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

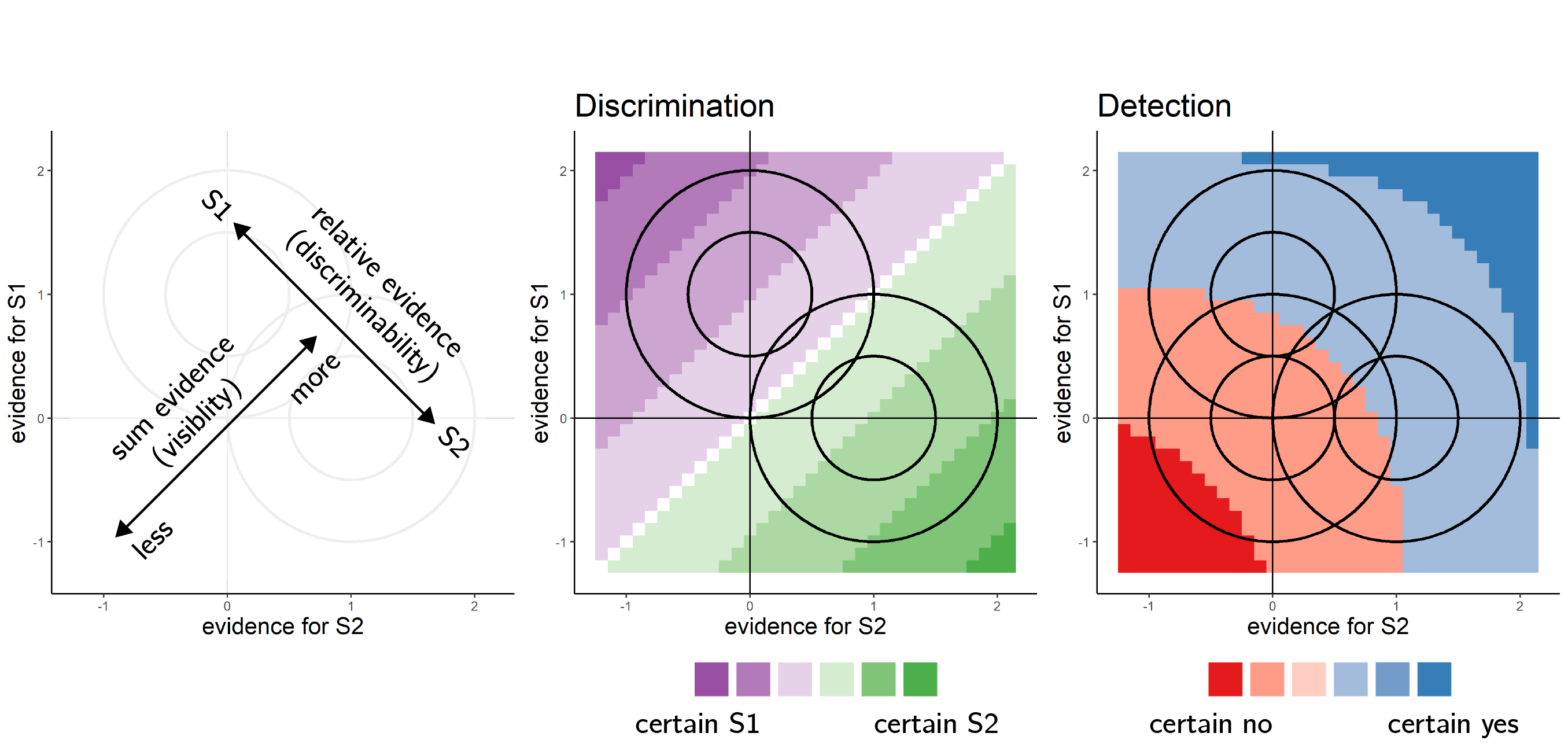
*Word count:* X

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.

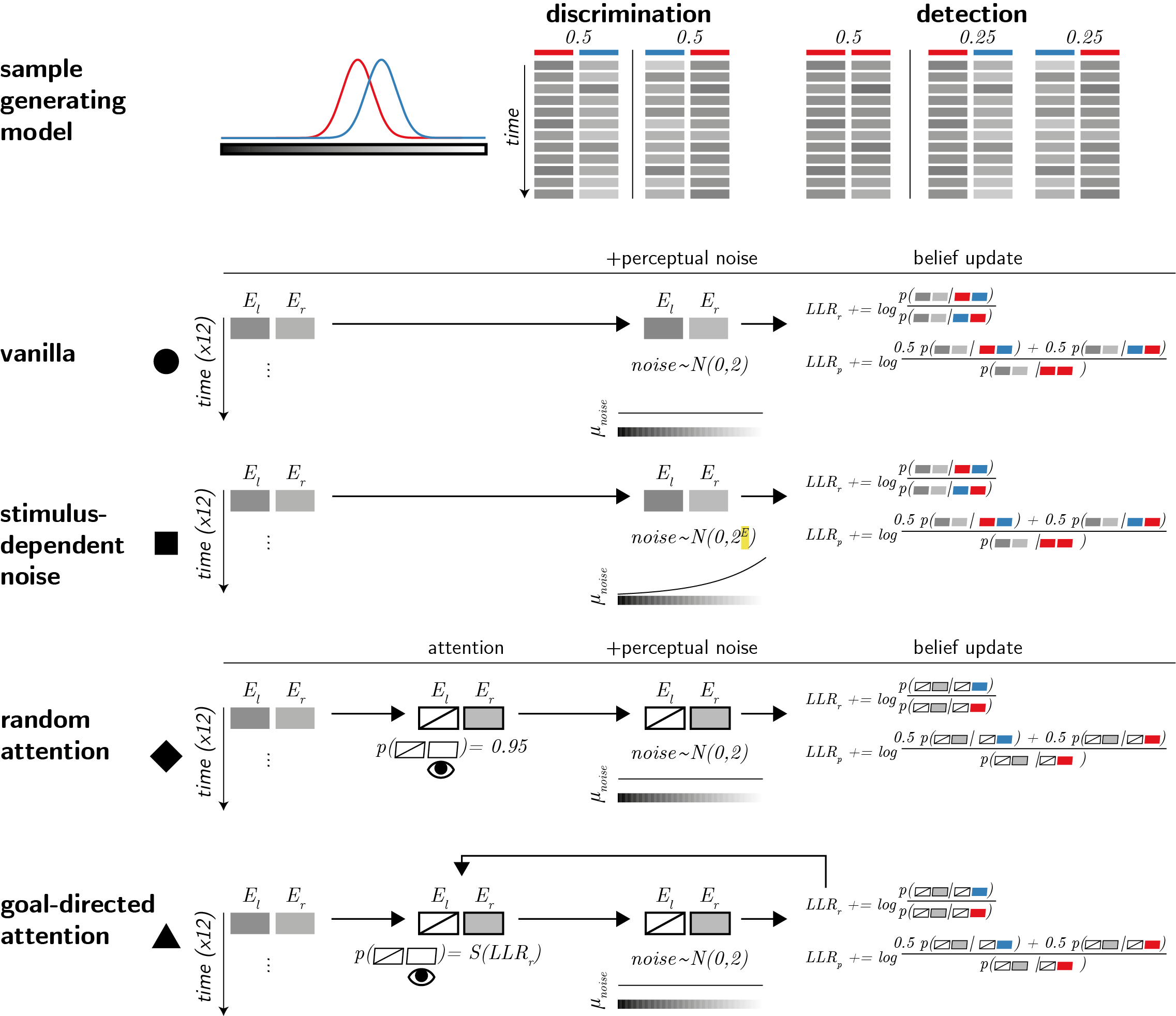
A surprising finding is that despite the irrelevance of sum evidence to the accuracy of discrimination decisions, people are systematically more confident in perceptual decisions when sum evidence is high. For example, Zylberberg, Barttfeld, and Sigman (2012) had subjects judge which of two flickering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

# 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/>)