Paradoxical evidence weighting in confidence judgments for detection and discrimination

Matan Mazor1, Lucie Charles4, Roni Maimon-Mor4, & 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

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

When making discrimination decisions between two stimulus categories, subjective confidecne judgments are more positively affected by evidence in support of a decision than negatively affected by evidence against it. Recent theoretical proposals suggest that this “positive evidence bias” may be 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, this evidence weighting asymmetry 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 positive evidence bias in discrimination confidence. We then show that detection decisions and confidence ratings paradoxically suffer from an opposite “negative evidence bias” to negatively weigh evidence even when it is optimal to assign it a positive weight. We show that the two effects are uncorrelated, and 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

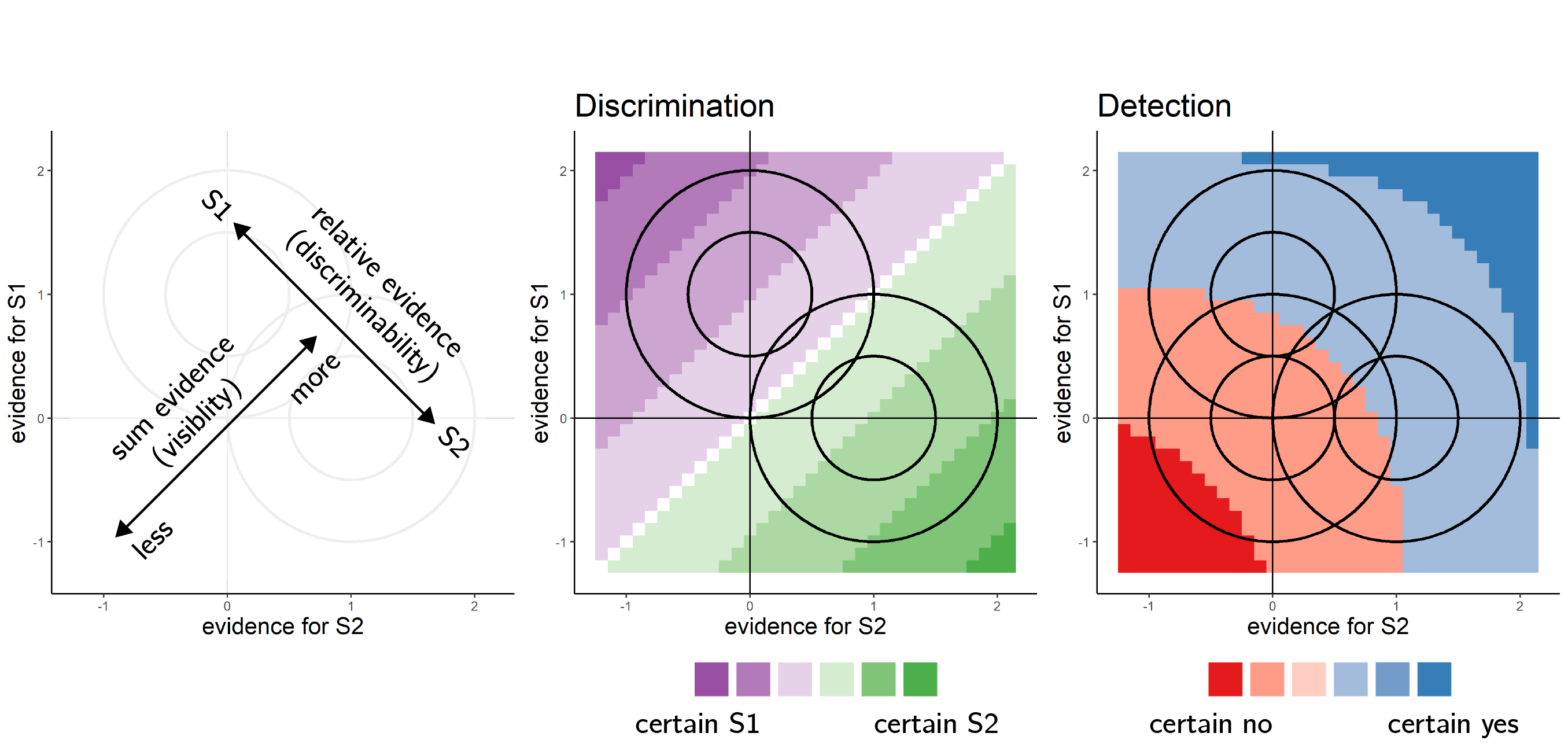
*Word count:* X

Paradoxical evidence weighting in confidence judgments for detection and discrimination

# Introduction

When considering two alternative hypotheses, the probability of the chosen hypothesis to be correct is a function of the availability of evidence supporting not only the chosen hypothesis, but also the unchosen one. For example, when deciding that there are more ants in the kitchen than in the living room, confidence should not only positively weigh the number of ants found in the kitchen (*positive evidence*), but also negatively weigh the number of ants found in the living room (*negative evidence*). Specifically, a decision should be based on the difference in the number of ants between the kitchen and the living room, but not on the total number of ants found in both rooms together (we refer to these quantities as *relative evidence* and *sum evidence*, respectively).

While sum evidence is irrelevant to discrimination decisions between two symmetrical hypotheses (e.g., kitchen or living room), it is highly informative with respect to detection decisions about the presence or absence of a signal. For example, when deciding that an ant colony is nesting in the house, the total number of ants found in both rooms is highly relevant, even more so than the difference between the number of ants found in the kitchen or living room (see Fig. 1 for a signal detection formulation).



*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: decision 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.

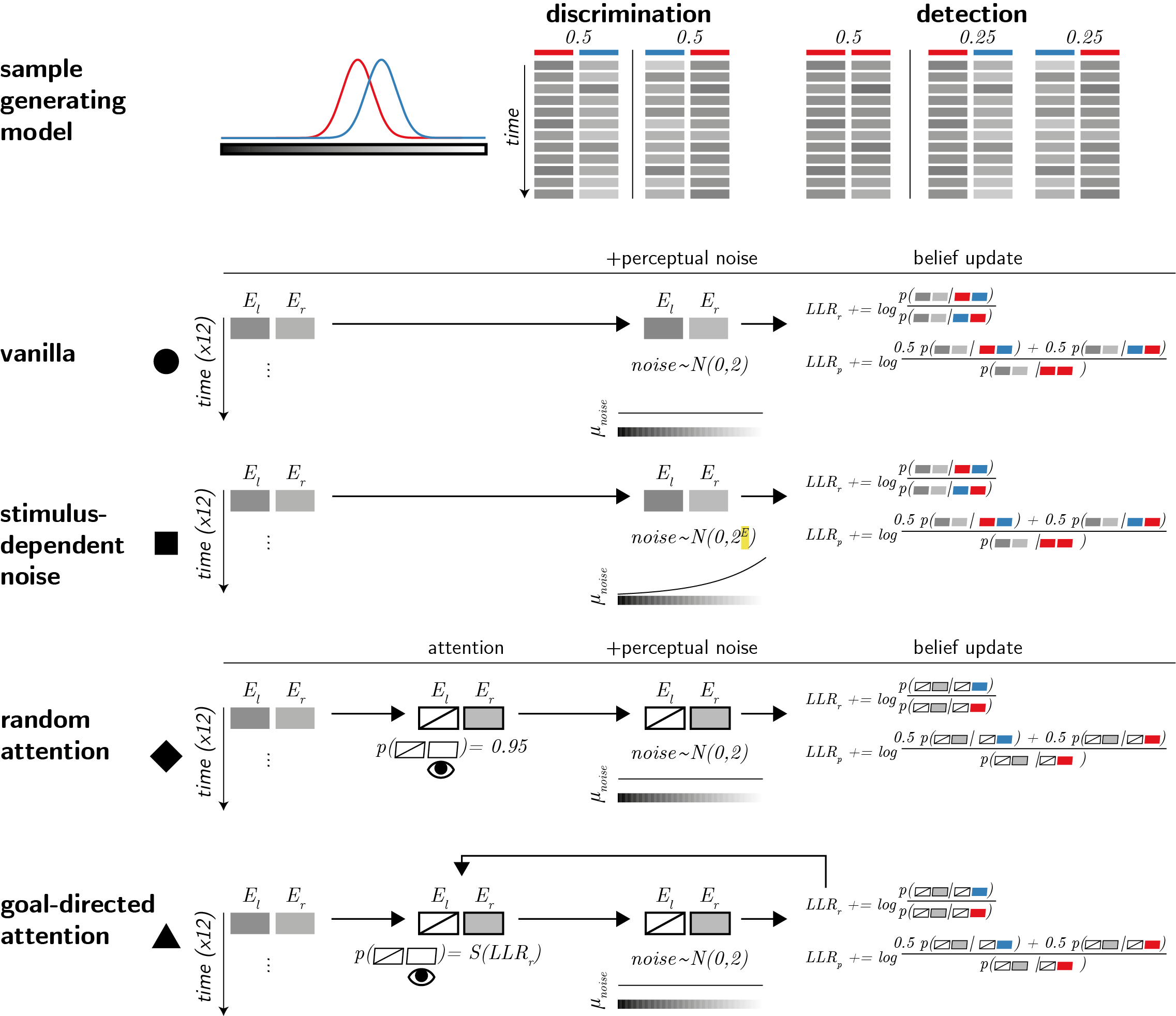
A surprising finding is that despite the irrelevance of sum evidence to the accuracy of discrimination decisions, people are systematically more confident in perceptual decisions when sum evidence is high. For example, Zylberberg, Barttfeld, and Sigman (2012) had subjects judge which of two flickering stimuli was brighter on average. Subjects were more confident in their decisions when both stimuli were bright, indicating an effect of sum evidence (here, overall luminance) on decision confidence. A positive effect of sum evidence on decision confidence is mathematically equivalent to a disproportional weighting of positive evidence over negative evidence, also known as a positive evidence bias (Koizumi, Maniscalco, & Lau, 2015; Peters et al., 2017; Rollwage et al., 2020; Samaha & Denison, 2020; Sepulveda et al., 2020; Zylberberg et al., 2012). The two are equivalent because positively weighing the sum of positive and negative evidence effectively weakens the negative contribution of negative evidence to decision confidence, while strengthening the contribution of positive evidence. Notably, an effect of sum evidence on discrimination confidence may indicate a profound link between discrimination confidence and detection decisions (Rausch, Hellmann, & Zehetleitner, 2018).

Different models identify the origin of this evidence weighting asymmetry at different levels of the cognitive hierarchy, going from a metacognitive bias to ignore conflicting information [metacognitive level; Peters et al. (2017); Maniscalco, Peters, and Lau (2016)], through asymmetries in the active sampling of evidence [attention allocation level; Sepulveda et al. (2020)], down to perceptual asymmetries between the representations of signal and noise [perception level; Miyoshi and Lau (2020); Webb, Miyoshi, So, and Lau (2021)]. These models vary in whether they postulate separate evidence accumulation processes for decision and confidence judgments, and in whether they model confidence formation as following a suboptimal heuristic, or alternatively as being optimal with respect to available information (information which may be limited or corrupted by noise).

Here we focus on a subset of models which assume that subjects are rational decision makers equipped with veridical beliefs about the world, but that they only have limited access to noisy evidence. Our models further assume that subjects’ confidence ratings are Bayesian estimates of the probability of being correct, given the exact same evidence that was used to make the decision. The models do not postulate any metacognitive biases, heuristics, or suboptimalities. We show that two of these models reproduce a positive evidence bias (that is, a positive effect of sum evidence) in discrimination confidence. The same models also make predictions for evidence weighting in detection judgments and confidence ratings. In four experiments, reverse correlation reveals evidence weighting patterns that only partly agree with the predictions of our models. Most notably, our four models fail to account for a negative evidence bias in detection decisions and confidence. In what follows we first describe the four models and the predictions they make, before turning to empirical findings from our four experiments.

# Computational models

We model a setting in which agents are presented with a sequence of samples from two sensory channels: and . The agents’ task is to decide which of the two channels was the signal channel (discrimination), or whether any of the channels had signal in it at all (detection). When a signal is present in a channel, evidence E is sampled from a normal distribution , and when a signal is absent evidence is sampled from (see Fig. 2, upper panel). In all four models agents only have access to a noisy version of these samples , corrupted by sensory noise. After each time step, they update their belief about the relative likelihood of the observed samples under the two possible world states (signal in channel 1 versus 2, or signal presence versus absence), and given full knowledge of the true sample-generating process, including the properties of sensory noise. Each trial comprises 12 time steps. At the end of a trial, agents report the world state that maximizes the likelihood of the observed evidence, and rate their confidence as the objective probability that their decision was correct given likelihood estimates. The four models vary in the properties of sensory noise, and in the selection of some channels for inspection by selection mechanisms.



*Figure* *2.*  Computational models. Upper panel: True world model. Stimuli span 12 timepoints, each comprising values from two sensory channels (here presented as luminance values). In discrimination blocks, values in one channel are sampled from the noise distribution (red), and values in the other channel are sampled from the signal distribution (blue). In detection blocks, on half of the trials all values are sampled from the noise distribution (red). Vanilla model: on each timepoint, participants perceive both channels, corrupted by sensory noise that is sampled from a normal distribution. They then update their beliefs accordingly. Stimulus-dependent noise model: the standard deviation of the sensory noise distribution is exponential with respect to signal intensity. Random attention model: agents only attend one channel at a time. The attended channel is chosen at random per timepoint, with a strong bias to be consistent within a trial. Goal-directed attention model: channels that are likely to include signal based on previous samples are more likely to be attended.

## Vanilla

In the basic, vanilla model, sensory noise is sampled from a normal distribution . This model corresponds to a standard equal-variance signal detection setting, as illustrated in Fig. 1.

## Stimulus-dependent noise

The stimulus-dependent noise model is similar to the vanilla model, with the exception that perceptual noise increases with signal intensity (formally, perceptual noise is sampled from a normal distribution ). This assumption is biologically plausible: 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. Identifying the origin of the positive evidence bias at the perceptual level, this model shares family resemblance with the perceptual model of Miyoshi and Lau (2020). However, two important differences are worth pointing out. First, here perceptual noise is conditioned not on stimulus class, but on the perceptual sample. This seems more plausible, as the perceptual system has no access to stimulus class beyond the information that is available in perceptual samples. And second, here we assume that confidence ratings are made based on all available evidence in a Bayes-rational way, rather than following a heuristic.

## Random attention

In this model, sensory noise is again sampled from , irrespective of signal intensity. Unlike the vanilla model, however, here agents have access to one channel per timepoint only (they “attend” to one channel at a time). On each trial, agents are biased to attend to one preferred channel, chosen at random. Then, on each timepoint, they attend to one channel (the preferred channel with probability 0.95), and update their beliefs accordingly. We include this model because it is inherently asymmetrical: on each trial, evidence from the preferred channel contributes more to decision and confidence.

## Goal-directed attention

This model is similar to the sensory noise model, except that here attention is biased to channels that are more likely to include signal. Specifically, agents track the log likelihood ratio between signal in the left or in the right channels, and the probability of attending the right channel is set to where is a sigmoid function with a steep slope of 5. A highly similar drift diffusion model was shown to produce a positive evidence bias in confidence ratings (Sepulveda et al., 2020).

We simulated 10,000 discrimination and 10,000 detection trials per model (100 trials x 100 simulated agents per model). On each discrimination trial, the signal channel could be right or left with equal probability. On half of the detection trials both channels were noise channels. We then sampled, for each trial, 12 values from each channel. These 24 values were then passed on to the simulated agent, who returned a decision and a confidence rating. We then subjected the agents’ decisions and confidence ratings to a reverse correlation analysis. We now turn to describe this analysis, which will also be used to analyze the behaviour of (actual!) participants in Exp. 1-4.

## Reverse correlation analysis

Following Zylberberg et al. (2012), we took a reverse correlation approach and asked which sources of evidence (positive, negative, relative, and sum evidence) contribute to agents’ decisions and confidence ratings. This analysis focuses on random fluctuations in signal intensity, and asks how they affect behaviour (here, decisions and confidence in these decisions). For simplicity, here we collapse across timepoints instead of correlating noise with decision or confidence per timepoint, as we do for human data.

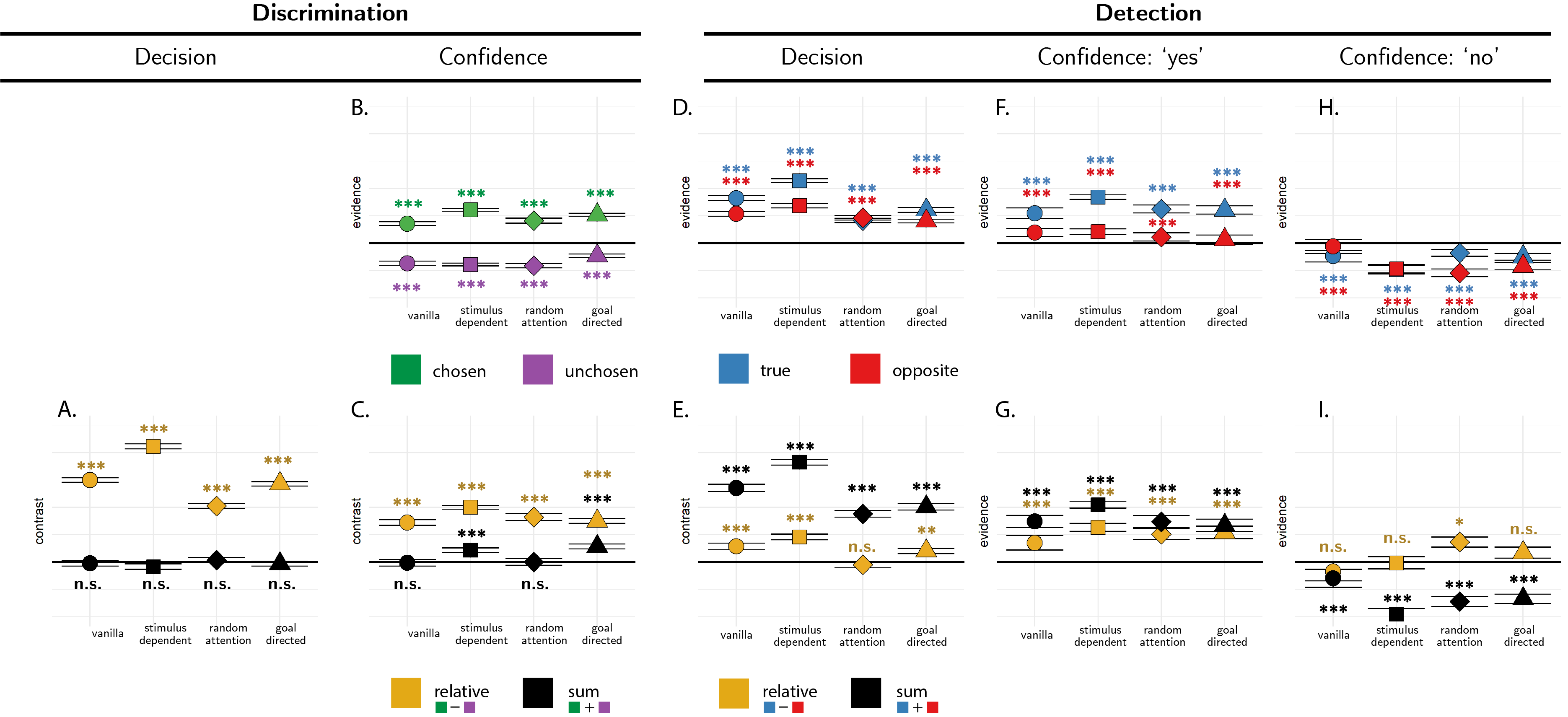
## Methodological note: positive evidence bias in perceptual decisions

The positive evidence bias in decision confidence is often seen as particularly striking, given that positive and negative evidence are equally weighted in forming a decision (Peters et al., 2017; Zylberberg et al., 2012). For example, using reverse correlation, Zylberberg et al. (2012) showed that momentary fluctuations in the availability of perceptual evidence for and against a decision were equally predictive of the decision itself. Similarly, Peters et al. (2017) showed that in classifying rapidly presented images as ‘face’ or ‘house’, decisions are not solely guided by positive evidence (e.g., face-related brain activity when deciding ‘face’), but also by negative evidence (e.g., house-related brain activity when deciding ‘face’).

In both cases, it is useful to ask what it would look like for an agent to only consider positive evidence in making a decision. This soon becomes circular, because positive and negative evidence are defined with respect to the decision itself. For example, when analyzing the decisions of an agent that consistently ignores evidence for one alternative (similar to the random attention model above), both positive and negative evidence should still be predictive of decisions. The effect of positive evidence is then driven by those trials in which the agent selected the attended alternative, and the effect of negative evidence by those trials in which the agent selected the ignored alternative (because the evidence for the attended alternative was insufficient). Put differently, asymmetries of positive and negative evidence cannot affect the decision itself, because at the time of making the decision there is no positive and negative evidence, but two sources of evidence that may become positive or negative, depending on the decision. For this reason, in measuring evidence weighting in decision formation, we defined positive and negative evidence relative to the ground truth rather than the agents’ decision.

## Discrimination decisions

First, to make sure we are measuring true random fluctuations and not systematic differences between noise and signal channel, we mean centered all signal channels to 0. ‘Sum evidence’ was defined as the total sum of noise terms across both channels . ‘Relative evidence’ was defined as the difference in noise terms between the signal and non-signal channels ( when was the signal channel and when was the signal channel). All four models predicted that the probability of correctly identifying the signal channel in the discrimination task should increase with relative evidence (Fig. 3A, orange markers). The models predicted no effect of sum evidence on discrimination decisions.



*Figure* *3.*  Predictions for the reverse correlation analysis, derived from the four models. A: the effect of relative (orange markers) and sum (black markers) evidence on discrimination decision. B: the effect of evidence for the chosen (green markers) and unchosen (purple markers) alternatives on discrimination confidence. C: the effect of sum and relative evidence (defined with respect to participants’ decisions) on discrimination confidence. Panels D, F and H: The effect of evidence in the signal channel (blue markers) and in the non-signal channel (red markers) on detection decisions, confidence in yes responses, and confidence in no responses, respectively. Panels E, G, and I: The effect of relative evidence (orange markers) and sum evidence (black markers) on detection decisions, confidence in yes responses, and confidence in no responses, respectively. Stars represent significance in a two-sided t-test: \*: p<0.05, \*\*: p<0.01, \*\*\*: p<0.001.

## Discrimination confidence

The median confidence rating was used to split evidence channels into four sets, according to decision (chosen or unchosen, depending on the agent’s decision) and confidence level (high or low). Confidence kernels for the chosen and unchosen channels were then extracted by subtracting the mean low-confidence from the mean high-confidence vectors for each channel. In all four models, agents were more confident in their discrimination decisions when confidence strengthened evidence in the chosen direction (Fig. 3B, green markers) and weakened evidence in the unchosen direction (Fig. 3B, purple markers). As expected, this translated to an effect of relative evidence on decision confidence: agents were more confident when the evidence difference between the chosen and unchosen channels was high (Fig. 3C, orange markers).

Critically, only the stimulus-dependent noise and goal-directed attention models produced an effect of sum evidence on decision confidence, such that agents were more confident when overall evidence was high (Fig. 3C, black markers). As reviewed above, this is a positive evidence bias in discrimination confidence.

## Detection decisions

For the reverse correlation analysis of detection decisions, we focused on trials in which a signal was present. This allowed us to disentangle the effects of evidence in the signal and non-signal channels on detection decisions and confidence. We subtracted evidence in trials that resulted in a ‘no’ (target absent) decision from evidence in trials that resulted in a ‘yes’ (target present) decision, separately for the signal and non-signal channels. All four models predicted that agents should be more likely to respond ‘yes’ when evidence is stronger in the signal channel (Fig. 3D, blue markers). Importantly, the same was true for evidence in the non-signal channel: agents were more likely to respond ‘yes’ when evidence was stronger in this channel too (Fig. 3D, red markers). Together, these two positive effects translated to a strong effect of sum evidence on detection decisions: agents were more likely to respond ‘yes’ when the total sum of evidence was high (Fig. 3E, black markers). A weaker effect of relative evidence on detection decisions was observed in all models except for the random attention model (Fig. 3E, orange markers).

## Detection confidence

Similar to the discrimination task, the median confidence rating was used to split evidence channels into four sets, according to signal (signal channel or non-signal channel) and confidence level (high or low). This was done separately for ‘yes’ and ‘no’ responses. Confidence kernels for the signal and non-signal channels were then extracted by subtracting the mean low-confidence from the mean high-confidence evidence values for each channel and decision.

In all four models, agents were more confident in their decisions about signal presence when evidence in the signal channel was stronger (Fig. 3F, blue markers). Mirroring the detection decision kernels, confidence in signal presence was also positively affected by evidence for signal in the non-signal channel (Fig. 3F, red markers). Together, these two positive effects produced an overall positive effect of sum evidence on confidence in presence (Fig. 3G, black markers). All four models predicted a weaker effect of relative evidence (Fig. 3G, orange markers).

Finally, in all four models, agents were more confident in decisions about signal absence when evidence in both signal and non-signal channels was weaker (Fig. 3H, blue and red markers, respectively). Together, these negative effects translated to a total negative effect of sum evidence on confidence in absence (Fig. 3I, black markers). None of the four models predicted a negative effect of relative evidence on confidence in absence, but the random attention model predicted a subtle positive effect (Fig. 3I, orange markers).

# Experiment 1

## Methods

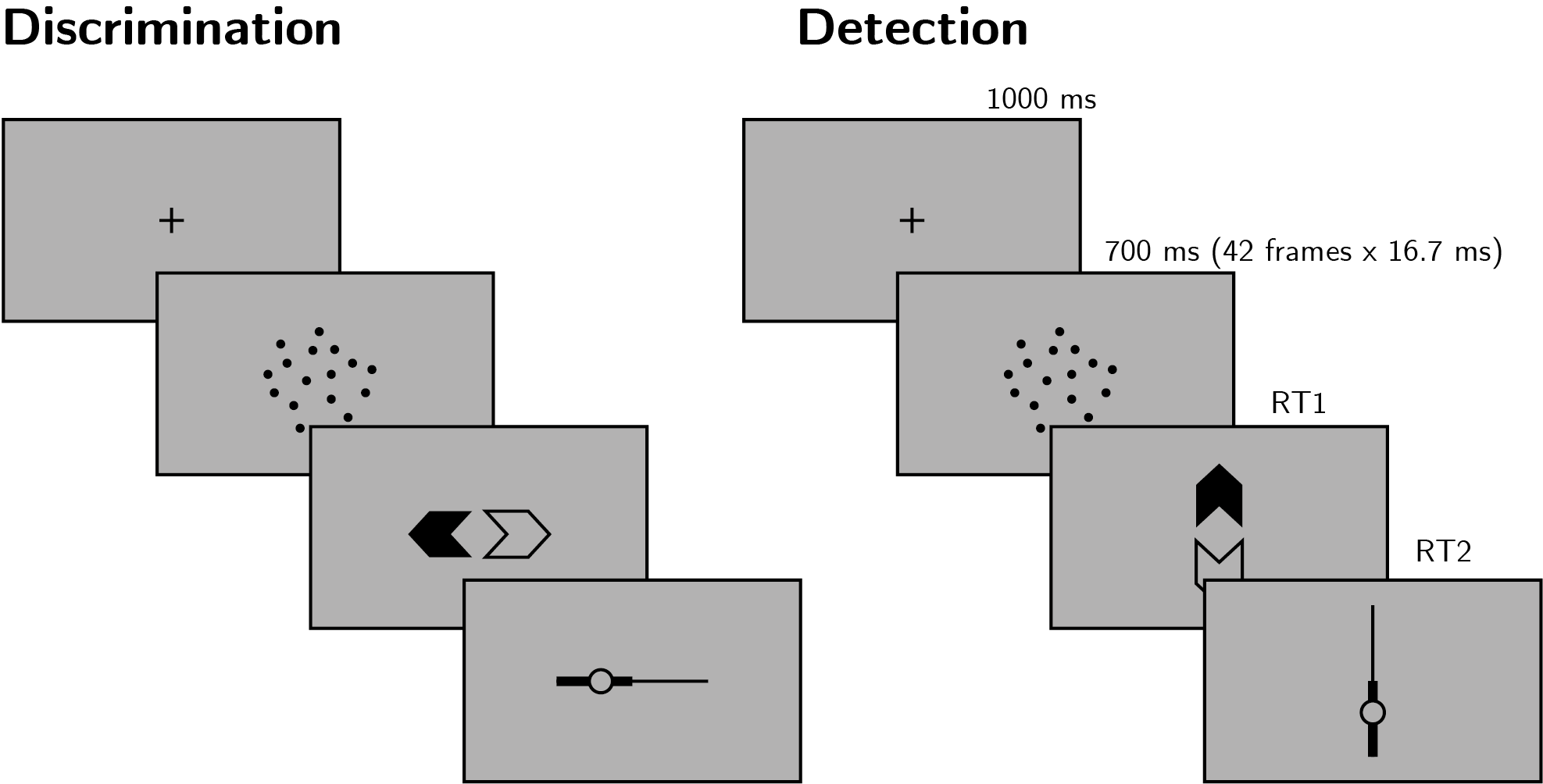
### Participants.

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.

### Experimental procedure.

The experimental procedure for Exp. 1 largely followed the procedure described in Zylberberg et al. (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 any 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 all dots 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. 4).



*Figure* *4.*  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.

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 et al. (2012)[[1]](#footnote-46). 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, where on a random half of the trials coherence was set to 0%, without random sampling of coherence values for different frames.

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 ??.

## 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 et al. (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 et al., 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. 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. Additionally, for every time point we extracted the variance along the x and y dimensions, but given a high correlation between our mean and variance estimates, we focused our analysis on the mean motion energy.

In order to distill 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 et al. (2012) in order to be sensitive to the possibility of nonlinear effects of coherence on motion energy.

# Results

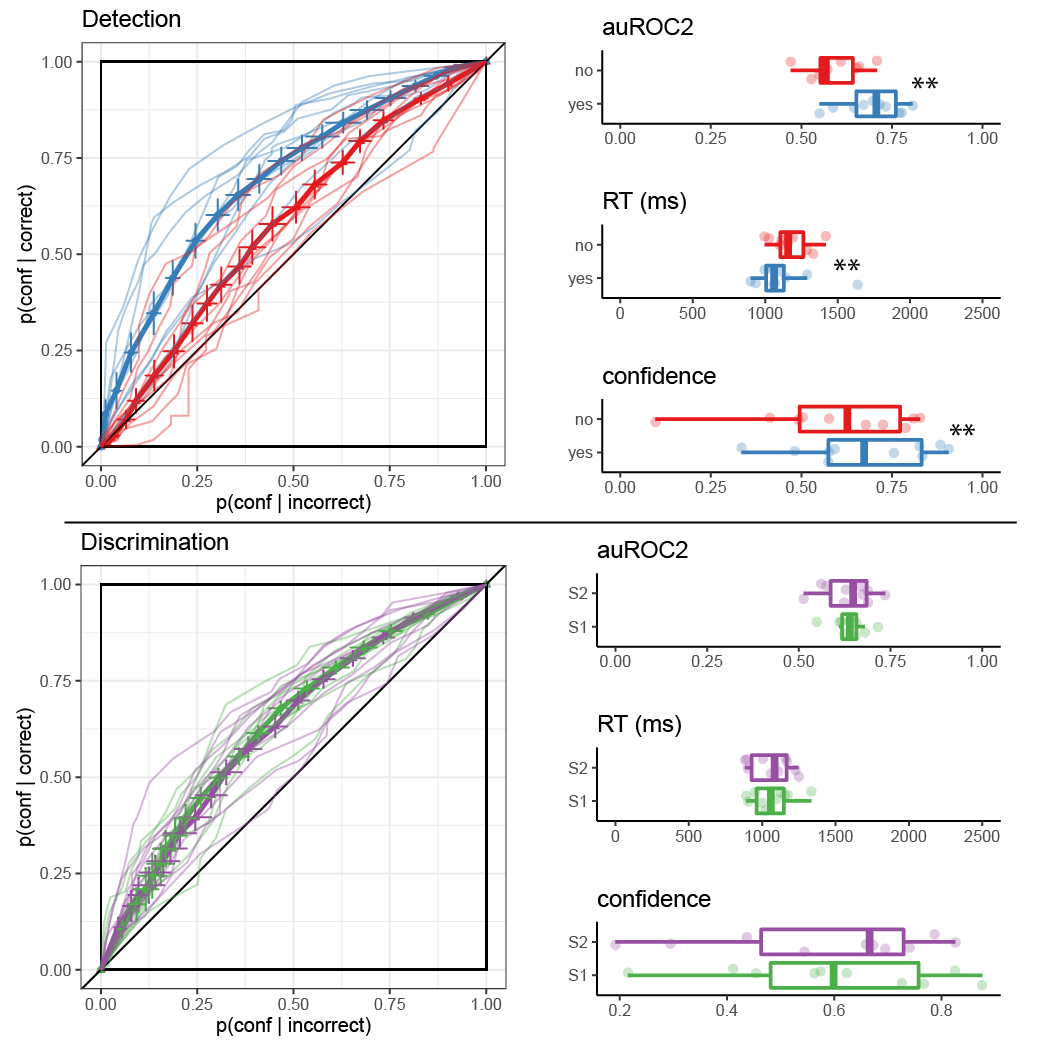
### Decision accuracy.

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.

### 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. 5). ‘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 ??.



*Figure* *5.*  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

### Reverse correlation.

#### Discrimination.

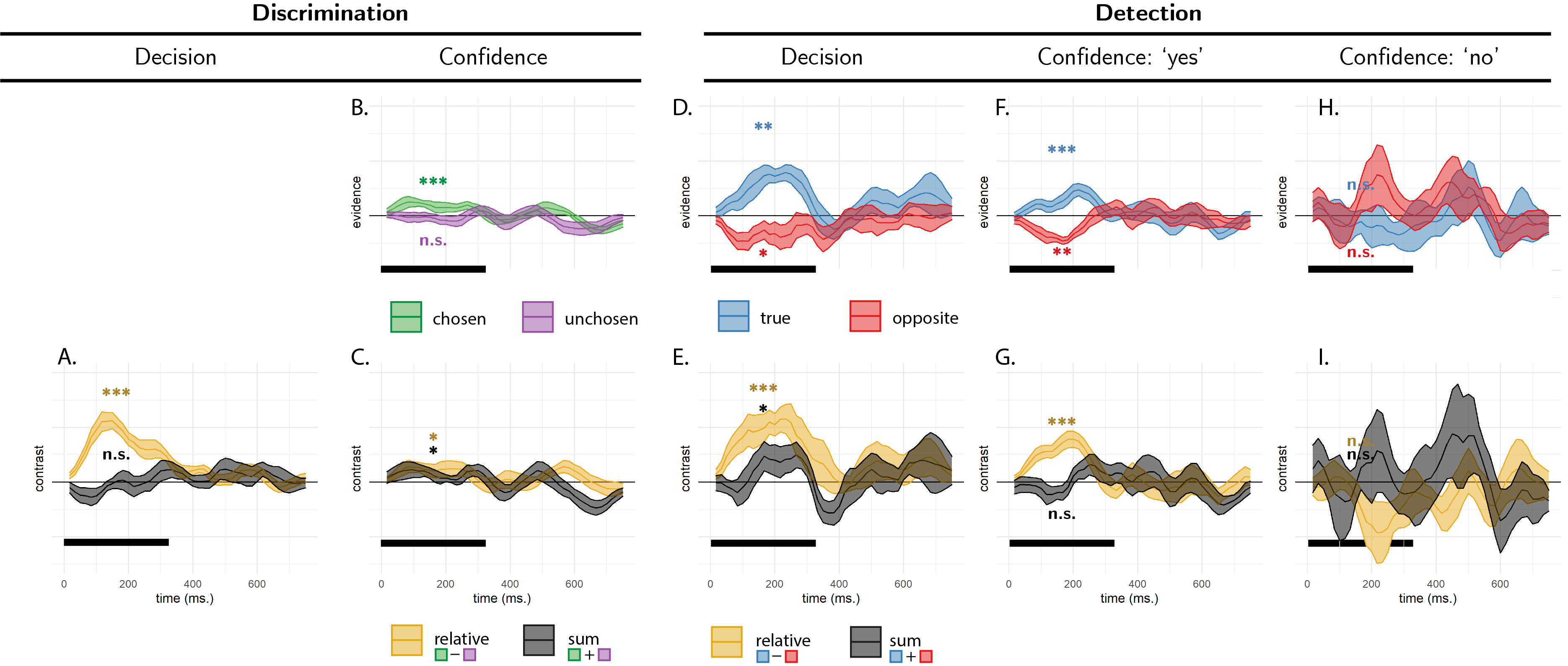
Using reverse correlation we quantified the effect of random fluctuations in motion energy on the probability of correctly identifying the true direction of motion, and on the temporal dynamics of decision formation. Following Zylberberg et al. (2012), we focused our analysis on the first 300 ms of the trial. Participants’ discrimination responses were significantly affected by the relative evidence for the true direction of motion compared to the opposite direction (, ), whereas sum evidence (the total amount of energy in both directions) had no effect on discrimination accuracy (, ; see Fig. 6A). This is consistent with a symmetric weighting of positive and negative evidence in decision formation, and with the predictions of all four models.

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 from the mean high-confidence vectors for both the chosen and unchosen directions. Motion energy in the chosen direction (positive evidence) significantly increased confidence (, ), but we found no significant decrease in confidence with stronger motion energy in the opposite direction (negative evidence; , ; see Fig. 6B). Equivalently, both relative and sum evidence positively contributed to decision confidence (relative evidence: , ; sum evidence: , ; see Fig. 6C). This is a replication of the Positive Evidence Bias observed in Zylberberg et al. (2012), and consistent with the predictions of the stimulus-dependent noise and goal-directed attention models.

#### 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 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’.

In order to approximate participants’ belief states during ‘yes’ responses, we focused our analysis on trials in which coherent motion was presented in one of the two directions (signal trials). 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.



*Figure* *6.*  Reverse correlation, Exp. 1. A: the effect of relative (black curve) and sum (orange curve) evidence on discrimination decision. Note that relative evidence here is defined with respect to the true direction of motion, not participants’ decisions. B: the effect of evidence for the chosen (green curve) and unchosen (purple curve) alternative on discrimination confidence. C: the effect of sum and relative evidence (defined with respect to participants’ decisions) on discrimination confidence. Panels D, F and H: The effect of evidence for the true direction of motion (blue curve) and for the opposite direction of motion (red curve) on detection decisions, confidence in yes responses, and confidence in no responses, respectively. Panels E, G, and I: The effect of relative evidence (black curve) and sum evidence (orange curve) on detection decisions, confidence in yes responses, and confidence in no responses, respectively. The first 300 milliseconds of the trial are marked in yellow. Stars represent significance in a two-sided t-test for the first 300 milliseconds of the trial: \*: p<0.05, \*\*: p<0.01, \*\*\*: p<0.001.

Participants were significantly more likely to respond ‘yes’ when fluctuations in motion energy during the first 300 milliseconds of the trial strengthened motion energy in the true direction of motion (, ; see Fig. 6D, blue curve). Critically, and in contrast to the predictions of all four models, motion energy in the opposite direction had a negative, rather than a positive effect on detection decision (, ; see Fig. 6D, red curve).

Confidence ratings were higher in detection ‘yes’ responses when random noise strengthened the motion energy in the true direction of motion (, ; see Fig. 6F, blue curve). Again, in contrast to our model predictions, motion energy in the opposite direction had a negative, rather than a positive effect on detection decisions (, ; see Fig. 6F, red curve).

Furthermore, unlike in the discrimination task, here we found no effect of sum evidence on confidence ratings (, ; see Fig. 6G, black curve). 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. 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.

Finally, confidence in ‘no’ responses was independent of relative, sum, positive, and negative evidence (all p’s>0.1; see Fig. 6H).

# Experiment 2

In Exp. 1, we replicated previous observations of a positive evidence bias in discrimination confidence, such that confidence scaled with the total sum of evidence for both hypotheses. In contrast, in detection an effect of sum evidence was apparent for the decision, but not for the confidence kernels. Furthermore, confidence in detection ‘no’ responses was not at all affected by fluctuations in motion energy.

In Exp. 2 we tested the generalizability of these findings to 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/>) 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.

## Methods

### 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/>). 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.

### Experimental paradigm.

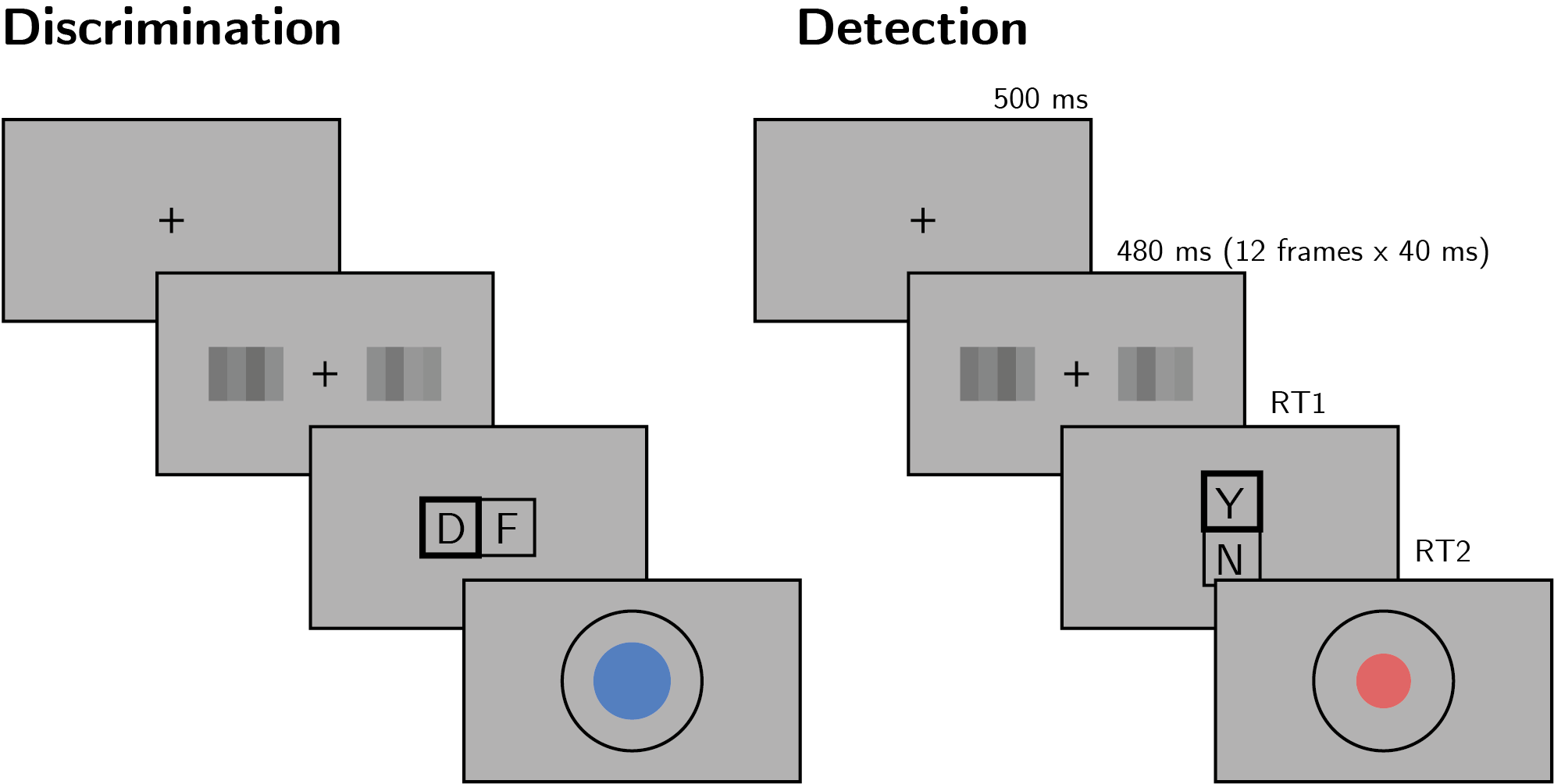
The experiment was programmed using the jsPsych and P5 JavaScript packages (De Leeuw, 2015; McCarthy, 2015), and was hosted on a JATOS server (Lange, Kuhn, & Filevich, 2015). It consisted of two tasks (Detection and Discrimination) presented in separate blocks. A total of 56 trials of each task was delivered 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 et al. (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. 7). 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 decision accuracy was delivered after the confidence rating phase.



*Figure* *7.*  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.

### 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).

## Results

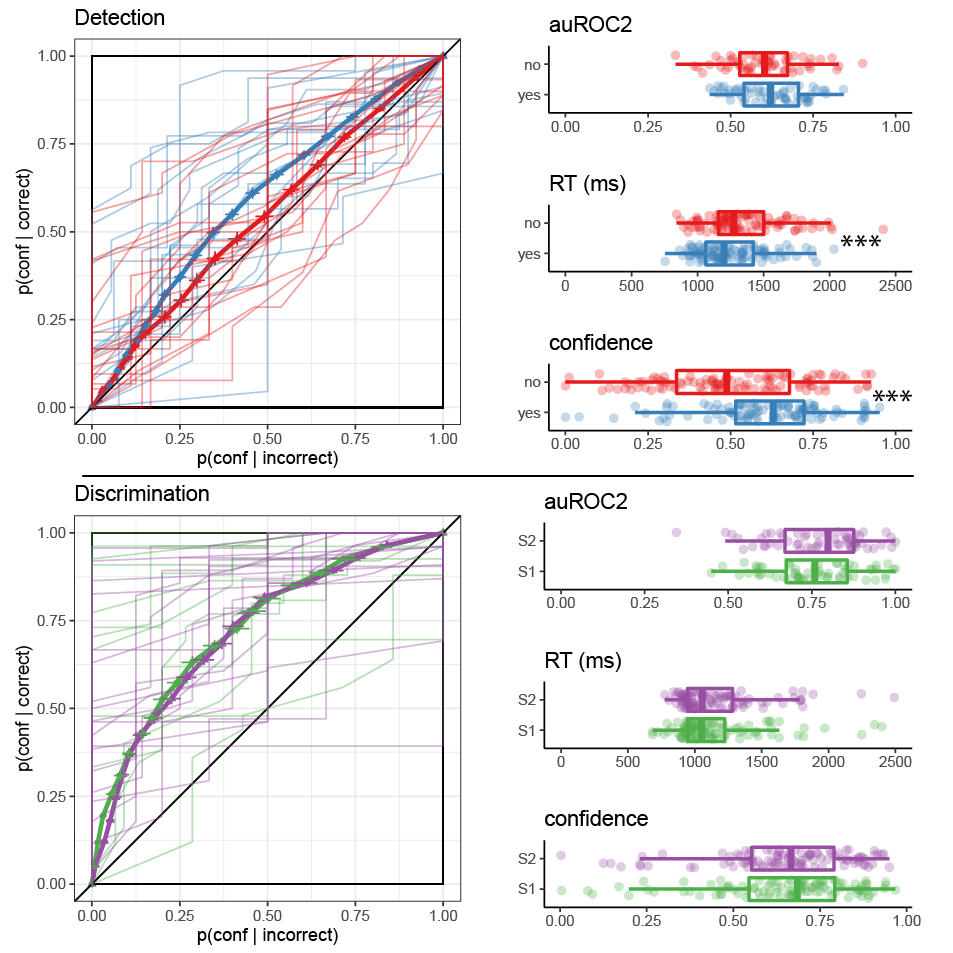
### Decision accuracy.

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).

### Overall properties of decision 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. 8), 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 ??.



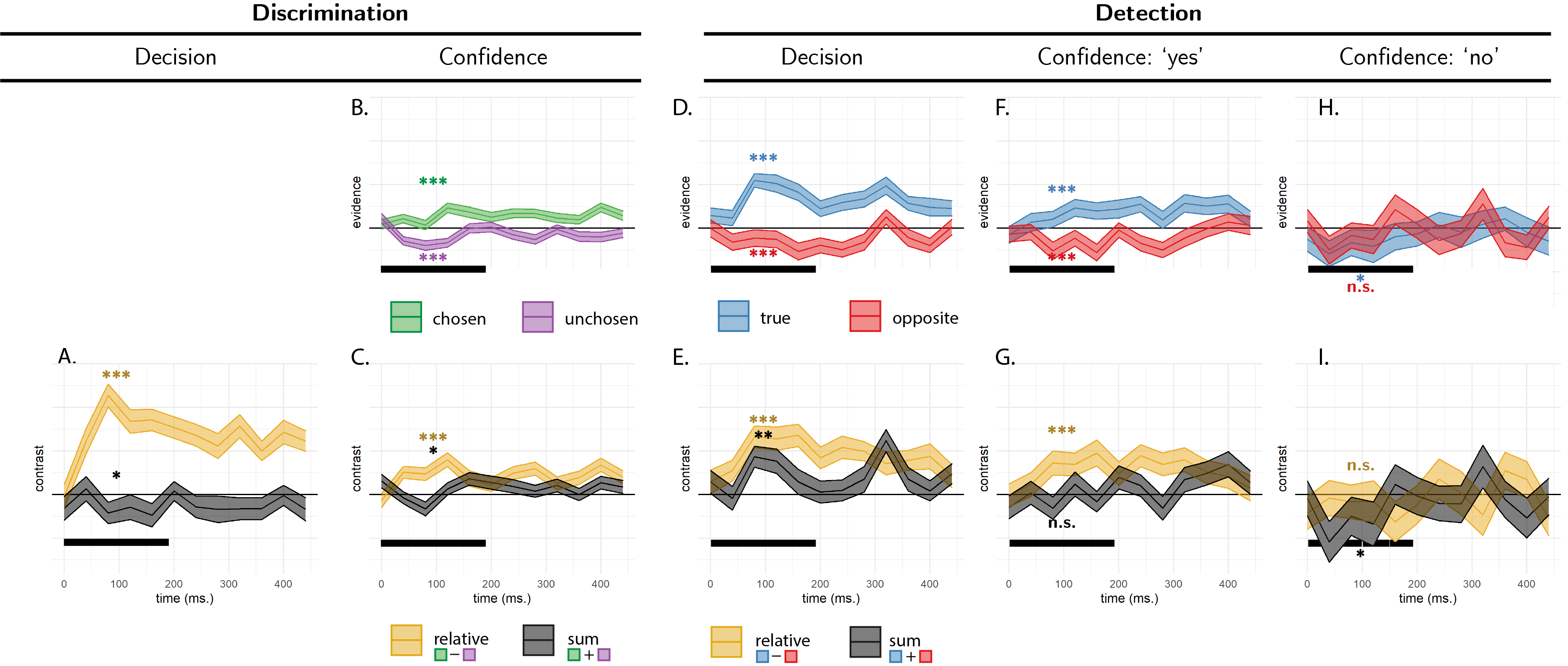
*Figure* *8.*  Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence, in Exp. 2. Same conventions as in Fig. 5.

### 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 et al. (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 relative evidence (the difference in mean luminance between the left and right stimulus) during the first 300 milliseconds of the trial (, ; see Fig. 9A, orange curve). Furthermore, participants’ decisions were surprisingly more sensitive to negative compared to positive evidence within the same time window, resulting in a negative effect of sum evidence (, ; see Fig. 9A, black curve). Importantly, this negative effect of sum evidence on decision accuracy was not replicated in Exp. 3 and 4.



*Figure* *9.*  Reverse correlation, Exp. 2. Same conventions as in Fig. 6.

#### Discrimination confidence.

Similar to Exp. 1, here too we observed a significant effect of positive (, ) and negative (, ) evidence on decision confidence within the first 300 milliseconds of the stimulus (see Fig. 9B). Replicating Zylberberg et al. (2012), this effect was significantly stronger for the chosen stimulus, producing an effect of sum evidence on decision confidence that is consistent with a positive evidence bias(, ; see Fig. 9C, black curve).

#### Detection.

Participants’ detection decisions were sensitive to fluctuations in the luminance of the target stimulus, such that ‘yes’ responses were associated with a brighter target stimulus (, ; see Fig. 9D, blue curve). Similar to Exp. 1, and in contrast to the behaviour of Bayes-rational simulated agents, the luminance of the non-target stimulus had a negative effect on the probability of responding ‘yes’ (, ; see Fig. 9D, red curve).

Confidence in detection ‘yes’ responses was similarly sensitive to fluctuations in the luminance of the target stimulus (, ; see Fig. 9F, blue curve). Again, brighter non-target stimuli made participants less, rather than more, confident in the presence of a signal (, ; see Fig. 9F, red curve). As in Exp. 1, here too sum evidence (overall luminance) had no significant effect on confidence in detection ‘yes’ responses (, ; see Fig. 9G, black curve). This surprising result was not replicated in Experiments 3 and 4.

Finally, confidence in detection ‘no’ responses was sensitive to random fluctuations in the luminance of the target, such that participants were more confident in the absence of a signal when the target stimulus was darker (t(96)=-2.28, p=.025; see Fig. (**ref?**)(fig:RC\_exp2-RC)G). The overall luminance of the display also had a negative effect on confidence in detection ‘no’ responses (, ; see Fig. (**ref?**)(fig:RC-exp2-RC)I). The luminance of the non-target stimulus (, ) and the difference in luminance between the two stimuli (, ) had no significant effects on confidence in detection ‘no’ responses.

# 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.

## Methods

### 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.

### 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. 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.

## Results

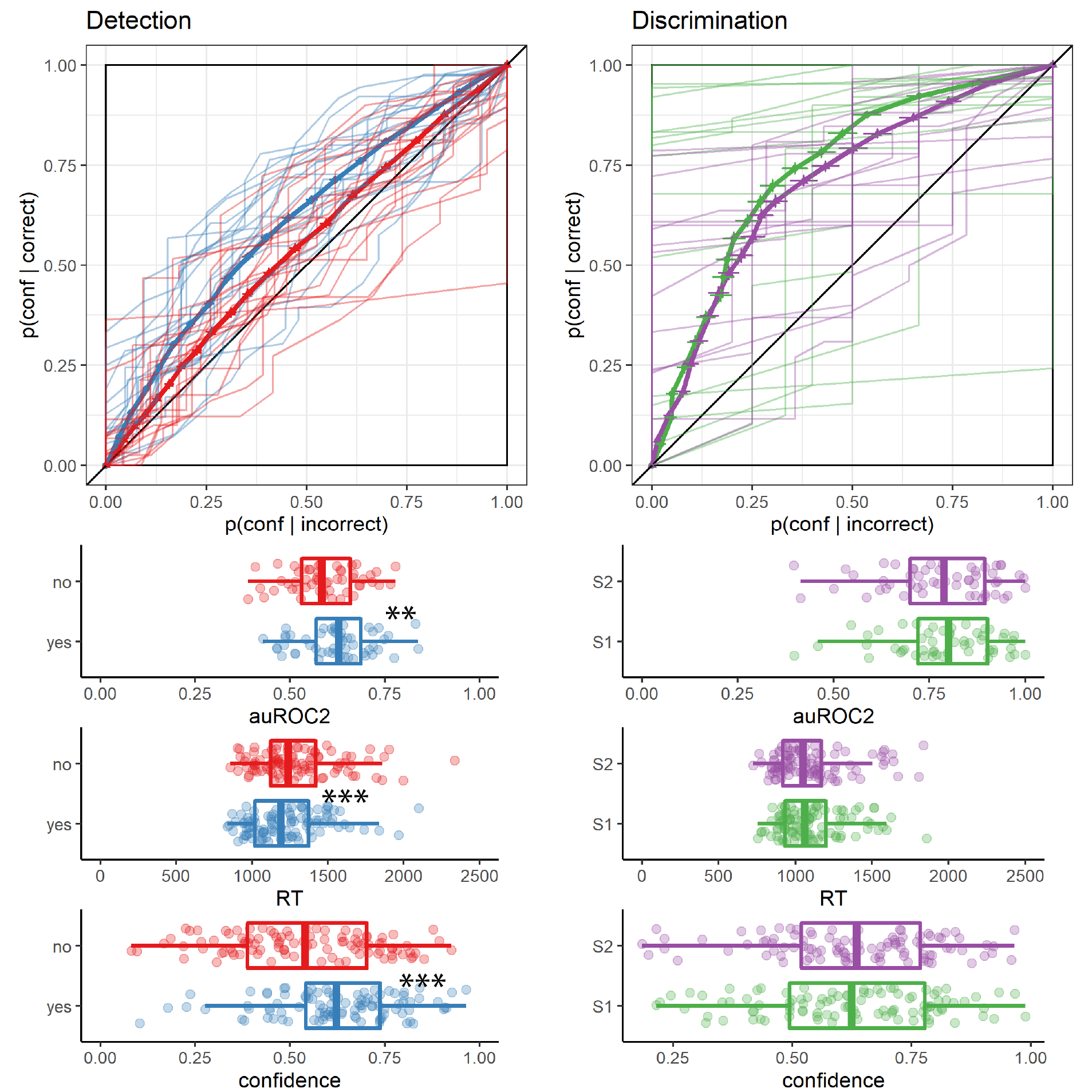
### Decision accuracy.

Overall proportion correct was 0.86 in the discrimination and 0.66 in the detection task. Performance for discrimination was significantly higher than for detection (, 95% CI , , ), as expected.

### Overall properties of decision and confidence distributions.

Similar to Exp. 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 Exp. 1 and 2, we also found behavioural asymmetries between the two detection responses (see Fig. 10), with ‘yes’ responses being faster (median difference of 68.90 ms) and accompanied by higher levels of confidence (mean difference of 0.07 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 ??.



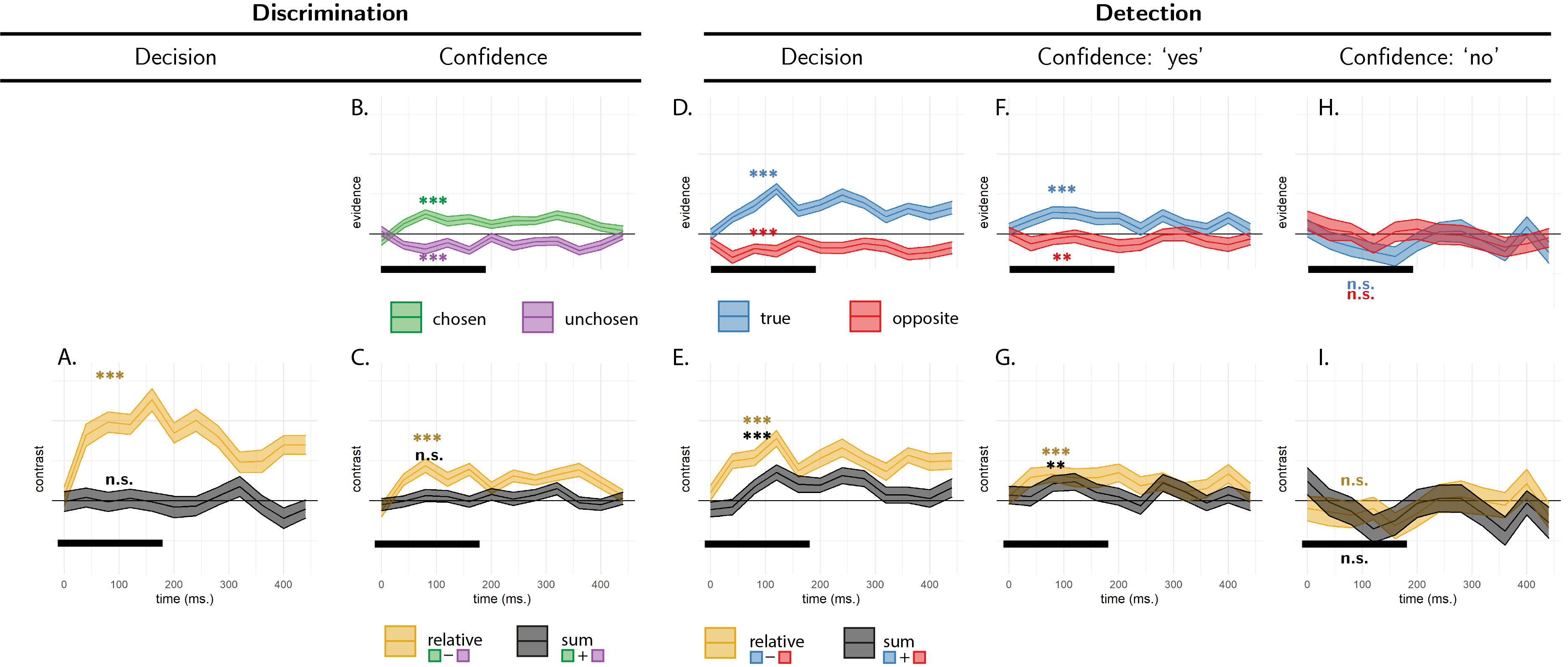
*Figure* *10.*  Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence, in Exp. 3. Same conventions as in Fig. 5.

### Reverse correlation.

We first focused on reverse correlation analyses pooling data from high-luminance and standard trials (after mean-centering luminance in both), 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 ??).

#### Discrimination decisions.

Discrimination decisions were sensitive to relative evidence during the first 300 milliseconds of the trial (, ; see Fig. 11A) with no effect of sum evidence (, ).



*Figure* *11.*  Reverse correlation, Exp. 3. Same conventions as in Fig. 6.

#### Discrimination confidence.

Decision confidence was sensitive to positive (, ) and negative (, ) evidence within the first 300 milliseconds of the stimulus (see Fig. 11B). Reverse correlation revealed no effect of random fluctuations in sum evidence on decision confidence (, ), but an effect of sum evidence was found when directly contrasting high- and low-luminance trials, as we show in the “Evidence-weighting” Section below.

#### Detection.

Participants’ detection decisions were sensitive to fluctuations in the luminance of the target stimulus, such that ‘yes’ responses were associated with brighter target stimuli (, ; see Fig. 11D, blue curve). Replicating the surprising results of Exp. 1 and 2, the luminance of the non-target stimulus had a negative effect on the probability of responding ‘yes’ in the detection task (, ; see Fig. 11D, red curve) Together, detection decisions were sensitive to relative evidence (the difference in luminance between the target and non-target stimuli;, ), and to sum evidence (the overall luminance of the display; , ; see Fig. 11E, orange and black curves, respectively).

Confidence in detection ‘yes’ responses was similarly positively correlated with the luminance of the target stimulus (, ), and negatively correlated with the luminance of the non-target stimulus (, ; see Fig. 11F). This is again in contrast to what is expected from a Bayes-rational agent: the proportion of being correct is positively correlatd with evidence intensith in both signal and non-signal channels. 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; , ; see Fig. 11G, black curve).

Finally, confidence in detection ‘no’ responses was marginally sensitive to random fluctuations in the luminance of the target, such that participants were more confident in the absence of a signal when the target stimulus was darker (, ; see Fig. @ref(fig:RC\_exp3-RC)H). Relative evidence also had a marginally significant negative effect on confidence in decisions about absence (, ). The luminance of the non-target stimulus and the overall luminance had no significant effects on confidence in detection ‘no’ responses (p’s>0.3).

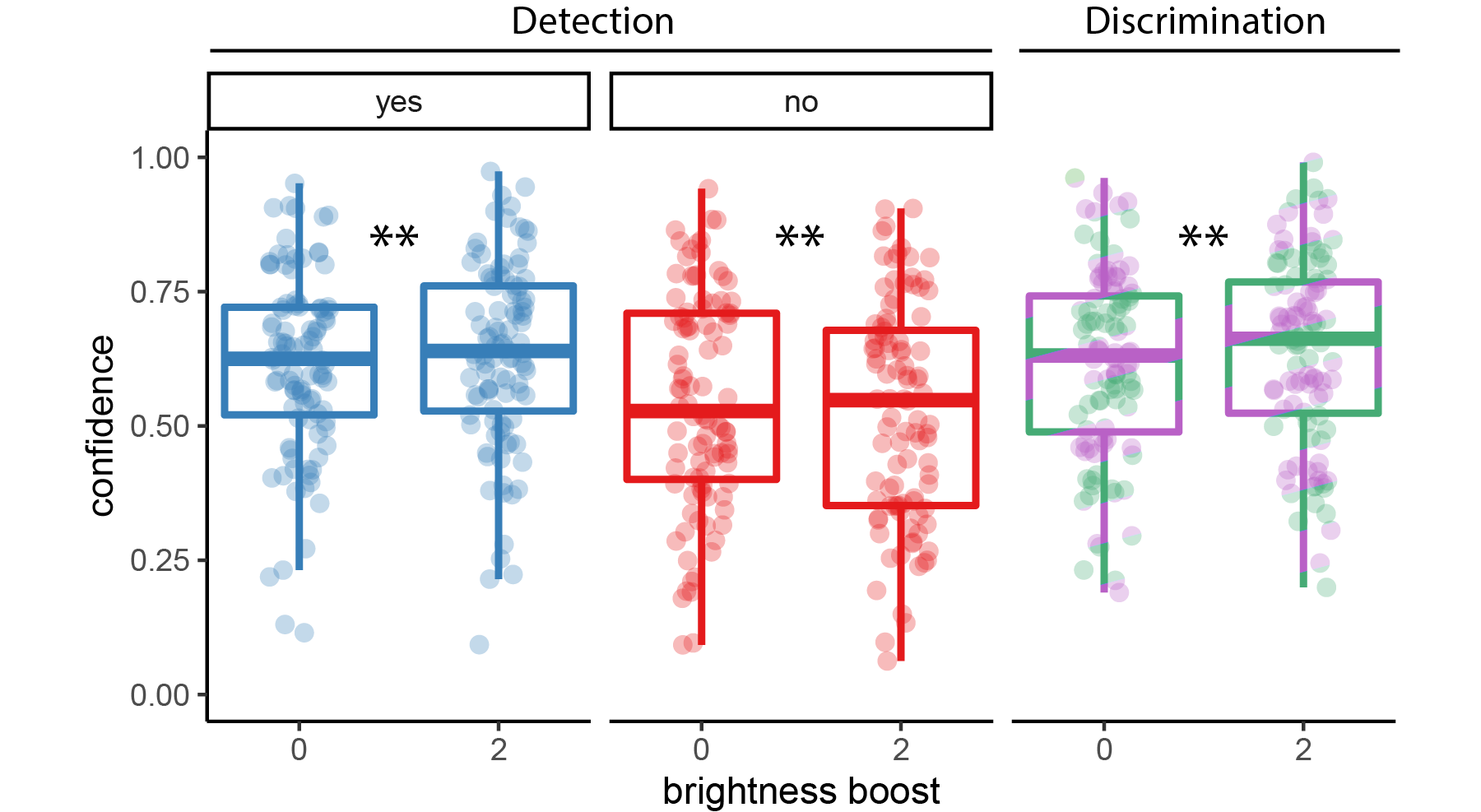
### Evidence-weighting.

In Experiments 1 and 2, confidence in 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 evidence 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 only partly replicated in Exp. 3: subjects still negatively weighted the luminance of the non-target stimulus (despite being irrational), but this negative effect was weaker than the positive effect of the luminance of the target stimulus, resulting in an overall positive effect of sum evidence on detection confidence.

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 the effects of sum luminance on discrimination and detection decisions and confidence.

First, we established that 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.



*Figure* *12.*  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

# Experiment 4

A limitation of Exp. 2 and 3, raised by an anonymous Reviewer, is that apparent asymmetries in the weighting of positive and negative evidence may result from a nonlinear mapping between luminance in RGB space and screen brightness. For example, a dark bar that is -2 RGB units from the mean does not necessarily cancel out a bright bar that is +2 RGB units from the mean, making positive evidence objectively more salient than negative evidence.

To attend this concern, we include an additional Experiment where evidence is sampled from a perceptually uniform space. Specifically, Exp. 4 was similar to Exp. 3 with the exception that flickering stimuli varied in their hue rather than luminance, and where hue values were sampled from a Gaussian distribution in the CIE L\*a\*b\* colour space. Moreover, the roles of ‘target’ and ‘non-target’ hues were counterbalanced between participants, such that any built-in asymmetries in the perception of positive and negative evidence should cancel out at the group level.

## Methods

### 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). 117 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/9zbpc>). 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.

### Experimental paradigm.

Experiment 4 was identical to Experiment 3 with two changes. First, flickering bars were varied in hue, randomly sampled from a Gaussian distribution in the CIE L\*a\*b\* colour space, centred at , and , with a radius of 49 (Schurgin, Wixted, & Brady, 2020). For half of the participants, non-target hues were sampled around an orientation of 1.85 radians, with a standard deviation of 0.35 and target hues were sampled around an orientation of 2.1 with a standard deviation of 0.35. For the first group, target patches were little more orange than non-target patches, and for the second group target patches were little more green than non-target patches. To make sure non-target patches were perceived as the absence of signal relative to the background, the RSVP display was overlaid on top of a rectangle with the mean color of a non-target patch. Second, in order to avoid interference, the confidence circle was presented in gray. Third, subjects were allowed to repeat the multiple choice questions up to three times. Finally, feedback about overall performance and mean confidence in correct and incorrect responses was displayed at the end of each block. # References

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., 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., et al.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.

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

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172.

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. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *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.

# Open Practices Statement

The data and materials for all experiments are available at github.com/matanmazor/reverseCorrelation. All three experiments were pre-registered (Exp. 1: <https://osf.io/z2s93/> ; Exp. 2: <https://osf.io/d3vkm/>; Exp. 3: <https://osf.io/hm3fn/>)

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-46)