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

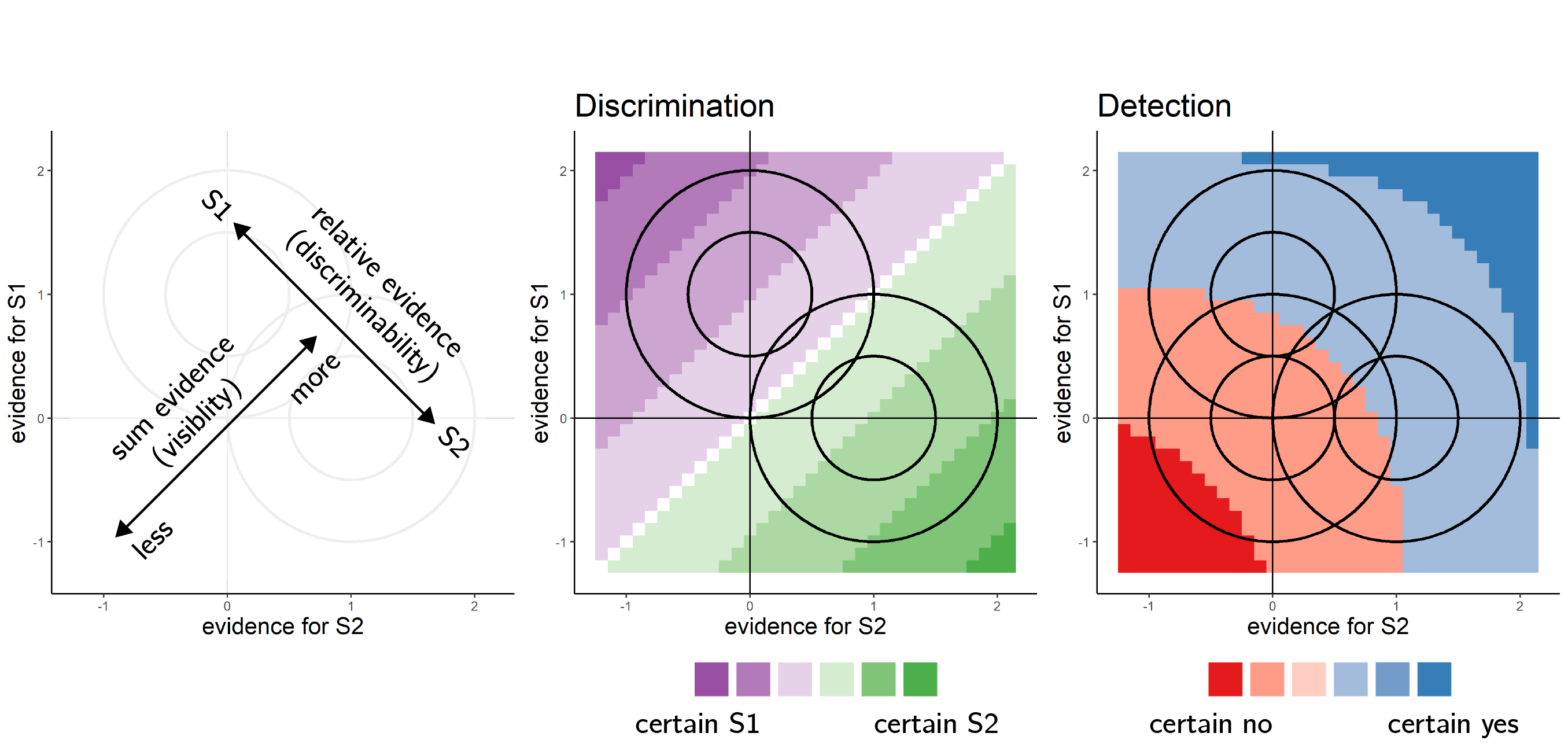
*Word count:* X

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: 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.

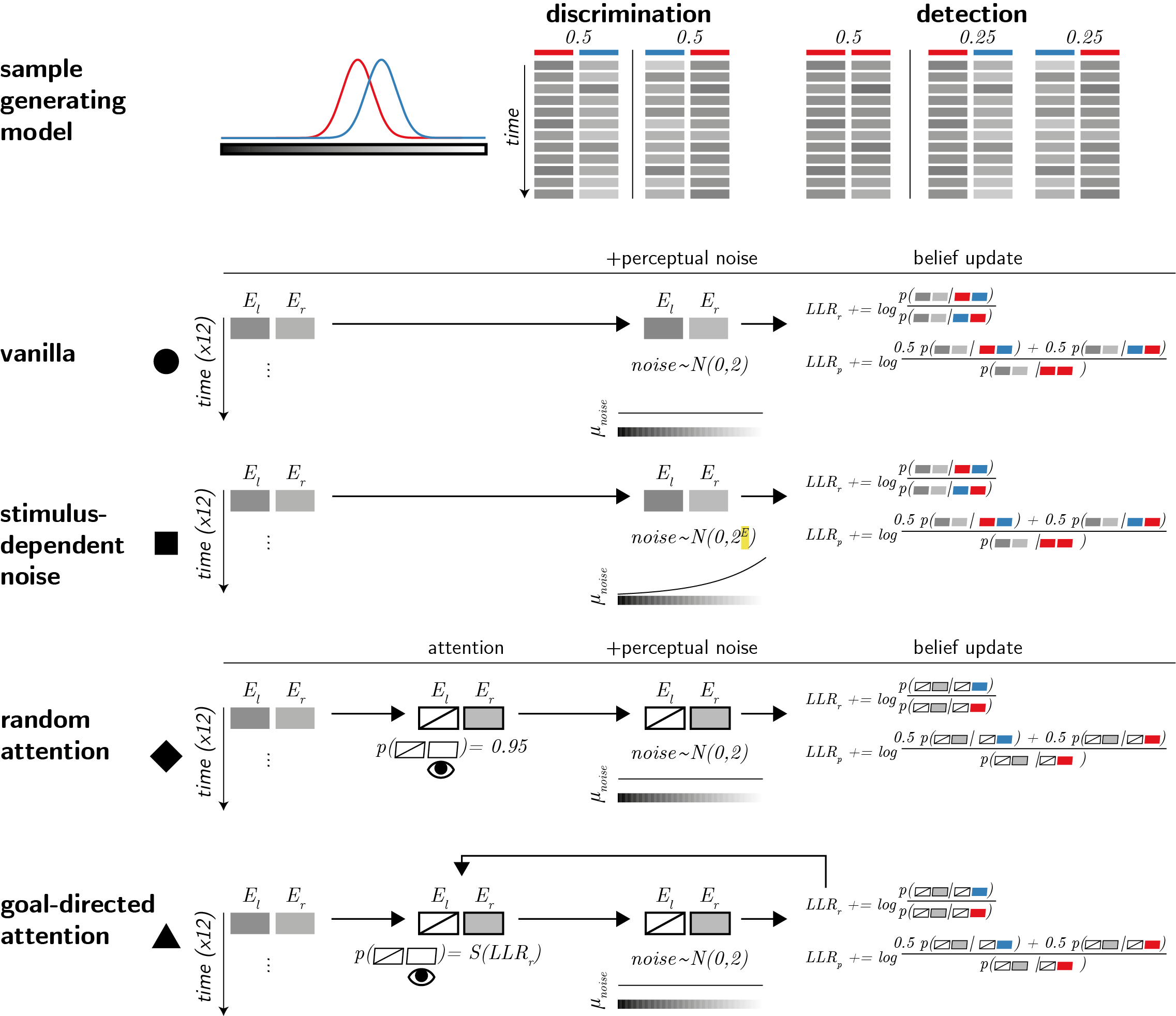
A surprising finding is that despite the irrelevance of sum evidence to the accuracy of discrimination decisions, people are systematically more confident in perceptual decisions when sum evidence is high. For example, Zylberberg, Barttfeld, and Sigman (2012) had subjects judge which of two flickering 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.

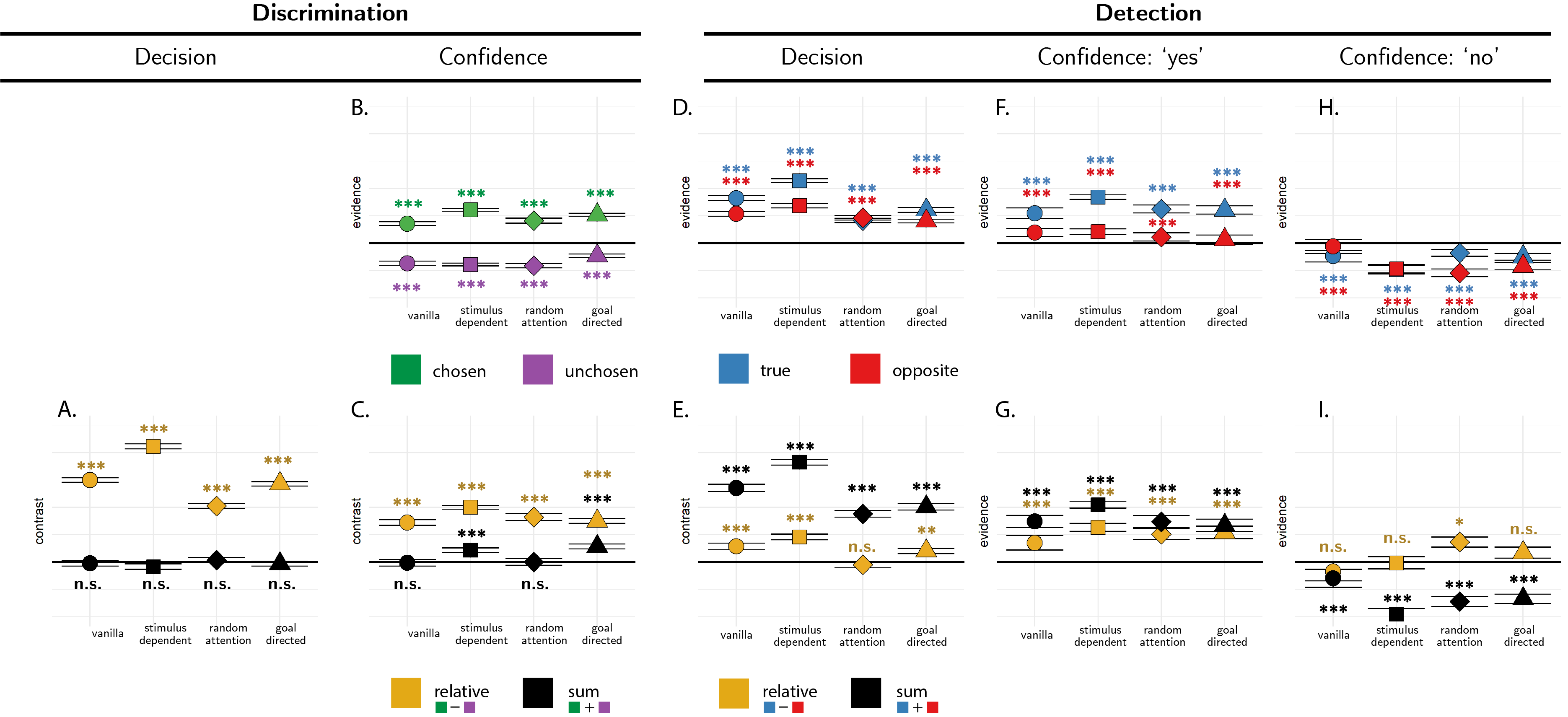
## Methodological note: positive evidence bias in perceptual decisions

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

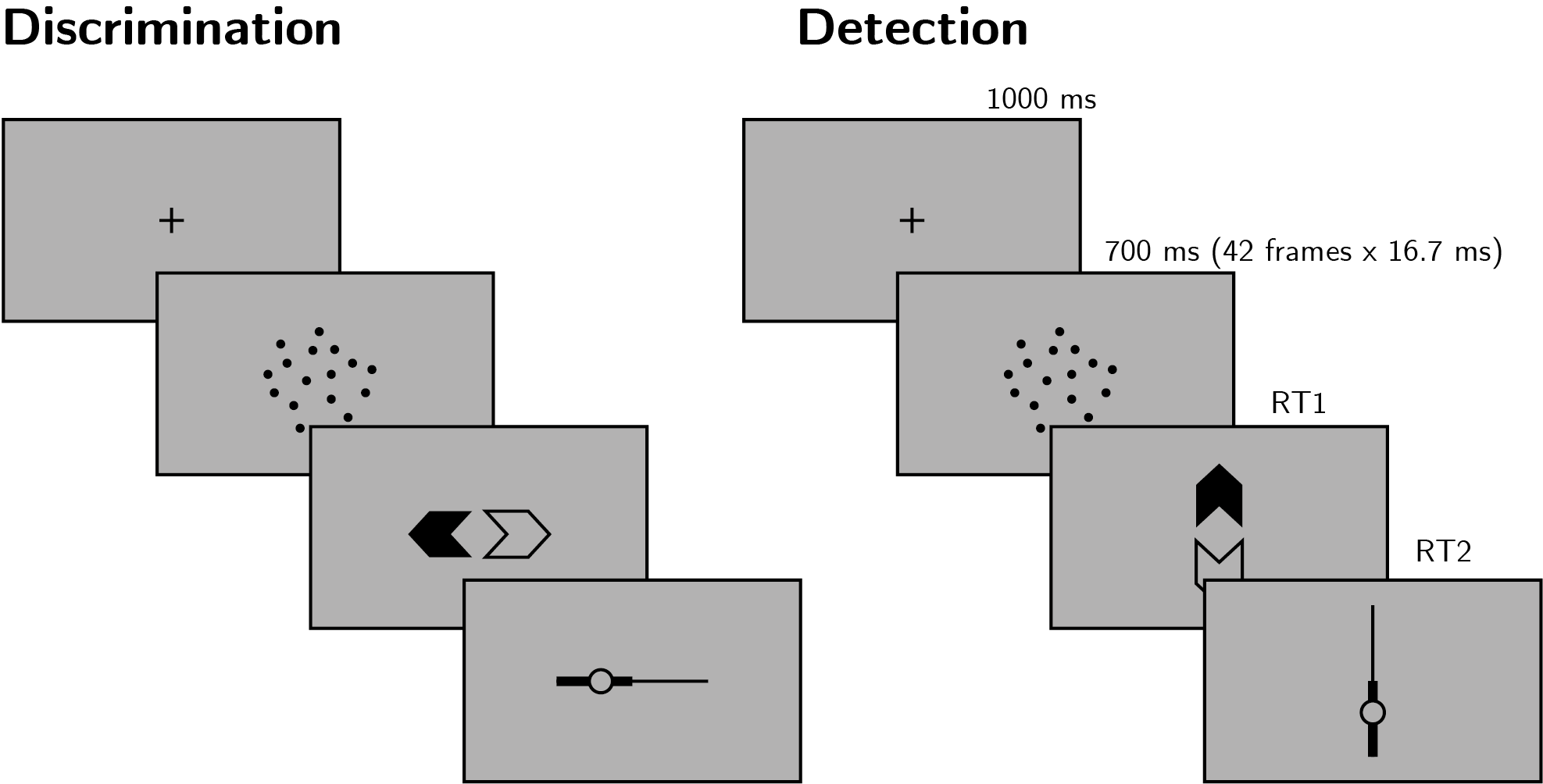
### Participants.

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

# References

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

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

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.

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*.

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*.

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)