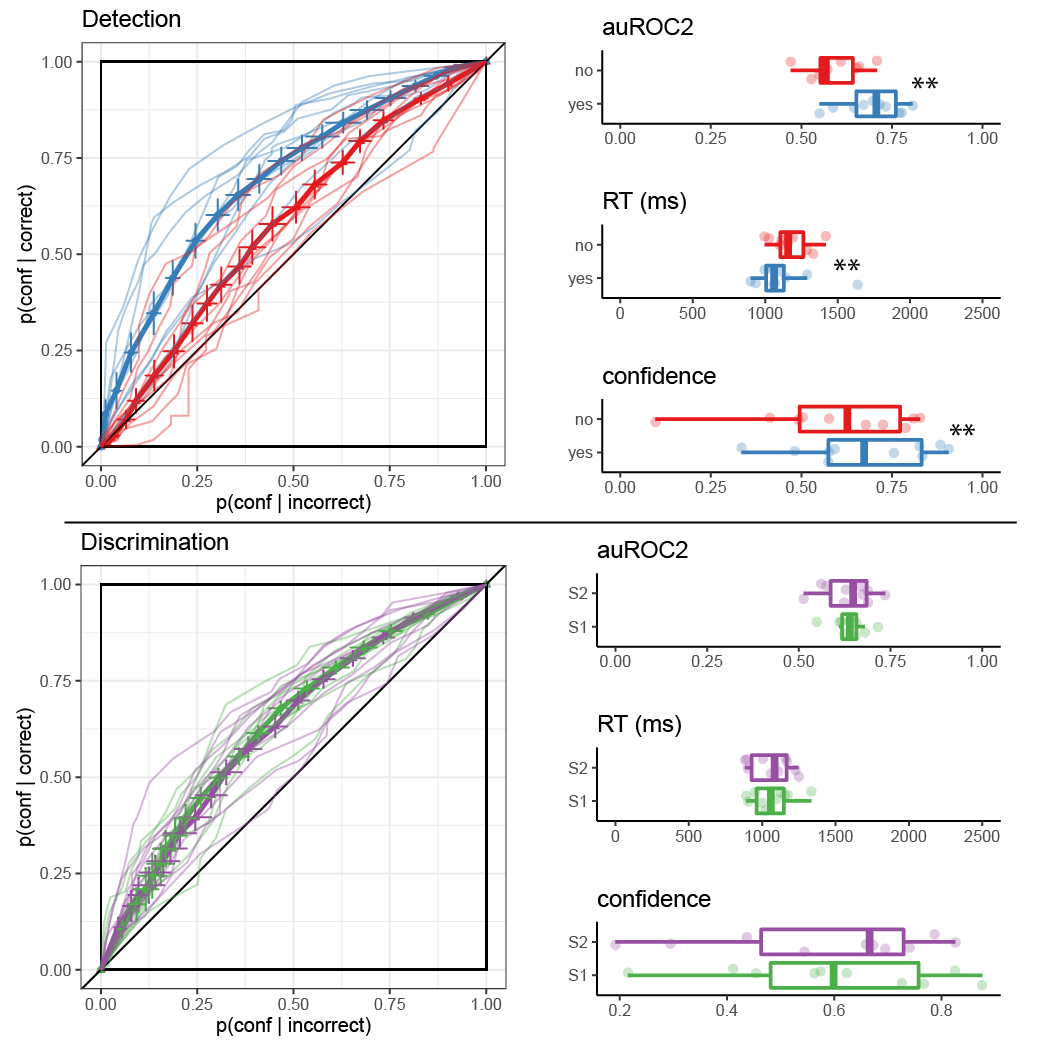
### 2.2.1 Response accuracy.

Overall proportion correct was 0.74 in the discrimination and 0.72 in the detection task. Performance for discrimination was significantly higher than for detection (, 95% CI , , , ). This difference in task performance reflected a slower convergence of the staircasing procedure for the discrimination task during the first session. When discarding all data from the first session and analyzing only data from the last three sessions (1800 trials per participant), task performance was equated between the two tasks at the group level (, 95% CI , , , ; ). In order to avoid confounding differences between discrimination and detection decision and confidence profiles with more general task performance effects, the first session was excluded from all subsequent analyses.

### 2.2.2 Overall properties of response time and confidence distributions.

In detection, participants were more likely to respond ‘yes’ than ‘no’ (mean proportion of ‘yes’ responses: , 95% CI , , , ). We did not observe a consistent response bias for the discrimination data (mean proportion of ‘rightward’ or ‘upward’ responses: , 95% CI , , , ).

Replicating previous studies (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2021; Mazor, Friston, & Fleming, 2020; Mazor, Moran, & Fleming, 2021; Meuwese, Loon, Lamme, & Fahrenfort, 2014), we find the typical asymmetries between detection ‘yes’ and ‘no’ responses in response time, overall confidence, and the alignment between subjective confidence and objective accuracy (also termed metacognitive sensitivity, here measured as the area under the response-conditional type 2 ROC curve; see Fig. 3). ‘No’ responses were slower compared to ‘yes’ responses (median difference: 85.37 ms), and accompanied by lower levels of subjective confidence (mean difference of 0.08 on a 0-1 scale). Metacognitive sensitivity was higher for detection ‘yes’ compared with detection ‘no’ responses (mean difference in area under the curve units: 0.11). No difference in response time, confidence, or metacognitive sensitivity was found between the two discrimination responses. For a detailed statistical analysis of these behavioural asymmetries see Appendix A.1.



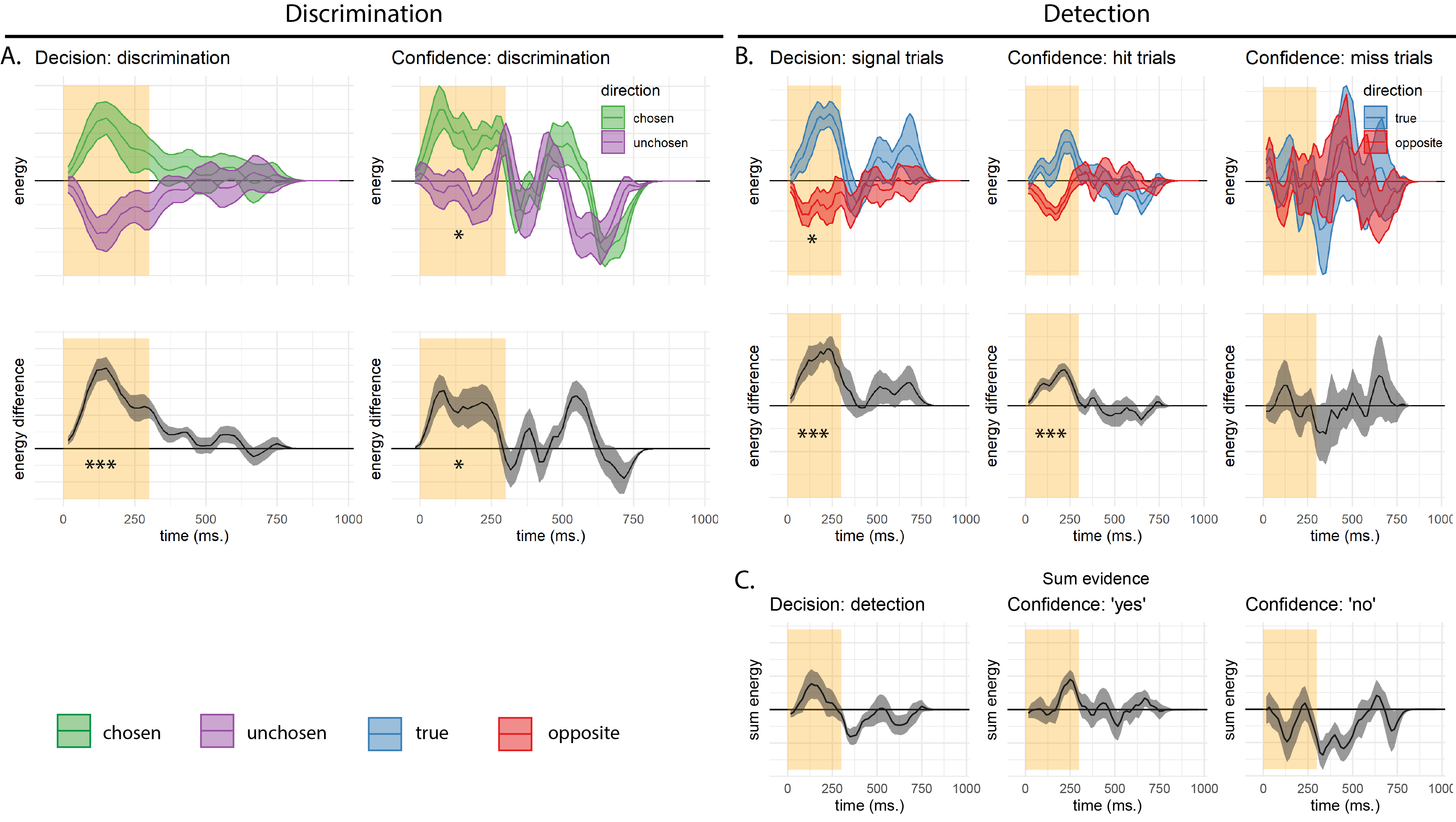
*Figure* *3.*  Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence in detection (upper panel) and discrimination (lower panel), in Exp. 1. Left: Response conditional type 2 ROC curves for the two tasks and four responses in Exp. 1. The area under the type 2 ROC curve is a measure of metacognitive sensitivity, and the difference in areas between the two responses a measure of metacognitive asymmetry. Single-subject curves are presented in low opacity. Right: distributions of the area under the type 2 ROC curve, median response time, and mean confidence for the four responses, across participants. Box edges and central lines represent the 25, 50 and 75 quantiles. Whiskers cover data points within four inter-quartile ranges around the median. Stars represent significance in a two-sided t-test: \*\*: p<0.01, \*\*\*: p<0.001

### 2.2.3 Reverse Correlation.

Random fluctuations in motion energy made it possible to apply reverse correlation to test which stimulus features are incorporated into decisions and confidence ratings in detection and discrimination. Following Zylberberg, Barttfeld, and Sigman (2012), our statistical analysis focused on the first 300 milliseconds after stimulus onset.

#### Discrimination.

Using reverse correlation analysis we quantified the effect of random fluctuations in motion energy on the probability of responding ‘right’ and ‘left’ (or ‘up’ and ‘down’), and on the temporal dynamics of decision formation. Similar to the results obtained by Zylberberg, Barttfeld, and Sigman (2012), participants’ decisions were sensitive to motion energy fluctuations during the first 300 milliseconds of the trial (, ; see Fig. 4A, left panels). Note that the green and purple lines are mathematically bound to be symmetric due to the demeaning procedure. To test for a potential asymmetry in evidence weighting in discrimination decisions, we contrasted the contribution of motion energy in the true and opposite directions of motion (defined with respect to the stimulus, and independently of decision). Fluctuations in motion energy in both directions contributed significantly to discrimination decisions (, ), with no significant difference between them (, ). In other words, positive and negative evidence equally contributed to discrimination decisions, even when defined independently of the decision.



*Figure* *4.*  Reverse correlation, Exp. 1. A: Decision and confidence discrimination kernels. Upper left: motion energy in the chosen (green) and unchosen (purple) direction as a function of time. Lower left: a subtraction between energy in the chosen and unchosen directions. Upper right: confidence effects for motion energy in the chosen (green) and unchosen (purple) directions. Lower right: a subtraction between confidence effects in the chosen and unchosen directions. B: Decision and confidence detection kernels in signal trials. Upper left: difference in motion energy between ‘yes’ and ‘no’ responses in the true (blue) and opposite (red) directions as a function of time. Upper middle and right: confidence effects for motion energy in the true and opposite directions for ‘yes’ and ‘no’ responses, respectively. Lower panels: the subtraction of decision and confidence kernels for the true and opposite directions. C: Decision and confidence detection kernels. Left: difference in sum motion energy between detection ‘yes’ and ‘no’ responses. Middle and right: difference in sum motion energy between high and low confidence trials in ‘yes’ and ‘no’ responses. Shaded areas represent the mean one standard error. The first 300 milliseconds of the trial are marked in yellow. Stars represent significance in a two-sided t-test: \*: p<0.05, \*\*: p<0.01, \*\*\*: p<0.001. In the upper rows of panels A and B, stars represent the significance of a positive evidence bias in evidence weighting.

We then turned to the contribution of motion energy to subjective confidence ratings. The median confidence rating in each experimental session was used to split all motion energy vectors into four groups, according to decision (chosen or unchosen directions) and confidence level (high or low). Confidence kernels for the chosen and unchosen directions were then extracted by subtracting the mean low confidence vectors from the mean high confidence vectors for both the chosen and unchosen directions. We observed a significant effect of motion energy on confidence within the first 300 milliseconds of the trial (, ; see Fig. 4A, right panels). Furthermore, confidence ratings in the discrimination task were more sensitive to motion energy in the chosen direction (positive evidence) than to motion energy in the opposite direction (negative evidence; , ). This is a replication of the Positive Evidence Bias observed in Zylberberg, Barttfeld, and Sigman (2012).

#### Detection.

Carrying out an analogous reverse correlation analysis for detection introduces a challenge: while ‘no’ responses reflect a belief in the absence of any coherent motion, ‘yes’ responses can result from detection of any type of coherent motion going in either direction (or both). We chose to have two possible motion directions in the detection task in order to prevent participants from making ‘no’ responses based on significant motion in an unexpected direction. While this choice ensured that participants cannot trivially accumulate evidence for absence, it also made the reverse correlation analysis more difficult, as we did not have full access to participants’ beliefs about the stimulus when they responded ‘yes.’

As a first approximation, we tested whether sum motion energy along the relevant dimension (horizontal or vertical), regardless of direction (up/down or left/right), affected the probability of a ‘yes’ response. Sum motion energy did not have a significant effect on participants’ responses during the first 300 milliseconds (, ; see Fig. 4C, left panel) or at any other time point. The effect of sum motion energy on decision confidence during the first 300 milliseconds was positive and marginally significant (, ; see 4C, middle and right panels). Response-specific effects of sum motion energy on decision confidence were not significant for either response.

#### Detection signal trials.

A lack of effect of sum motion energy on detection decisions and confidence may be due to the fact that participants were sensitive to relative evidence (e.g., ‘more dots are moving to the right’) rather than to the sum motion along the relevant axis. However, as described above, on any single trial, we cannot tell whether a ‘yes’ response means ‘I perceived coherent motion to the right’ or ‘I perceived coherent motion to the left.’ Instead, in order to approximate participants’ belief states during ‘yes’ responses, we focused only on trials in which coherent motion was presented in one of the two directions (signal trials). In these trials, we reasoned that a ‘yes’ response is most likely to reflect the detection of the true direction of motion. We then asked whether fluctuations in the true and opposite directions of motion contributed to detection decision and confidence. This was done by subtracting the motion energy vectors for ‘yes’ and ‘no’ responses in the true and opposite motion directions.

Similar to discrimination decisions, detection decisions were sensitive to perceptual evidence in the first 300 milliseconds of the trial (see Fig. 4B, left panels). However, in contrast to discrimination, an asymmetric evidence weighting was apparent in the decision itself: when deciding whether a stimulus contained coherent motion, participants were more sensitive to fluctuations in motion energy that strengthened the true direction of motion, in comparison to fluctuations that weakened motion in the opposite direction (, ).

Motion fluctuations in the first 300 milliseconds of the trial also contributed to confidence in detection ‘yes’ responses (contrasting high and low confidence hit trials; , ). However, unlike in the discrimination task, here we found no positive evidence bias in confidence ratings (, ; see Fig. 4B, middle panels). To reiterate, while detection decisions were mostly sensitive to fluctuations in motion energy toward the true direction of motion, confidence in detection ‘yes’ responses was equally sensitive to fluctuations in the true and opposite directions of motion. Confidence in ‘miss’ trials was independent of motion energy (, ). This was true both for motion energy in the true direction of motion (, ) as well as for motion energy in the opposite direction (, ). However, and to anticipate the results of Exp. 3 presented below, we note that this equal weighting of positive and negative evidence in detection confidence was not replicated in a subsequent experiment designed to directly test this surprising result.

# 3 Experiment 2

In Exp. 1, we replicated previous observations of a positive evidence bias in discrimination confidence, such that evidence in support of a decision was given more weight in the construction of confidence than evidence against it. In contrast, in detection a positive evidence bias was apparent for the decision, but not for the confidence kernels. Equal weighting of positive and negative evidence suggests that detection confidence followed not sum evidence (visibility), but relative evidence (discriminability). Furthermore, confidence in detection ‘no’ responses was not at all affected by fluctuations in motion energy.

In Exp. 2 we tested the robustness of these findings by employing a different type of stimuli (flickering patches) and mode of data collection (a ~10 minute online experiment). Our pre-registered objectives (documented here: [https://osf.io/d3vkm/](https://osf.io/8u7dk/)) were 1) to replicate a positive evidence bias in discrimination confidence, 2) to replicate the absence of a positive evidence bias in detection confidence, 3) to replicate the absence of an effect of either positive or negative evidence on confidence in ‘no’ judgments.

## 3.1 Methods

### 3.1.1 Participants.

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 147 participants were recruited via Prolific (prolific.co) and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. Following our pre-registration, we aimed to collect data until we had reached 100 included participants based on our pre-specified inclusion criteria (see [https://osf.io/d3vkm/](https://osf.io/8u7dk/)). Our final data set includes observations from 102 included participants. The entire experiment took around 10 minutes to complete. Participants were paid £1.25 for their participation, equivalent to an hourly wage of £7.5.

### 3.1.2 Experimental paradigm.

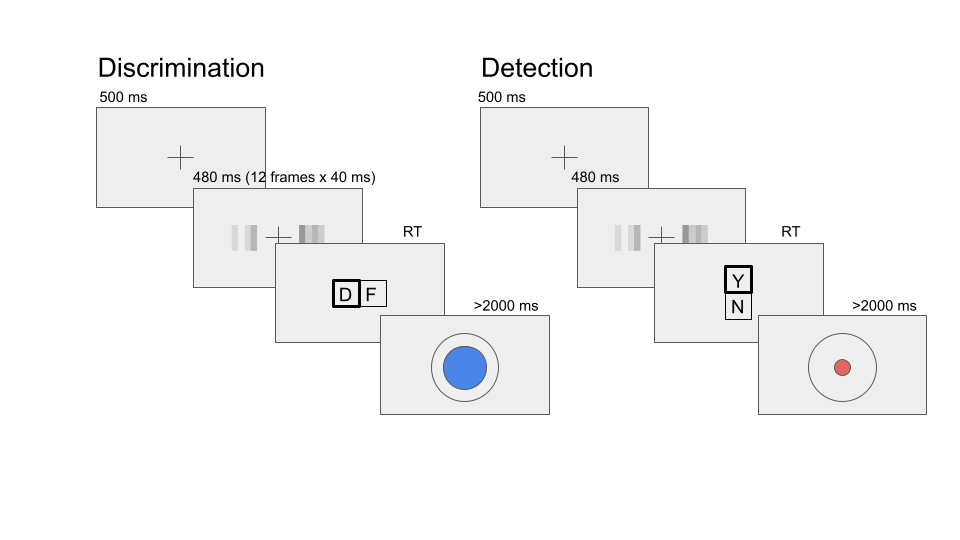
The experiment was programmed using the jsPsych and P5 JavaScript packages (De Leeuw, 2015; McCarthy, 2015), and was hosted on a JATOS server (Lange, Kuhn, & Filevich, 2015). It consisted of two tasks (Detection and Discrimination) presented in separate blocks. A total of 56 trials of each task was delivered in 2 blocks of 28 trials each. The order of experimental blocks was interleaved, starting with discrimination.

The first discrimination block started after an instruction section, which included instructions about the stimuli and confidence scale, four practice trials and four confidence practice trials. Further instructions were presented before the second block. Instruction sections were followed by multiple-choice comprehension questions, to monitor participants’ understanding of the main task and confidence reporting interface. To encourage concentration, feedback was delivered at the end of the second and fourth blocks about overall performance and mean confidence in the task.

Importantly, unlike the lab-based experiment, there was no calibration of difficulty for the two tasks. The rationale for this is that in Exp. 1 perceptual thresholds for motion discrimination were highly consistent across participants, and staircasing took a long time to converge. Furthermore, in Exp. 1 we aimed to control for task difficulty, but this introduced differences between the stimulus intensity in detection and discrimination. To complement our findings, here we aimed to match stimulus intensity between the two tasks, and accept that task performance might vary.

#### Trial structure.

In discrimination blocks, trial structure closely followed Exp. 2 from Zylberberg, Barttfeld, and Sigman (2012), with a few adaptations. Following a fixation cross (500 ms), two sets of four adjacent vertical gray bars were presented as a rapid serial visual presentation (RSVP; 12 frames, presented at 25Hz), displayed to the left and right of the fixation cross (see Fig. 5). On each frame, the luminance of each bar was randomly sampled from a Gaussian distribution with a standard deviation of 10/255 units in the standard RGB 0-255 coordinate system. For one set of bars, this Gaussian distribution was centered at the same luminance value as the background (128/255). For the other set, it was centered at 133/255, making it brighter on average. Participants then reported which of the two sets was brighter on average using the ‘D’ and ‘F’ keys on the keyboard. After their response, they rated their confidence on a continuous scale, by controlling the size of a colored circle with their mouse. High confidence was mapped to a big, blue circle, and low confidence to a small, red circle. To discourage hasty confidence ratings, the confidence rating scale stayed on the screen for at least 2000 milliseconds. Feedback about response accuracy was delivered after the confidence rating phase.



*Figure* *5.*  Task design for Experiment 2. In both tasks, participants viewed 480 milliseconds of two flickering patches, after which they made a keyboard response to indicate which of the patches was brighter (discrimination) or whether any of the patches was brighter than the background (detection).

Detection blocks were similar to discrimination blocks, with the exception that decisions were made about whether the average luminance of either of the two sets was brighter than the gray background, or not. In ‘different’ trials, the luminance of the four bars in one of the sets was sampled from a Gaussian distribution with mean 133/255, and the luminance of the other set from a Gaussian distribution with mean 128/255. In ‘same’ trials, the luminance of both sets was sampled from a distribution centered at 128/255. Decisions in Detection trials were reported using the ‘Y’ and ‘N’ keys. Confidence ratings and feedback were as in the discrimination task.

### 3.1.3 Randomization.

The order and timing of experimental events was determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized in a way that ensures registration time-locking (Mazor, Mazor, & Mukamel, 2019).

## 3.2 Results

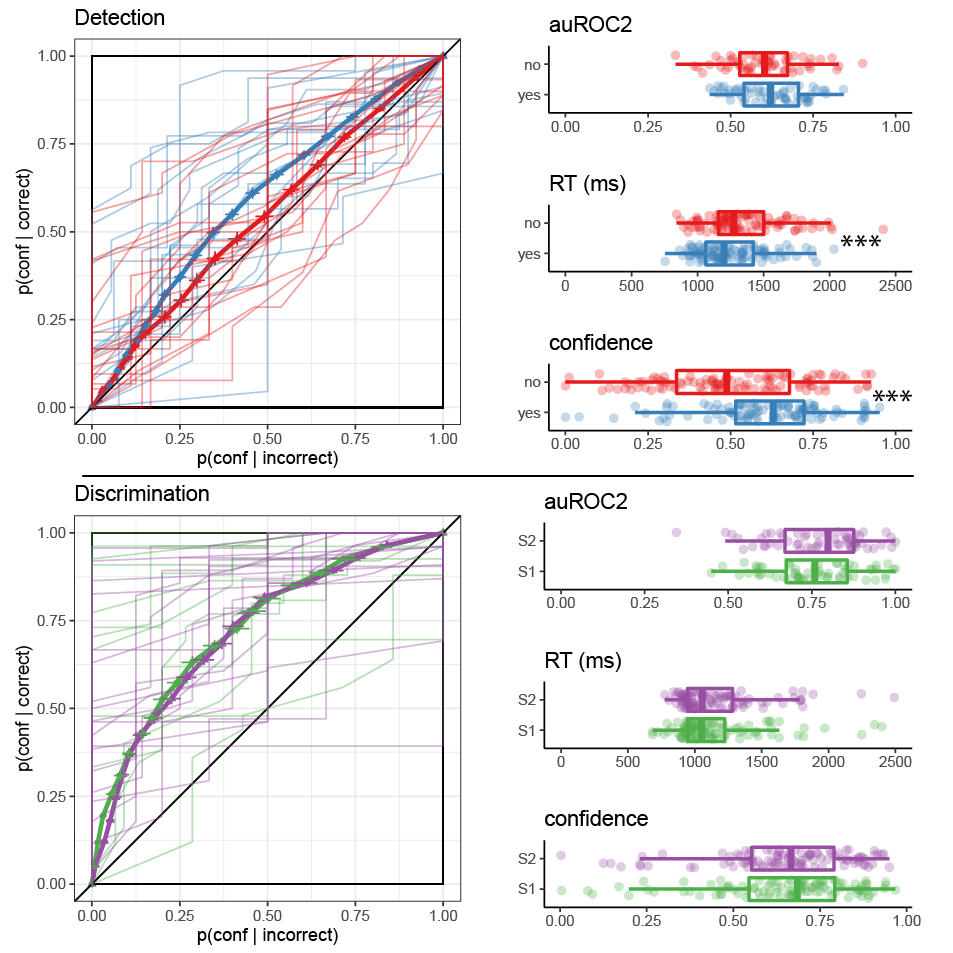
### 3.2.1 Response accuracy.

Overall proportion correct was 0.85 in the discrimination and 0.67 in the detection task. Performance for discrimination was significantly higher than for detection (, 95% CI , , , ). Unlike in Exp. 1, where we aimed to control for task difficulty, here we decided to match stimulus intensity between the two tasks, so a difference between detection and discrimination performance was expected (Wickens, 2002, p. 104).

### 3.2.2 Overall properties of response and confidence distributions.

Similar to Exp. 1, participants were more likely to respond ‘yes’ than ‘no’ in the detection task (mean proportion of ‘yes’ responses: , 95% CI , , , ). We did not observe a consistent response bias in discrimination (mean proportion of ‘right’ responses: , 95% CI , , , ).

As in Exp. 1, we also found behavioural asymmetries between the two detection responses (see Fig. 6), with ‘yes’ responses being faster (median difference of 77.12 ms) and accompanied by higher levels of confidence (mean difference of 0.10 on a 0-1 scale). Unlike in Exp. 1, here we found no evidence for a difference in metacognitive sensitivity between ‘yes’ and ‘no’ responses (mean difference of 0.02 in AUC units). No asymmetries were observed between the two discrimination responses. For a detailed statistical analysis see Appendix B.1.



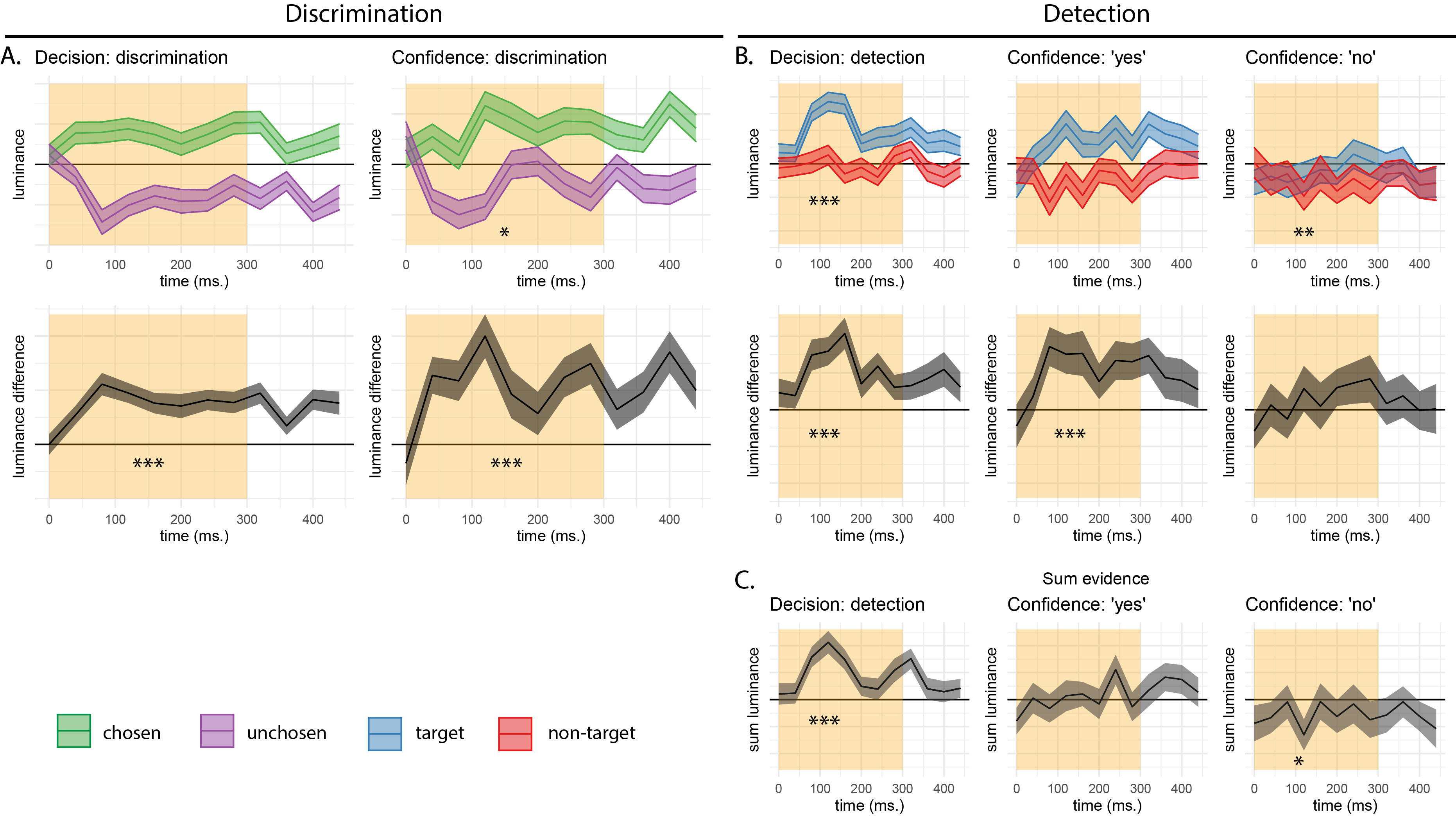
*Figure* *6.*  Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence, in Exp. 2. Same conventions as in Fig. 3.

### 3.2.3 Reverse Correlation.

Stimuli in Exp. 2 consisted of two flickering patches, each comprising 4 gray bars presented for 12 frames. Together, this summed to 96 random luminance values per trial, which we subjected to reverse correlation analysis, following the analysis procedure of Exp 2. in Zylberberg, Barttfeld, and Sigman (2012).

#### Discrimination decisions.

First, we asked whether random fluctuations in luminance had an effect on participants’ discrimination responses. Similar to the results obtained by Zylberberg et. al., discrimination decisions were sensitive to fluctuations in luminance during the first 300 milliseconds of the trial (, ; see Fig. 7A, left panels). As per our approach to the reverse correlation analysis of Exp. 1, in order to test for asymmetries in evidence weighting in the decision stage we classify positive and negative evidence independently of participants’ decision. When sorting evidence based on the location of the true signal, participants’ decisions were surprisingly more sensitive to fluctuations in luminance in the non-signal compared with the signal set of bars within the first 300 milliseconds of the trial (, ). Importantly, this asymmetry (effectively, a negative evidence bias) is in the opposite direction to what we later find in discrimination confidence and detection decisions, and was not replicated in Exp. 3.



*Figure* *7.*  Reverse correlation, Exp. 2. A: Decision and confidence discrimination kernels. Upper left: mean luminance in the target (chosen) and non-target (unchosen) flickering patches as a function of time. Lower left: subtraction between luminance in the chosen and unchosen flickering patches. Upper right: confidence effects for luminance in the chosen (green) and unchosen (purple) flickering patches. Lower right: subtraction between confidence effects in the chosen and unchosen patches. B: Decision and confidence detection kernels in signal trials. Upper left: difference in luminance between ‘yes’ and ‘no’ responses in the target (blue) and non-target (red) patches as a function of time. Upper middle and right: confidence effects for luminance in the target and non-target patches for ‘yes’ and ‘no’ responses, respectively. Lower panels: subtraction between decision and confidence kernels for target and non-target patches. C: Decision and confidence detection kernels using sum luminance. Left: difference in sum luminance between detection ‘yes’ and ‘no’ responses. Middle and right: difference in sum luminance between high and low confidence trials in ‘yes’ and ‘no’ responses. Shaded areas represent the mean one standard error. The first 300 milliseconds of the trial are marked in yellow. Stars represent significance in a two-sided t-test: \*: p<0.05, \*\*: p<0.01, \*\*\*: p<0.001. In the upper rows of panels A and B, stars represent the significance of a positive evidence bias in evidence weighting.

#### Discrimination confidence.

We observed a significant effect of luminance on confidence within the first 300 milliseconds of the stimulus (, ; see Fig. 7A, right panels). Replicating Zylberberg, Barttfeld, and Sigman (2012), this effect was significantly stronger for luminance in the chosen stimulus, compared to the unchosen one (, ), consistent with a positive evidence bias.

#### Detection decisions.

We pooled luminance values from both right and left stimuli and contrasted the resulting values as a function of detection response. Unlike in Exp. 1, here sum luminance had a significant effect on participants’ detection responses during the first 300 milliseconds (, ; see Fig. 7C, left panel). This sensitivity of detection responses to sum evidence (visibility) is expected from a model in which detection is rationally based on the likelihood ratio between signal presence and absence (see Fig. 1).

We then asked if overall luminance had an effect on detection confidence, such that participants are more confident in their ‘yes’ responses for brighter displays, and more confident in their ‘no’ responses for darker displays. Interestingly, and in contrast with our hypothesis and to what is expected based on the likelihood ratio, sum luminance had no effect on decision confidence in ‘yes’ responses (, ), but had a significant negative effect on confidence in ‘no’ responses (, ; see Fig. 7B, middle and right panels). However, to again anticipate our pre-registered Exp. 3, when using the same type of stimuli with higher statistical power, we find an effect of sum luminance on confidence in both ‘yes’ and ‘no’ detection responses.

#### 3.2.3.1 Detection signal trials.

We next focused on detection signal trials only (i.e., hits and misses). This analysis diverged from our pre-registered plan; for the pre-registered analysis, please see Appendix D). In these trials, we could separate stimuli into a signal channel (the brighter, target stimulus) and a noise channel (the darker, non-target stimulus), and ask how random fluctuations in luminance in each channel affected detection decisions and confidence. As in Exp. 1, and in line with what is expected based on a likelihood ratio decision rule, a positive evidence bias effect in detection was apparent in the decision itself: when deciding whether one of the flickering patches was brighter, participants were more sensitive to positive noise in the brighter patch than to negative noise in the darker patch (, ; see Fig. 7B, left panels).

Random fluctuations in the luminance of the target stimulus relative to the non-target stimulus contributed positively to confidence in detection ‘yes’ responses (hits; , ; see Fig. 7B, middle panels). In contrast, and consistent with the results of Exp. 1, confidence in ‘miss’ trials was independent of the difference in luminance between the target and non-target stimuli (, ). Higher confidence in missing a signal was instead associated with lower overall luminance in both stimuli, in line with our observation that confidence in detection ‘no’ responses was based on the overall darkness of the display (visibility), rather than on relative evidence (discriminability). A negative effect of luminance on confidence in ‘no’ responses was significant for the non-target stimulus (, ), and marginally significant for the target stimulus (, ; see Fig. 7B, right panels). Finally, and similar to the results of Exp. 1, detection confidence was not susceptible to a positive evidence bias (, ). Exp. 3 below was designed to replicate this surprising symmetric weighting of positive and negative evidence in detection confidence (the absence of a positive evidence bias) in a high-powered design.

# 4 Experiment 3

In Exp. 3 we aimed to replicate our findings using an experimental manipulation in addition to employing reverse-correlation analysis. Our pre-registered objectives (see our pre-registration document: <https://osf.io/hm3fn/>) were 1) to replicate a positive evidence bias in discrimination, 2) to replicate a positive evidence bias in detection decisions, 3) to replicate the absence of a positive evidence bias in detection confidence.

## 4.1 Methods

### 4.1.1 Participants.

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 173 participants were recruited via Prolific (prolific.co), and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. Following our pre-registration, we aimed to collect data until we had reached 100 included participants based on our pre-specified inclusion criteria (see <https://osf.io/hm3fn/>). Our final data set includes observations from 100 included participants. The entire experiment took around 20 minutes to complete. Participants were paid £2.5 for their participation, equivalent to an hourly wage of £7.5.

### 4.1.2 Experimental paradigm.

Experiment 3 was identical to Experiment 2 with two changes. First, on half of the trials (*high-luminance* trials) the luminance of both sets of bars was increased by 2/255 for the entire duration of the display, thereby increasing sum evidence without affecting relative evidence on these trials. Second, in order to increase our statistical power for detecting response-specific effects in detection, participants performed four detection blocks and two discrimination blocks. Each block comprised 56 trials. The order of blocks was [detection, discrimination, detection, discrimination, detection, detection] for all participants.

## 4.2 Results

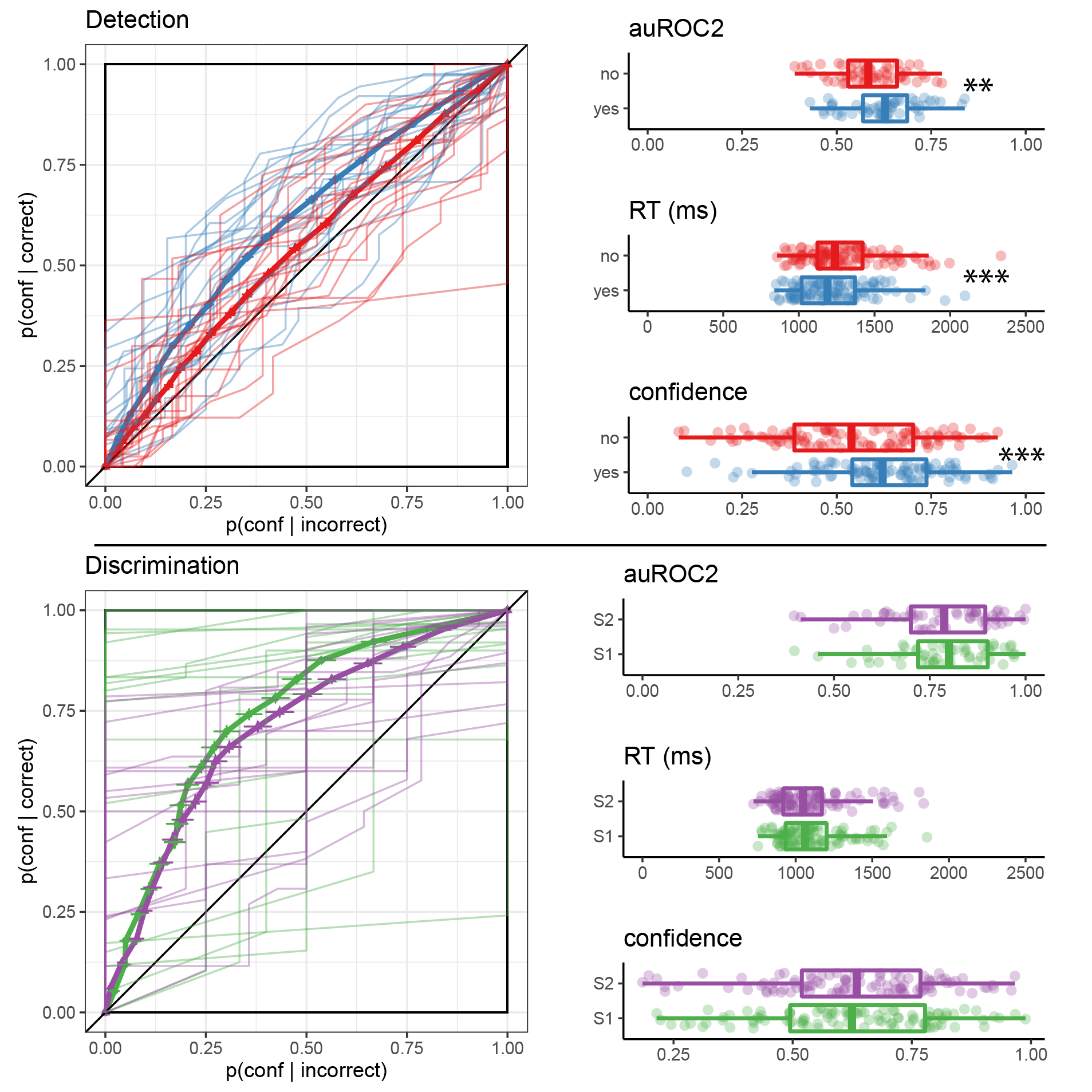
### 4.2.1 Response accuracy.

Overall proportion correct was 0.87 in the discrimination and 0.67 in the detection task. Performance for discrimination was significantly higher than for detection (, 95% CI , , , ), as expected.

### 4.2.2 Overall properties of response and confidence distributions.

Similar to Experiments 1 and 2, participants were more likely to respond ‘yes’ than ‘no’ in the detection task (mean proportion of ‘yes’ responses: , 95% CI , , , ). We did not observe a consistent response bias in discrimination (mean proportion of ‘right’ responses: , 95% CI , , , ).

As in both Experiments 1 and 2, we found behavioural asymmetries between the two detection responses, with ‘yes’ responses being faster (median difference of 68.90 ms), and accompanied by higher levels of confidence (mean difference of 0.09 on a 0-1 scale). As in Exp. 1, metacognitive sensitivity was higher for ‘yes’ than for ‘no’ responses (mean difference of 0.03 in AUC units). No asymmetries were observed between the two discrimination responses. For a detailed statistical analysis see Appendix C.1.



*Figure* *8.*  Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence, in Exp. 3. Same conventions as in Fig. 3.

### 4.2.3 Reverse correlation.

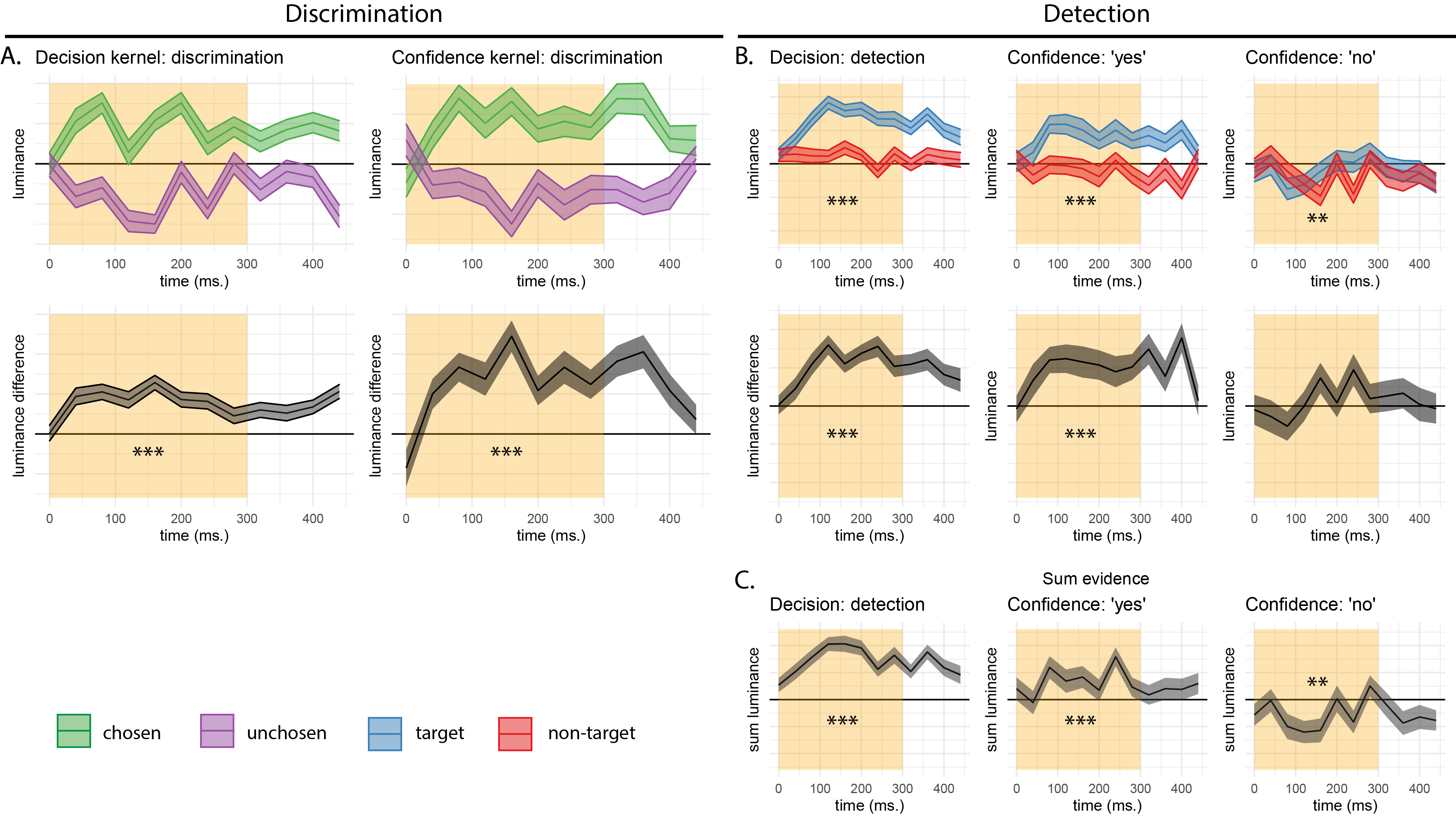
We first focused on reverse correlation analyses that collapsed across high-luminance and standard trials, in order to replicate the findings of Exps. 1 and 2. We note the results are qualitatively similar when including standard trials only, with the exception of confidence in detection ‘no’ responses (see Appendix C.2).

#### Discrimination decisions.

Discrimination decisions were sensitive to fluctuations in luminance during the first 300 milliseconds of the trial (, ; see Fig. 9A, left panels). As expected, we found no evidence for a positive evidence bias in discrimination decisions, even when grouping evidence based on the location of the true signal rather than subjects’ decisions (, ).

#### Discrimination confidence.

Luminance within the first 300 milliseconds had a significant effect on confidence ratings (, ; see Fig. 9A, right panels). A positive evidence bias for discrimination confidence was only marginally significant in this sample (, ).



*Figure* *9.*  Decision and confidence discrimination kernels, Experiment 3. Same conventions as Fig. 4.

#### 4.2.3.1 Detection.

Similar to Exp. 2, sum luminance had a significant effect on participants’ detection responses during the first 300 milliseconds (, ; see Fig. 9C, left panel). Recall that a surprising finding in Exp. 1 and 2 was that sum evidence (motion energy or luminance) had no effect on participants’ confidence in their judgments of stimulus presence. In contrast, in Exp. 3 sum luminance had a significant positive effect on decision confidence when reporting target presence (‘yes’ responses; , ), and a significant negative effect on confidence when reporting target absence (‘no’ responses; , ; see Fig. 9C, middle and right panels).

#### 4.2.3.2 Detection signal trials.

As in Exp. 2, here we also asked how random variability in luminance in the target (brighter) and non-target (darker) channels affected detection decisions and confidence. When deciding whether one of the two flickering patches was brighter than the background, participants were sensitive to positive fluctuations in the target patch more than to negative fluctuations in the non-target patch (, ), consistent with a positive evidence bias in detection decisions and replicating findings from Exps. 1 and 2. Random fluctuations in luminance in the first 300 milliseconds of the trial also contributed to confidence in detection ‘yes’ responses (hits; , ). Importantly, however, and in contrast to the results of Exp. 1 and 2, confidence in ‘yes’ responses was more sensitive to positive evidence than to negative evidence (i.e., the relative darkness of the non-target patch; , ). Together these results are consistent with a positive evidence bias not only for detection decisions, but also for detection confidence.

Similar to our findings in Exp. 2, higher confidence in ‘no’ responses (‘misses’) was associated with lower luminance values in both patches. This is consistent with our observation that confidence in judgments about stimulus absence is based on the overall darkness of the display. This negative effect of luminance on confidence in ‘no’ responses was significant for the non-target stimulus (, ), and marginally significant for the target stimulus (, ). Confidence in misses was independent of the contrast in luminance between the target and non-target patches (, ).

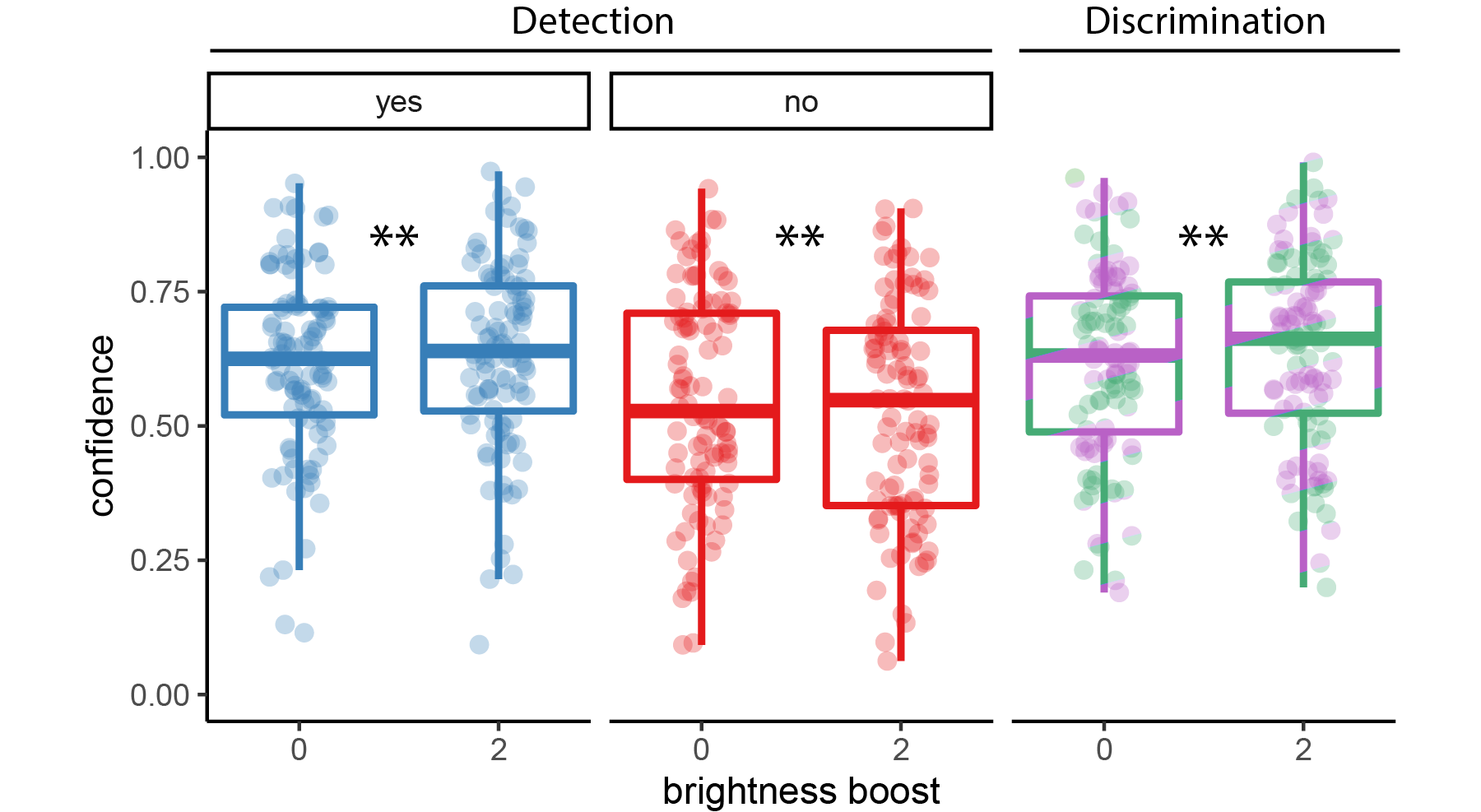
### 4.2.4 Evidence-weighting.

In Experiments 1 and 2, confidence in stimulus presence was invariant to sum evidence (overall motion energy in Exp.1, sum luminance in Exp. 2). This was surprising for two reasons. First, in both cases sum motion energy did have a significant effect on detection decisions. Second, incorporating information about sum evidence into confidence in the presence of a stimulus is rational: a bright target stimulus is more likely to be present when both target *and* non-target stimuli are bright compared to when both are dark. As we document above, however, the counterintuitive findings of Exps. 1 and 2 did not replicate in Exp. 3, where detection confidence was now sensitive to the overall brightness of the display.

To shed further light on this issue, in Exp. 3 half of the trials consisted of slightly brighter stimuli, thereby increasing statistical power for tests of a positive evidence bias. Participants were more likely to respond ‘yes’ on higher- compared to lower-luminance trials (, 95% CI , , , ), consistent with overall luminance providing a valid cue for signal presence.

We next turned to the effects of our luminance manipulation on confidence. For discrimination judgments, participants were also more confident in higher- compared to lower- luminance trials (, 95% CI , , , ), replicating a positive evidence bias for discrimination confidence. For detection judgments, in line with the reverse correlation analysis of Exp. 3 (and in contrast to the findings of Experiments 1 and 2), participants were more confident in their ‘yes’ responses when overall luminance was higher (, 95% CI , , , ). Our pre-registered Bayesian analysis provided strong evidence for the alternative hypothesis that detection confidence is affected by this manipulation (). Furthermore, this increase in ‘yes’ response confidence as a function of the brightness manipulation was not significantly different from that observed for discrimination confidence (, 95% CI , , , ). Finally, and in line with Exp. 2, overall luminance had a significant negative effect on confidence in ‘no’ responses (, 95% CI , , , ), indicating that participants were more confident in the absence of a target when overall luminance was lower.

In an exploratory analysis, we examined the correlation between the effects of our luminance manipulation on detection decisions and on confidence in the detection and discrimination tasks. Participants who showed a more pronounced decision PEB in detection also showed a more pronounced confidence PEB in detection ‘yes’ responses (, 95% CI , , , ), and marginally in detection ‘no’ responses (, 95% CI , , , ; note that this correlation is expected to be negative, because in detection ‘no’ responses high luminance corresponds to low confidence). In contrast, we find no significant association between a decision PEB in detection and a confidence PEB in discrimination (, 95% CI , , , ). A Bayes Factor analysis with a scale factor set to the correlation strength between detection decision and confidence PEBs provided moderate evidence against an association between a decision PEB in detection and a confidence PEB in discrimination (). Similarly, no significant correlation was observed between detection decision and discrimination confidence PEBs derived from the reverse correlation analysis (Exp. 2: , 95% CI , , , ; ; Exp. 3: , 95% CI , , , ; for a medium-narrow scale factor of ). Together, a PEB in detection decisions was related to a PEB in detection confidence, but not to a PEB in discrimination confidence.



*Figure* *10.*  Difference in confidence between standard and higher-luminance trials for the three response categories (detection ‘yes’ and ‘no’ responses, and discrimination responses) in Exp. 3. Box edges and central lines represent the 25, 50 and 75 quantiles. Whiskers cover data points within four inter-quartile ranges around the median. Stars represent significance in a two-sided t-test: \*\*: p<0.01

# 5 Discussion

In three experiments, we compared the perceptual drivers of decisions and confidence ratings in discrimination and detection, matched for difficulty (Exp. 1) and signal strength (Exp. 2 and 3). In order to measure the contribution of perceptual evidence to confidence in detection and discrimination judgments, we followed Zylberberg, Barttfeld, and Sigman (2012) and applied reverse correlation analysis to noisy stimuli in perceptual decision making tasks. Fig. 11 provides a concise summary of the findings from all three experiments. We fully replicated the main results of Zylberberg and colleagues: decisions and confidence were affected by perceptual evidence in the first 300 milliseconds of the trial, peaking at around 200 milliseconds. We also successfully replicated a positive evidence bias (PEB) for discrimination confidence: confidence in the discrimination task was more affected by supporting than by conflicting evidence in all experiments (although only marginally significant in Exp. 3). A PEB in discrimination confidence judgments may indicate that participants adopt a detection-like disposition in their metacognitive judgments, focusing on sum evidence rather than relative evidence when rating their confidence.



*Figure* *11.*  Summary of reverse correlation findings from Exp. 1, 2, and 3. A positive evidence bias in discrimination confidence ratings is evident in the effect of sum evidence (visibility) on discrimination confidence (rows 4-6). In detection, overall visibility had an effect on decisions and confidence in absence in all three experiments, but contributed to confidence in presence only in Exp. 3.

Detection decisions also showed a PEB: when making a detection response participants mostly ignored random fluctuations in stimulus energy that were not aligned with the true, presented signal. However, in Exps. 1 and 2, these fluctuations were later negatively weighed when rating their confidence, as evidences by the absence of a positive evidence bias for confidence in detection ‘yes’ responses. Based on this surprising finding, in Experiment 3 we pre-registered an hypothesis that detection confidence should be equally sensitive to positive and negative evidence. To increase our statistical sensitivity, we doubled the number of detection trials and included a direct manipulation of positive evidence. Results from Experiment 3 provided clear evidence against this hypothesis derived from Experiments 1 and 2, and support an unequal weighting of positive and negative evidence not only for detection decisions, but also for detection confidence judgments.

We considered whether this difference in detection PEB between the results of Experiments 1 and 2 versus 3 merely reflected the higher statistical power of Exp. 3, or alternatively, emerged due to a more interesting difference between these experiments. The similar design of experiments 2 and 3 allowed us to test this hypothesis by directly comparing the observed PEB in Exp. 2 with the one observed in Exp. 3. A PEB for detection confidence was significantly stronger in Exp. 3 than in Exp. 2 (, ), indicating a non-trivial difference between the two experiments. This direct comparison indicates that one or more unique properties of Exp. 3, ones that were not present in Exp. 2, produced a PEB in detection confidence. Two candidates for such features are the inclusion of high-luminance trials, which may have affected participants’ beliefs about the underlying generative model, and the difference in block order: starting with discrimination in Exp. 2, but with detection in Exp. 3. Either option would have important implications for our understanding of this phenomenon (that it is affected by the distribution of sensory samples in the first case, or by task context in the second). We leave it for future research to further investigate the sensitivity of a detection PEB to these different features.

What do our results mean for theoretical accounts of the PEB? Previous accounts have a hypothesized that a PEB may reflect a heuristic that participants adopt due to cognitive constraints in the face of unreasonably vast representational spaces (Maniscalco, Peters, & Lau, 2016) or due to an asymmetric encoding of signal and noise (Miyoshi & Lau, 2020). A heuristic use of evidence in confidence ratings, but not in the decision itself, in turn implies that different processes are involved in the generation of decisions and confidence ratings, and that participants are in some sense being irrational when discarding relevant evidence that could be used in constructing confidence.

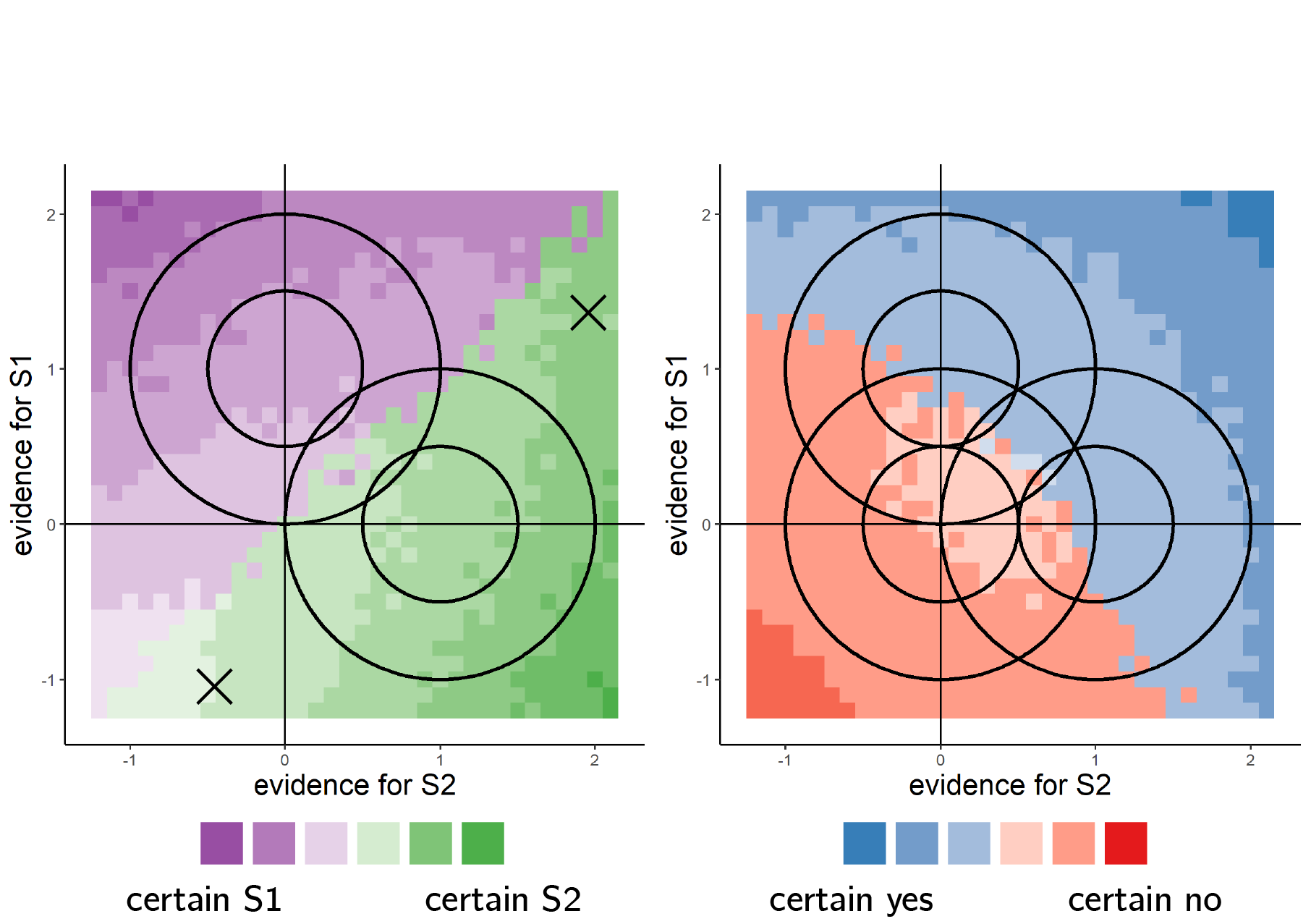
Similarly, a PEB was recently demonstrated in an artificial neural network trained to classify hand-written digits and, in parallel, predict its classification accuracy (Webb, Miyoshi, So, & Lau, 2021). The network was trained on images varying in contrast and visual noise, and later tested on overlays of two digits, varying in contrast only. Under this training regime, the network was more confident for high-contrast images, controlling for classification accuracy. Similar to the heuristic account of Maniscalco, Peters, and Lau (2016), here also a PEB reflected the application of an inductive bias acquired in real life, or in training, to a test setting where doing so is maladaptive.

An alternative possibility is that a single, Bayes-rational model with valid priors governs both choice and confidence ratings, but that the form of that model is yet to be specified. For instance, one possible driver of a PEB in discrimination confidence ratings is the higher informational value of signal than noise. If the signal channel holds more information about signal identity or stimulus presence, giving more weight to information from this channel is rational. This is the case in unequal-variance SDT settings, where signal is sampled from a wider range of values than noise. As an example, if noise is sampled from a Gaussian distribution with mean 0 and variance 1 and signal from a Gaussian distribution with mean 2 and variance 3, sampling the value 6 is much more informative about the presence of a signal than sampling the value -2, because the first is only likely if sampled from the signal distribution (likelihood ratio > 1,000,000), but the second is likely under both distributions (likelihood ratio = 1).[[2]](#footnote-2) Similarly, if the representation of coherent motion is more variable across trials than the representation of random motion, participants would be rational to give more weight to evidence for coherent motion in one channel than evidence for its absence in the other channel.

Higher variability in the representation of signal is often built into the experiment itself. For example, in our Exp. 1, following Zylberberg, Barttfeld, and Sigman (2012), the number of coherently moving dots was itself randomly determined, sampled from a Gaussian distribution once every four frames. This means that there were two sources of variability for the true direction of motion (variability in the direction of randomly moving dots and variability in the number of coherently moving dots), but only one source of variability for the opposite direction (variability in the direction of randomly moving dots). But even when signal is not made more variable by design, the representation of signal is expected to be inherently more variable due to the Weber-Fechner law (Fechner & Adler, 1860) and the coupling between firing rate mean and variability implied by a Poisson form for neuronal firing rates.

To obtain qualitative predictions for such effects, we simulated a stimulus-dependent noise model (full simulation details, including source code are available in appendix E). To model the unequal variance nature of signal and noise, perceptual noise was sampled from a normal distribution with mean 0 and a standard deviation proportional to the exponent of the sensory sample (). This perceptual noise was then added to the sensory sample. We chose to use the exponent of the sensory sample in order to restrict the standard deviation parameter to positive values, but qualitatively similar results are obtained for a linear mapping from sensory samples to sensory noise. A Bayes-rational agent had full knowledge of this generative model in basing their decision and confidence on the likelihood ratio between the two stimulus categories (discrimination) or between signal presence and absence (detection).

Notably, simulation gave rise to a pronounced positive evidence bias in discrimination confidence ratings and in detection decisions (see Fig. 12). The agent was more sensitive to variations in the signal channel when deciding whether a signal was present or not and when rating its confidence, both in detecting the presence of the signal and in discriminating between two stimulus classes. This is in line with our empirical findings. Importantly, in this model decision and confidence ratings are the output of the same Bayes-rational process applied to a situation where perceptual noise scales with signal strength, and do not reflect any suboptimality in evidence weighting.



*Figure* *12.*  Model predictions for a stimulus-dependent noise model. Sensory noise is higher for stronger sensory samples. For each point on the grid, we simulated 200 trials by sampling a sensory sample and extracting a decision and a confidence rating according to the log likelihood ratio of the two hypotheses. Left: two example sensory samples, marked with an X, have the same relative evidence, but absolute evidence is higher for the sample on the upper right. This sample was given a higher-confidence rating, consistent with a positive evidence bias in behaviour but here emerging from a Bayes-rational agent.

However, a stimulus-dependent-noise model makes predictions not only for the effect of sum evidence on discrimination and detection confidence ratings, but also for the effect of sum evidence on decision performance (). Specifically, if stronger stimuli are also noisier, sum evidence should have a positive effect on confidence, but a negative effect on response accuracy. In contrast with this prediction, in Exp. 3, an increase to the luminance of both stimuli had no effect on accuracy (, 95% CI , , , ), but boosted discrimination confidence nonetheless. Moreover, the effect of overall luminance on confidence was positively correlated with its effect on decision confidence (, 95% CI , , , ) and not negatively correlated as would be expected if degraded accuracy and higher confidence were both driven by higher sum evidence. This suggests that Bayes rational inference under stimulus-dependent noise is not all there is to the story.

A recent finding that the PEB can be reversed if the observer’s goal is altered also weigh against a signal detection theoretic, Bayes-rational account of the positive evidence bias. In a recent study, Sepulveda et al. (2020) presented participants with pairs of dot arrays and asked them to choose the array with more white dots. Participants were more confident when both arrays included more dots, replicating a positive evidence bias. Critically, the experiment also included a second condition, in which participants were asked to choose the array with *fewer* white dots. In this condition, confidence was higher when both arrays included fewer dots: a “negative,” rather than positive, evidence bias. This effect was also related to participants’ information sampling behaviour in the two conditions: they spent more time fixating their gaze at the array containing more dots under typical instructions, but the opposite was the case when they were instructed to select the array with fewer dots. In a Bayes-rational model, evidence weighting should be identical for these two equivalent ways of framing the task instructions. This finding is also inconsistent with heuristic models that are based on variance differences between the encoding of signal and noise, as those are not expected to change as task instructions change.

Instead, our findings, as well as those of Sepulveda et al. (2020), are generally in line with a heuristic account of positive evidence bias that posits limits on cognitive resources when coping with high-dimensional representations (Maniscalco, Peters, & Lau, 2016). Basing confidence on positive evidence in such a world frees agents from the need to consider an infinite number of alternative hypotheses. Similarly, the findings of Sepulveda et al. (2020) can be accounted for by a model in which participants flexibly allocate attention to the choice-consistent dimension of evidence (more vs. fewer dots), while ignoring other dimensions. What constitutes as positive evidence is then rationally dependent on an agent’s specific goals and attentional set at the time of performing a task.

# 6 References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, *2*(2), 284–299.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Fechner, G. T., & Adler, H. E. (1860). Elements of psychophysics [elemente der psychophysik]. *Leipzig, Germany: Breitkopf and Ha Rtel*.

Kellij, S., Fahrenfort, J., Lau, H., Peters, M. A., & Odegaard, B. (2021). An investigation of how relative precision of target encoding influences metacognitive performance. *Attention, Perception, & Psychophysics*, *83*(1), 512–524.

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics*, *77*(4), 1295–1306.

Lange, K., Kuhn, S., & Filevich, E. (2015). Just another tool for online studies (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6), e0130834.

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d’, response-specific meta-d’ and the unequal variance SDT model. In *The cognitive neuroscience of metacognition* (pp. 25–66). Springer.

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*(3), 923–937.

Mazor, M., Friston, K. J., & Fleming, S. M. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. *ELife*, *9*, e53900.

Mazor, M., Mazor, N., & Mukamel, R. (2019). A novel tool for time-locking study plans to results. *European Journal of Neuroscience*, *49*(9), 1149–1156.

Mazor, M., Moran, R., & Fleming, S. (2021). Stage 2 registered report: Metacognitive asymmetries in visual perception.

McCarthy, L. (2015). p5. js. *URL: Https://P5js. Org*, *3*.

Meuwese, J. D., Loon, A. M. van, Lamme, V. A., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*, *76*(4), 1057–1068.

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review*, *127*(5), 655.

Peters, M. A., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., … others. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 1–8.

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics*, *80*(1), 134–154.

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), 1–11.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Samaha, J., & Denison, R. (2020). The positive evidence bias in perceptual confidence is not post-decisional. *bioRxiv*.

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*, e60705.

Webb, T., Miyoshi, K., So, T. Y., & Lau, H. (2021). A task-optimized neural network model of decision confidence. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, USA.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*, 79.

Appendix

# A Additional analyses: Exp. 1

## A.1 Response time, confidence, and metacognitive sensitivity differences

In detection, participants were generally slower to deliver ‘no’ responses compared to ‘yes’ responses (median difference: 85.37 ms, , for a t-test on the log-transformed response times). No significant difference in response times was observed for the discrimination task (median difference: 6.16 ms, , ).

Confidence in detection was generally higher than in discrimination (, 95% CI , , , ). Within detection, confidence in ‘yes’ responses was generally higher than confidence in ‘no’ responses (, 95% CI , , , ). No difference in average confidence levels was found between the two discrimination responses (, 95% CI , , , ).

Following Meuwese, Loon, Lamme, and Fahrenfort (2014), we extracted response-conditional type-2 ROC (rc-ROC) curves for the two tasks. Unlike traditional type-I ROC curves that provide a summary of subjects’ ability to distinguish between two external world states, type 2 ROC curves represent their ability to track the accuracy of one’s own responses. The area under the response-conditional ROC curve (auROC2) is a measure of metacognitive sensitivity, with higher values corresponding to more accurate metacognitive monitoring.

Mean response-conditional ROC curves for the two responses in the discrimination task closely matched (, 95% CI , , , ), indicating that on average, participants had similar metacognitive insight into the accuracy of the two discrimination responses. In contrast, auROC2 estimates for ‘yes’ responses were significantly higher than for ‘no’ responses, indicating a metacognitive asymmetry between the two detection responses (group difference in auROC2: , 95% CI , , , ).

## A.2 zROC curves

A difference in response-conditional auROC estimates can emerge from higher-order differences in metacognitive monitoring for the two responses and/or from lower-level differences in the perceptual representations of signal and noise [such as in first-order signal detection models where the signal variance is higher; Maniscalco and Lau (2014)]. Importantly, a difference can also emerge in first-order signal-detection models that assume equal variance, in the presence of a response bias or insufficient variance in confidence ratings (Mazor, Moran, & Fleming, 2021). To test if the metacognitive asymmetry between ‘yes’ and ‘no’ responses could be accounted for by an equal-variance SDT model, we simulated data that was identical to our empirical data except for confidence ratings in correct responses, which were chosen to perfectly agree with the assumptions of an equal-variance SDT model given participants’ decision criterion, sensitivity, and their confidence in incorrect responses. We then compared subject-wise differences between the response-conditional auROCs with the differences in this simulated dataset (Mazor, Moran, & Fleming, 2021). The difference in differences was significant, indicating that the observed metacognitive asymmetry could not be accounted for by a first-order equal-variance SDT model (, 95% CI , , , ).

An asymmetry in metacognitive sensitivity for ‘yes’ and ‘no’ responses is also predicted by unequal-variance Signal Detection Theory (*uvSDT*). Specifically, if the signal distribution is wider than the noise distribution, the overlap between the distributions will be more pronounced for misses and correct rejections than for hits and false alarms, making metacognitive judgments for ‘no’ responses objectively more difficult. Unequal-variance SDT predicts that plotting the type-1 ROC curve in z-space (taking the inverse cumulative distribution of the confidence rating histogram) will result in a straight line with a slope equal to . Because the variance of the signal distribution is higher than that of the noise distribution, zROC slopes are typically shallow, with slopes below 1.

We used linear regression to estimate the slope of the zROC curve. To control for underestimation of the slope due to regression to the mean (Wickens, 2002, p. 56), we fitted two regression models for the task data of each participant: one predicting based on (slope ) and one predicting based on (slope ). We then used as a bias-free measure of the zROC slope. In equal-variance SDT, this value is predicted to be 0, corresponding to a slope of 1.

Indeed, slopes were generally shallower than 1 for detection zROC curves (as predicted by an unequal-variance SDT model; , 95% CI , , , ), and not significantly different from 1 for discrimination zROC curves (as predicted by equal-variance SDT; , 95% CI , , , ).

These results support a difference in the variance-structure of the representation of signal and noise, such that the representation of signal is more variable across trials. However, it is still possible that some of the metacognitive asymmetry in detection (the difference in auROC between ‘yes’ and ‘no’ responses) reflects additional higher-order processes that cannot be captured by a first-order signal-detection model. If this was the case, zROC curves for detection should not only be shallower, but also less linear than for discrimination, reflecting poorer fit of the signal-detection model to detection. To test if this was the case, we compared the subject-wise values for the detection and discrimination zROC regression lines. values reflect the goodness of fit of a linear model to the data. These values were similar for the two tasks (, 95% CI , , , ), suggesting that a first-order SDT model accounted equally well for the data from both tasks.

## A.3 Confidence-RT correlations

Following our pre-registered analysis plan, we extracted a Spearman correlation coefficient between confidence and response times separately for the two tasks and four responses. We find a negative correlation in all four cases (discrimination responses: -0.40 and -0.39, detection ‘yes’: -0.41, detection ‘no’: -0.33). As hypothesized, this negative correlation was significantly attenuated in detection ‘no’ responses compared to detection ‘yes’ responses (tested with a one-tailed t-test: , ). The difference in correlation strength between detection ‘no’ responses and discrimination responses was only marginally significant (, ).

## A.4 Global metacognitive estimates

At the end of each 100-trial block, participants estimated their block-wise accuracy. Mean estimated accuracy was 0.71 for discrimination and 0.69. These figures are close to true correct response rates: 0.74 in discrimination and 0.72 in detection.

A difference of 0.02 between mean accuracy estimates for discrimination and detection was not significant at the group level (, ).

# B Additional analyses: Exp. 2

## B.1 Response time, confidence, and metacognitive sensitivity differences

Participants were slower to deliver ‘no’ responses compared to ‘yes’ responses (median difference: 77.12 ms, , for a t-test on the log-transformed response times). No significant difference in response times was observed for the discrimination task (median difference: 10.90 ms, , ).

Confidence in detection was generally lower than in discrimination, consistent with lower accuracy in this task (, 95% CI , , , ). Within detection, confidence in ‘yes’ responses was generally higher than confidence in ‘no’ responses (, 95% CI , , , ). No difference in average confidence levels was observed between the two discrimination responses (, 95% CI , , , ).

In contrast to the results of Exp. 1, auROC2 values for ‘yes’ and ‘no’ responses were not significantly different (group difference in area under the response-conditional curve, auROC2: , 95% CI , , , ), first and second rows). auROC2s were also not significantly different when controlling for type-1 response and confidence biases (, 95% CI , , , ).

## B.2 zROC curves

Unlike in Experiment 1, detection zROC slopes were not significantly different from 1 (, 95% CI , , , ), whereas discrimination zROC slopes were significantly shallower than 1 (, 95% CI , , , ). This unexpected result indicates equal variance for the signal and noise distributions, but higher variance for targets presented on the right than on the left. Furthermore, a first-order SDT model fitted the data significantly better for the detection task than for the discrimination (difference in for the two tasks: , 95% CI , , , ).

# C Additional analyses: Exp. 3

## C.1 Response time, confidence, and metacognitive sensitivity differences

Participants were slower to deliver ‘no’ responses compared to ‘yes’ responses (median difference: 68.90 ms, , for a t-test on the log-transformed response times). No significant difference in response times was observed for the discrimination task (median difference: 11.28 ms, , ).

Confidence in detection was generally lower than in discrimination, consistent with lower accuracy in this task (, 95% CI , , , ). Within detection, confidence in ‘yes’ responses was generally higher than confidence in ‘no’ responses (, 95% CI , , , ). No difference in average confidence levels was observed between the two discrimination responses (, 95% CI , , , ).

## C.2 Reverse correlation analysis of standard trials only

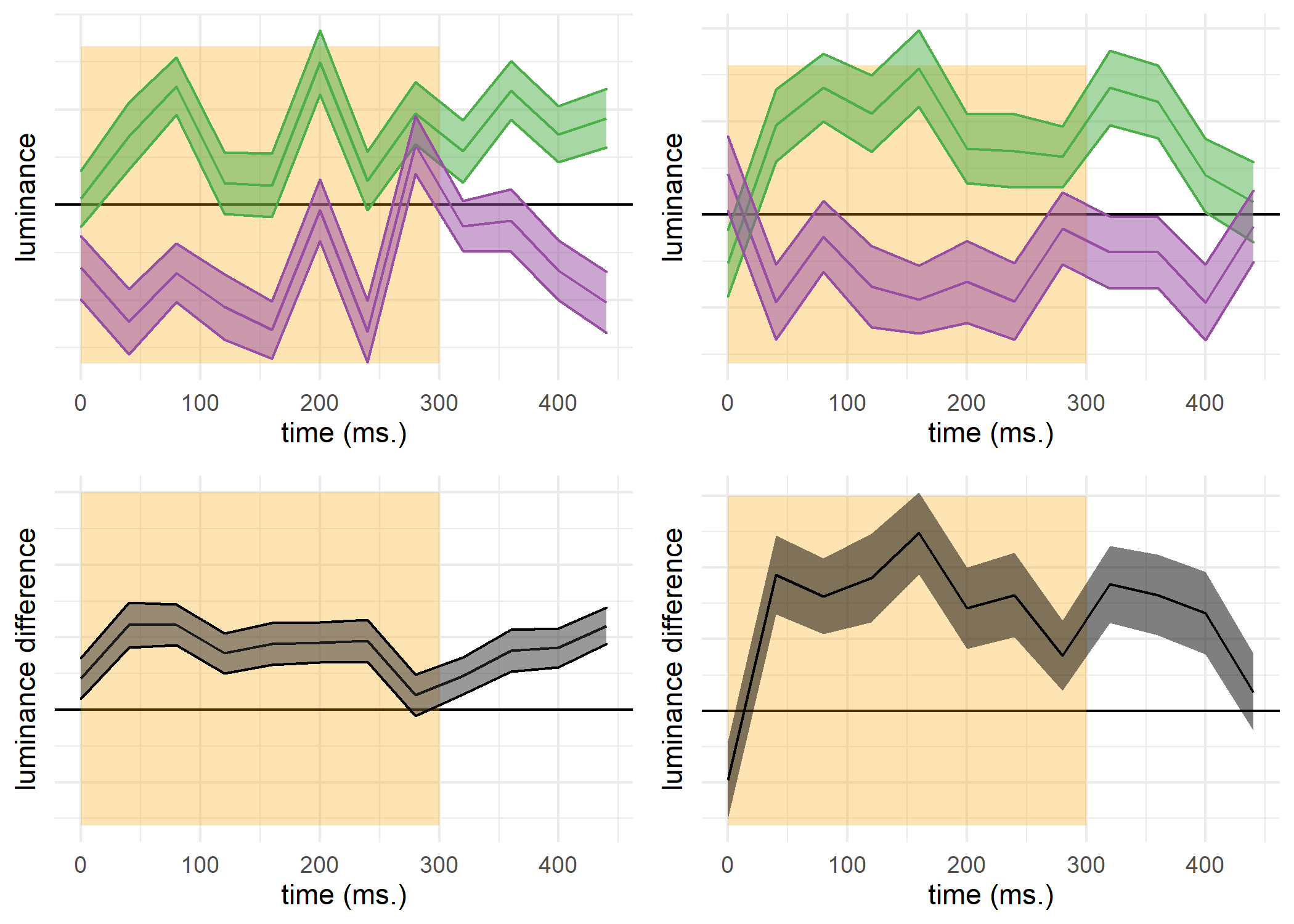
In the following, we repeat the reverse correlation analysis reported for Exp. 3, but here restricted to the subset of “standard” trials where luminance was not increased by 2/255.

### Discrimination decisions.

Discrimination decisions were sensitive to fluctuations in luminance during the first 300 milliseconds of the trial (, ). We found no evidence for a positive evidence bias in discrimination decisions, even when grouping evidence based on the location of the true signal rather than subjects’ decisions (, ).

### Discrimination confidence.

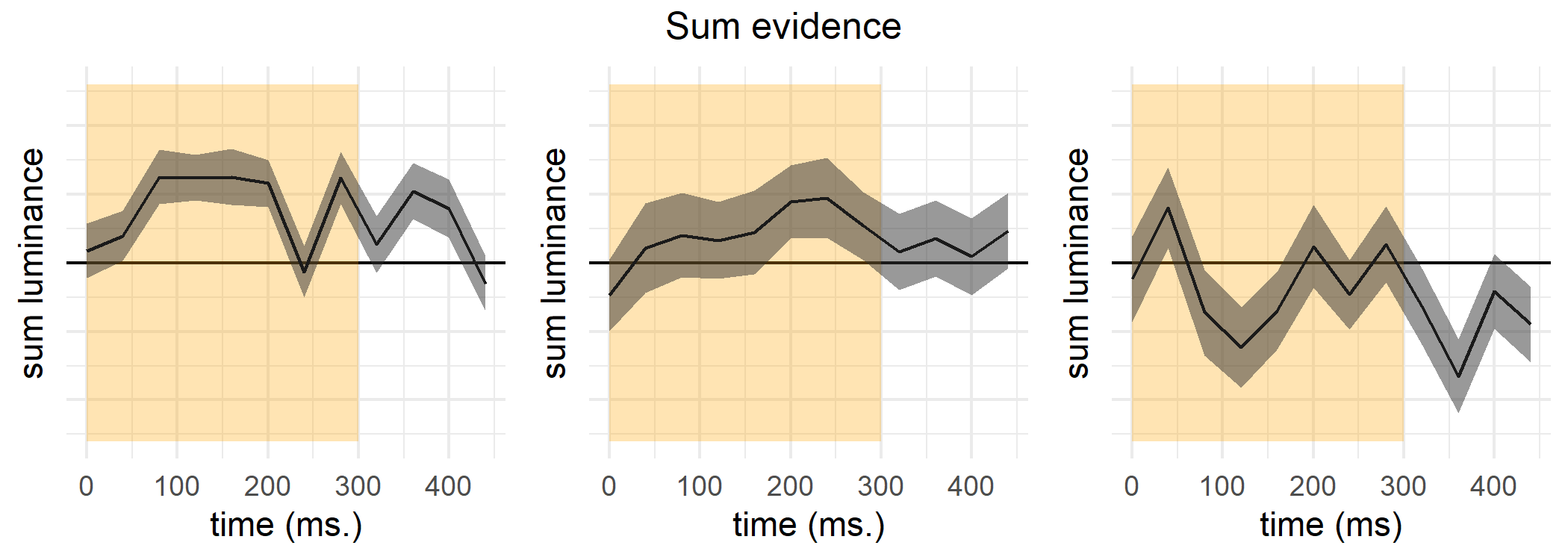
Luminance within the first 300 milliseconds had a significant effect on confidence ratings (, ; see Fig. 13, right panels). A positive evidence bias in discrimination confidence was not significant in this sample (, ).



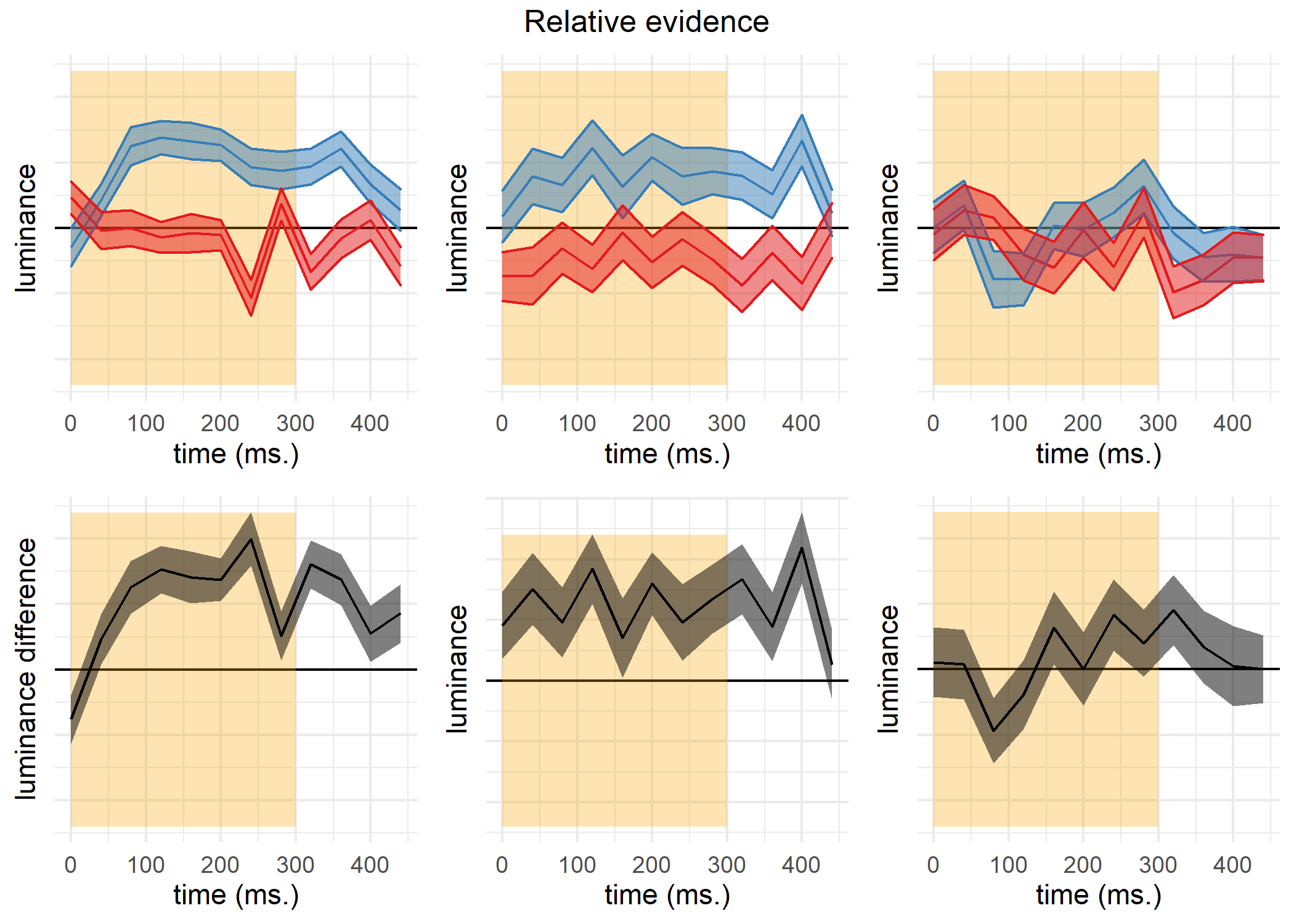
*Figure* *C1.*  Decision and confidence discrimination kernels, Experiment 3, standard trials only.

### C.2.1 Detection.

Similar to Exp. 2, sum luminance had a significant effect on participants’ detection responses during the first 300 milliseconds (, ; see Fig. 14, left panel). Sum luminance also had a significant positive effect on decision confidence when reporting target presence (‘yes’ responses; , ). In contrast, a negative effect on confidence when reporting target absence was not significant (, ; see Fig. 14, middle and right panels).



*Figure* *C2.*  Decision and confidence sum luminance detection kernels, Experiment 3, standard trials.



*Figure* *C3.*  Decision and confidence kernels for detection signal trials, Experiment 3, standard trials only.

### C.2.2 Detection signal trials.

When deciding whether one of the two flickering patches was brighter than the background, participants were sensitive to positive noise in the target patch more than to negative noise in the non-target patch (, ), consistent with a positive evidence bias in detection decisions and replicating findings from Exps. 1 and 2. Random fluctuations in luminance in the first 300 milliseconds of the trial also contributed to confidence in detection ‘yes’ responses (hit trials; , ). Importantly, however, and in contrast to the results of Exp. 1 and 2, confidence in ‘yes’ responses was more sensitive to positive evidence than to conflicting evidence (but only marginally so: , ). Confidence in ‘miss’ trials was independent of the contrast in luminance between the right and left stimuli (, ) but, as described above, confidence in ‘no’ responses was sensitive to the overall luminance of the display. This negative effect of luminance on confidence in ‘no’ responses was not significant either for the non-target stimulus (, ), or for the target stimulus (, ).

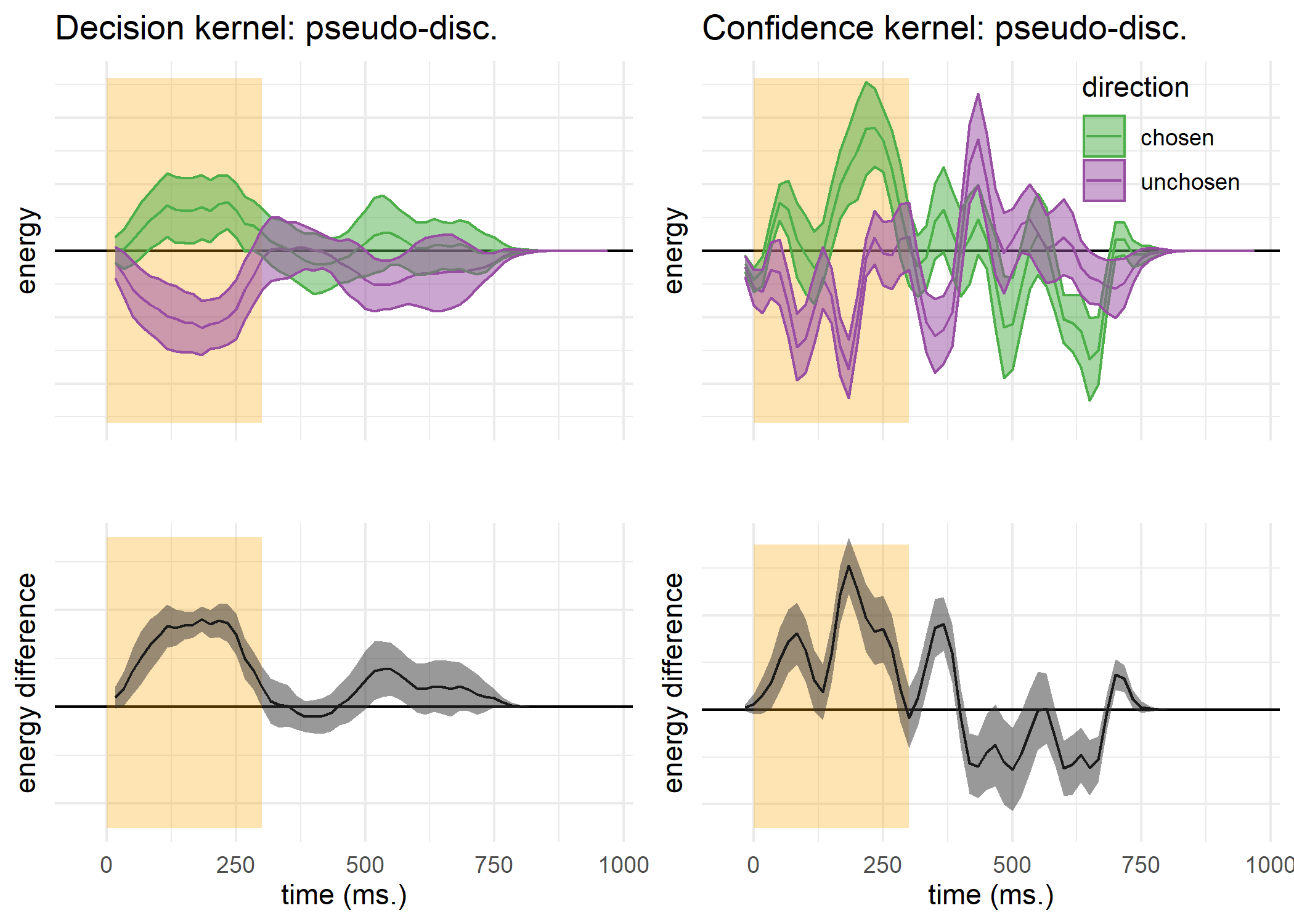
# D Pseudo-discrimination analysis

In our pre-registration document (https://osf.io/d3vkm/), we specified our plan for *pseudo-discrimination analysis*, where we analyze detection ‘signal’ trials as if they were discrimination trials:

In this analysis, we will assume that in the majority of ‘different’ trials, when participants responded ‘yes’ they correctly identified the brighter set. For example, a detection trial in which the brighter set was presented on the right and in which the participant responded ‘yes’ will be treated as a discrimination trial in which the participant responded ‘right.’ Conversely, a trial in which the brighter set was presented on the right and in which the participant responded ‘no’ will be treated as a discrimination trial in which the participant responded ‘left.’ These hypothetical responses will then be submitted to the same reverse correlation analysis described in the previous section confidence kernels.

We subsequently realized that a much simpler approach is to contrast ‘yes’ and ‘no’ responses for the true and opposite direction of motion (or flickering stimuli) in signal trials. This alternative approach does not entail treating ‘no’ responses as the successful detection of a wrong signal. The results of this analysis mostly agree with the pre-registered pseudo-discrimination analysis. For completeness, we include the pre-registered pseudo-discrimination analysis for both experiments here.

## D.1 Exp. 1

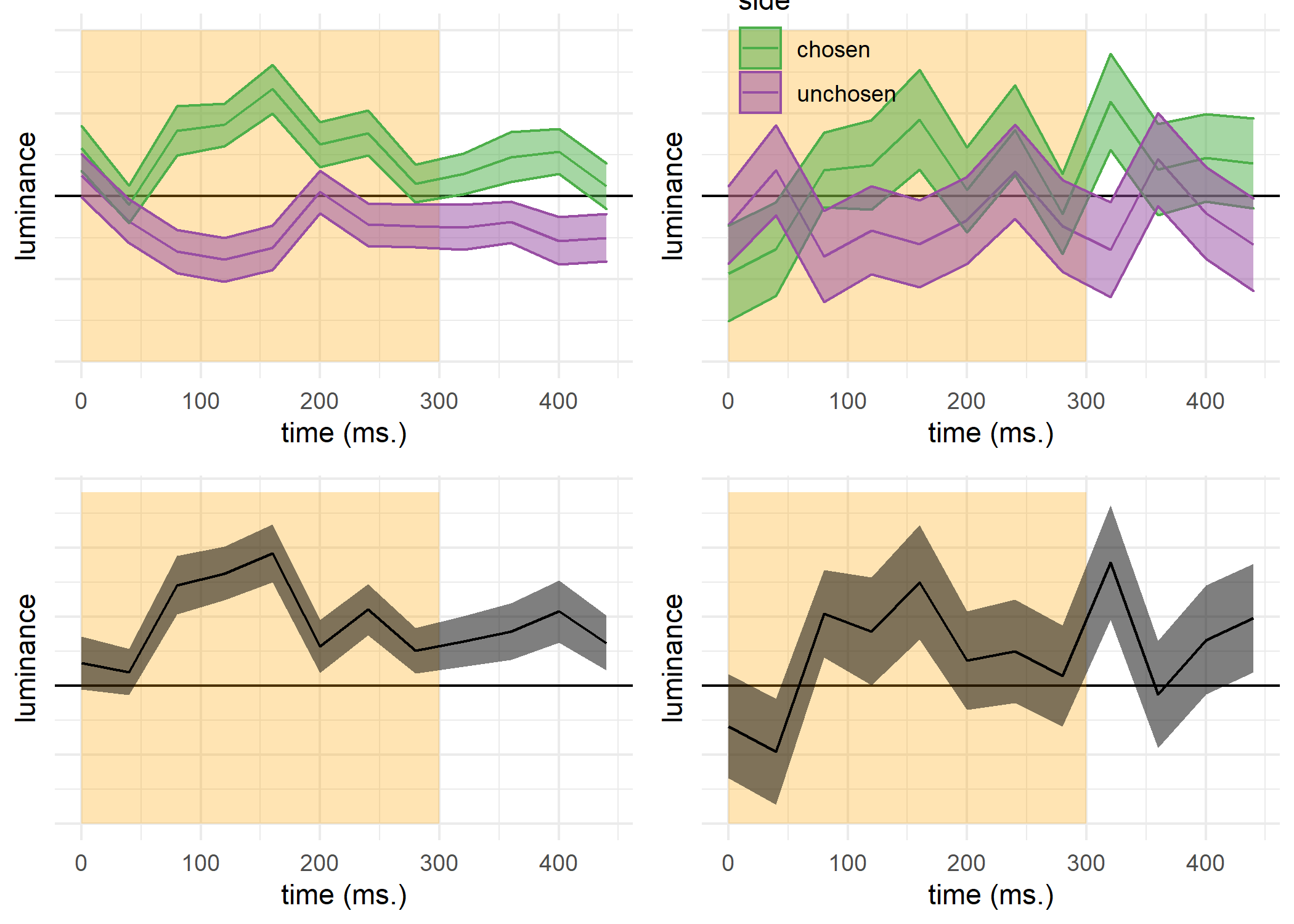


*Figure* *D1.*  Decision and confidence pseudo-discrimination kernels, Experiment 1. Upper left: motion energy in the “chosen” (green) and “unchosen” (purple) direction as a function of time. Bottom left: a subtraction between energy in the “chosen” and “unchosen” directions. Upper right: confidence effects for motion energy in the “chosen” (green) and “unchosen” (purple) directions. Lower right: a subtraction between confidence effects in the “chosen” and “unchosen” directions. Shaded areas represent the the mean +- one standard error. The first 300 milliseconds of the trial are marked in yellow.

Pseudo-discrimination decision kernels were highly similar to discrimination decision kernels. Here also, motion energy during the first 300 milliseconds of the stimulus had a significant effect on decision (, ) and on decision confidence (, ). However, unlike in discrimination, where motion energy in the chosen direction influenced decision confidence more than motion energy in the unchosen direction, no such positive evidence bias was observed for detection responses (, ).

While motion energy during the first 300 milliseconds of the trial significantly affected confidence in ‘yes’ responses (, ), it had no significant effect on confidence in ‘no’ responses (, ). However, given that the pseudo-discrimination analysis was performed on signal trials only, confidence kernels for ‘no’ responses were based on fewer trials than confidence kernels for ‘yes’ responses, such that the absence of a significant effect in ‘no’ responses may reflect insufficient statistical power to detect one.

## D.2 Exp. 2



*Figure D2.* Decision and confidence pseudo-discrimination kernels, Experiment 2. Upper left: luminance in the “chosen” (green) and “unchosen” (purple) stimulus as a function of time and spatial position. Bottom left: decision kernel averaged across the four spatial positions. Upper right: confidence effects for motion energy in the “chosen” (green) and “unchosen” (purple) stimuli. Bottom right: confidence effects averaged across the four spatial positions. Shaded areas represent the mean one standard error. The first 300 milliseconds of the trial are marked in yellow.

Similar to decision kernels in Exp. 2, random fluctuations in luminance during the first 300 milliseconds of the stimulus had a significant effect on decision (, ). However, in Exp. 2 this anslysis revealed no effect of luminance on decision confidence (, ), and no positive evidence bias in confidence judgments (, ).

# E Stimulus-dependent noise model

## E.1 Discrimination

### Generative model.

Stimuli were represented as pairs of numbers, corresponding to the two sensory channels (e.g., right and left motion). One sensory channel transmitted pure noise, and one channel had additional signal in it. The signal channel was chosen randomly for each trial with equal probability.

On top of the presented noise, we added perceptual noise to the stimulus. Importantly, this additional noise affected the agent’s decisions and confidence ratings, but did not affect the stimulus itself such that trial-wise estimates of stimulus energy were unaffected for the reverse correlation analysis. The noise was channel specific, and its magnitude dependent on the magnitude of the underlying signal:

### Inference.

The log likelihood ratio is computed to decide whether it is more likely that the signal was in channel 1 or 2.

### E.2 Detection.

### Generative model.

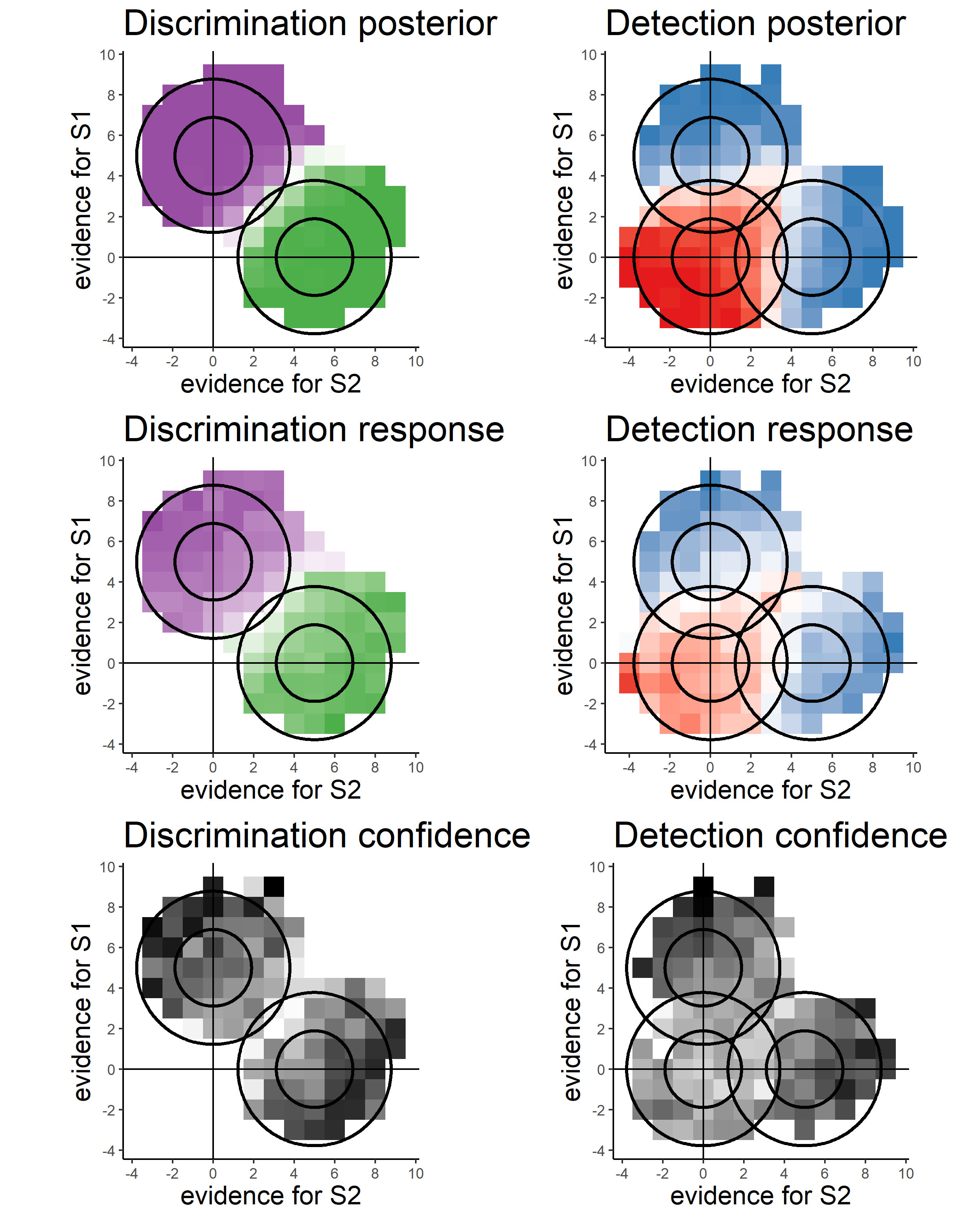
Similar to discrimination, except that on half of the trials both channels transmitted noise only.

### Inference.

The log likelihood ratio is computed to decide whether it is more likely that the signal was present or absent.

# D Effects of evidence on decision and confidence: Exp. 2 and 3

To compare participants’ empirical behaviour to our model simulations, we plotted optimal behaviour, participants’ responses, and confidence in correct responses, as a function of perceptual evidence in a two-dimensional representational space. First, for each trial we extracted mean luminance (minus background luminance) in the first 300 milliseconds for the right and left stimuli. These numbers were rounded to the closest integer. For each tuple of such integers, we extracted the posterior probability for stimulus category (Fig. 18, top row), participants’ empirical discrimination and detection decisions (middle row), and participants’ subjective confidence in correct responses (bottom row).



*Figure* *D1.*  Top row: posterior probability of stimulus category given perceptual evidence for discrimination (left) and detection (right). Middle row: decision probability as a function of perceptual evidence. Bottom row: mean confidence in correct responses as a function of perceptual evidence.

1. We reused the original Matlab code that was used for Exp. 1 in Zylberberg et. al. (2012), kindly shared by Ariel Zylberberg. [↑](#footnote-ref-1)
2. In other words, the expected log likelihood ratio when sampling from the signal distribution is higher (in absolute terms) than when sampling from the noise distribution, or . [↑](#footnote-ref-2)