Dear Prof. Shomstein,

We are grateful to both you and the four Reviewers for the detailed evaluation of our manuscript “*Evidence weighting in confidence judgments for detection and discrimination*”, and we apologize for our delay in submitting a revision. As you will see, we have taken your feedback very seriously and made substantial changes and additions to the manuscript in response, including running an additional experiment and re-writing large parts of the Introduction and Discussion.

Specifically, in this revised version we:

1. Include simulations from four Bayes-rational computational models, and compare their predictions to observed behaviour.
2. Include a fourth experiment in order to make sure previous findings are not solely driven by nonlinear effects of screen luminance on perception.
3. Identify a persistent suboptimality in subjects’ detection decisions and confidence, and organize the Discussion around this ‘negative evidence bias’.

Below, we address Reviewers’ comments and questions in detail, referring to specific changes we have made to the paper when relevant (all changes appear in red).

With best wishes,

Matan Mazor, on behalf of all authors

# Reviewer 1

“This paper looks at the weighting of evidence (from reverse-correlation analysis) over time guiding both first-order decisions and confidence judgements in both detection and discrimination experiments using random-dot motion and random-luminance flicker stimuli. The results are mostly consistent across experiments with some important exceptions. The works is carefully carried out and is a substantial contribution. I'm a slightly outsider to the specifics under examination (the "positive-evidence bias") and my main complaint is that I found the paper difficult to follow. It wasn't always clear how the data were cooked nor where in a given figure I was supposed to see the effect being described. This is all easily fixable, however. My other disappointment, unsurprisingly, is that the modeling was relegated to a final appendix and minimal discussion, rather than making an attempt to fit multiple alternative models to the data. I understand that the one model they describe made qualitatively incorrect predictions, but for me that was the most interesting bit ;^)”

We thank the Reviewer for these two comments. We now go into more detail in describing our analysis approach and re-plot key figures in a way that should facilitate understanding. We also present simulations from three additional computational models, and discuss their predictions in relation to empirical data. We elaborate on each of these changes below.

“5/42-6/24: At this point in reading the manuscript, I found all of this pretty opaque.”

We have now rewritten the Introduction to be more clear to all readers, including those who are not familiar with the literature on the PEB:

“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).

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).”

“8/40-44: What was the initial coherence value and how was it deterrmined? This is a very slow staircase procedure and it's unclear why you did it this way.”

We used an initial coherence value of 1. We opted for this slow staircasing procedure due to the difficulties with staircasing a detection task, where the optimal placement of a criterion depends on knowledge about the expected signal strength. In order to keep the expected motion energy stable across trials, and to avoid reductions in motion energy after a series of consecutive correct rejections, we used this (admittedly slow) block-based staircasing approach. We now mention this in the text:

“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%. We opted for a block-wise staircasing procedure in order to keep motion energy relatively stable across trials, allowing participants to optimally place their detection criterion. The staircasing procedure for both tasks started at a coherence value of 1.0.”

“10/51: "ensures registration time-locking": I have no idea what this means, nor why the initialization ensures it.”

We thank the Reviewer for prompting a clarification for our original terse description. We now expand on this in the Open Practices Statement:

“The data and materials for all experiments, including demos of the experiments, full analysis code, model simulations, and a fully-reproducible version of the manuscript in Rmarkdown are available at github.com/matanmazor/reverseCorrelation. All four experiments were pre-registered (Exp. 1: https://osf.io/z2s93/ ; Exp. 2: https://osf.io/d3vkm/; Exp. 3: https://osf.io/hm3fn/; Exp. 4: https://osf.io/9zbpc). To ensure preregistration time-locking (in other words, that preregistration preceded data collection), we employed randomization-based preregistration. We used the SHA256 cryptographic hash function to translate our preregistered protocol folder (including the pre-registration document) to a string of 256 bits. These bits were then combined with the unique identifiers of single subjects, and the resulting string was used as seed for initializing the Mersenne Twister pseudorandom number generator prior to determining all random aspects of the experiment, including the order of trials, motion energy in Exp. 1, random luminance values in Exp 2 and 3, and hue values in Exp. 4. This way, experimental randomization was causally dependent on, and therefore could not have been determined prior to, the specific contents of our preregistration document (Mazor, Mazor & Mukamel, 2019) . Protocol folders and their hashed sums,as well as the relevant lines of code in our Experiment code, are available on the project's [Github page](https://github.com/matanmazor/reverseCorrelation#pre-registration-time-locking-).”

“11/36-48: This is also opaque and a bit inside baseball (inside cricket? rounders?). First, what's a motion energy "vector" (I think you mean across time steps/frames, but it's never explained)? What does it mean to extract motion "at the group level"? You don't explain the Zylberberg et al. approach, so that doesn't help explain why what you did was better. You also don't justify a worry about nonlinear effects of coherence on motion energy (and I'd expect that to be pretty linear). And, this whole paragraph is about a computation on your stimuli without any preview of why you are doing that and how you plan on relating these vectors to responses, so the reader has no way of guessing what considerations are important here.”

We thank the Reviewer for prompting a clarification on these issues.

We now explain what we mean by ‘motion energy vectors’:

“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. We refer to these temporal estimates as motion energy vectors, where each such vector consists of 42 entries, one per timepoint.“

We also clarify how we extracted mean motion energy vectors:

“The mean motion energy vectors were extracted by averaging the motion energy vectors of all participants, separately for each motion coherence level and motion direction.“

More importantly, we now include a detailed description of the reverse correlation analysis in the Computational Modelling subsection.

“12/44: Are you pooling rightward with upward? Why? Seems ad hoc.”

In Exp. 1, half of the participants made discrimination decisions between up and right motion and used the up and down keys for reporting their detection decisions. These participants used the ‘up’ key for presence and the ‘down’ key for absence. The remaining participants made discrimination decisions between up and down motions, and used the right and left keys for reporting their detection decisions. These participants reported target presence with the right key, and target absence with the left key. Pooling right and up in the response bias analysis seemed natural because both directions were mapped to ‘present’ responses. We note that this analysis decision does not affect the reverse correlation analysis, which is contingent on the true direction of motion or on subjects’ decisions, and not on motion direction in absolute coordinates.

Figure 3 and nearby: Nowhere around here do you clearly describe how your response-conditional (I would have said "response-contingent") type 2 ROCs are computed. Later on somewhere (I forget where) you do say you split continuous confidence judgments around their median to get a binary value, thus explaining the notation "P(conf|(in)correct)". But, you don't say that here, and you don't say what "criterion" you are varying to map out this AUROC2. You just assume the reader knows what a type 2 ROC is and how it's computed for your experimental paradigm. This needs to be all laid out explicitly.

Following the Reviewer’s comment, and since this analysis is not at the center of this project, we moved the response-conditional type-2 ROC analysis to the Appendix, where we expand on the method in more detail:

“Following Meuwese et al. (2014), we extracted response-conditional type-2 ROC (rc-ROC) curves for the two tasks. For different values of x∈[0,1], we plotted p(confidence≥x) for correct against incorrect responses, separately for the two response options. Unlike traditional type-I ROC curves that provide a summary of subjects’ ability to distinguish between two external world states, type 2 ROC curves represent their ability to track the accuracy of one’s own responses. The area under the response-conditional ROC curve (auROC2) is a measure of metacognitive sensitivity, with higher values corresponding to more accurate metacognitive monitoring.”

15/25: "temporal dynamics of decision formation" seems an odd phrase for what you actually do. You correlate responses with momentary motion energies. You don't really concentrate on their dynamics, but only analyze (with details I still don't understand) whether the correlation is significant over the first 300 ms. And "decision formation" sounds like an interest in evidence accumulation (as in the DDM), but that isn't done here at all.

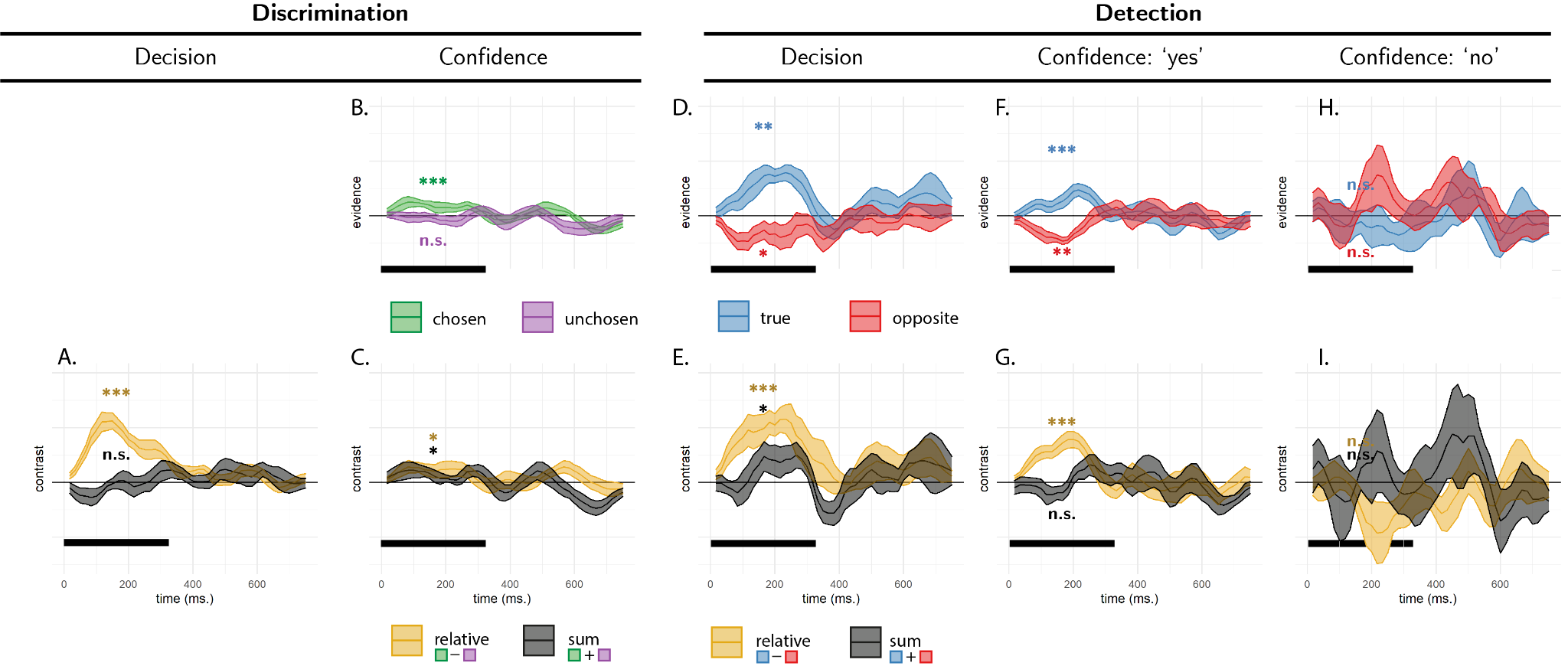
As the Reviewer rightly points out, our analysis in this paper mostly focused on comparisons between conditions rather than between timepoints. Nevertheless, the reverse correlation analysis reveals the temporal dynamics of evidence accumulation in that it allows us to see, for example, that subjects aggregate evidence during the first 300 ms of the trial, but not much after.

15/30: I have no idea what data went into this (and subsequent) statistical test and it's never clearly described. On any given trial you have a vector of motion energies and a binary response (or binarized confidence response?). How are these pooled over trials? Over participants? How does this result in 9 degrees of freedom???

The reverse correlation analysis consists of applying contrasts to motion energy vectors (or luminance vectors in Experiments 2 and 3) at the single subject level. The resulting subject-level contrast vectors are then used in a group-level contrast, and compared against 0 with a t-test. For this reason, statistical analyses for Experiment 1 (which had 10 participants) had 9 degrees of freedom.

Figure 4 confused me in many ways. First, I'm still not clear on what kind of mean-subtraction was done (pooled how over what) leading to the symmetrry in panel A upper-left. What exactly is plotted in the upper-right panels of A (how does it depend on the confidence response)? Why are you not remarking on the clearly significant effects here after 300 ms? For detection (panel B, upper left and perhaps others), isn't it a given that motion energy should be higher for "true" compared to "opposite", since that's the definition of the stimulus direction of a target being "true"? I don't understand how that asymmetry can mostly go away on hit trials. The legend for part B should say Lower left, not Upper left for the difference in motion energies.

We have now re-plotted the said Figure, including all other reverse-correlation figures, in a way that we hope will improve:



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.

We explain the reverse correlation analysis in more detail in the Computational Modelling section (we apologize for the corrupted mathematical notation):

“From each trial (tr) we extracted random fluctuations in perceptual evidence in the signal E\_s^tr (t) and non-signal E\_n^tr (t) sensory channels. To make sure we are measuring true random fluctuations and not systematic differences between noise and signal channel, we mean centered the signal channels to 0. For simplicity, in analyzing model simulations we averaged all timepoints in a trial to obtain trial-level noise estimates E\_s^tr and non-signal E\_n^tr. Human data was analyzed in a similar fashion, but separately for each timepoint. ‘Relative evidence’ was defined as the difference in noise terms between the signal and non-signal channels (E\_relative^tr=E\_s^tr-E\_n^tr). To obtain a decision kernel, we took the difference between the average relative evidence in trials where agents chose the signal and non-signal channels E\_relative=⟨E\_relative^tr ⟩\_CORRECT-⟨E\_relative^tr ⟩\_INCORRECT. This was done separately for each simulated agent, and the resulting values were tested against zero in a t-test. In all four models, relative evidence was higher on trials in which the agent correctly identified the signal channel (Fig. 3A, orange markers).”

Although we appreciate the figure gives an impression of significant effects (we thought so too!), results after the first 300 ms were not significant when correcting for multiple comparisons. We focused on the first 300 ms in light of previous findings by Zylberberg and colleagues.

Since we mean-centered all motion energy vectors as a function of motion direction, there is no longer a difference between the average motion energy vector in the true and opposite directions of motion. The results reflect the effects of random fluctuations on top of these systematic differences.

18/39-43: I'm not clear on how this was computed. Did you average separately over yes and over no trials and then subtract?

Yes, exactly.

19/4-8: Isn't it true that motion energy that strengthened the true direction of motion was already bigger to begin with, so how can you talk about differential sensitivity when the energy content was (by definition) not equated?

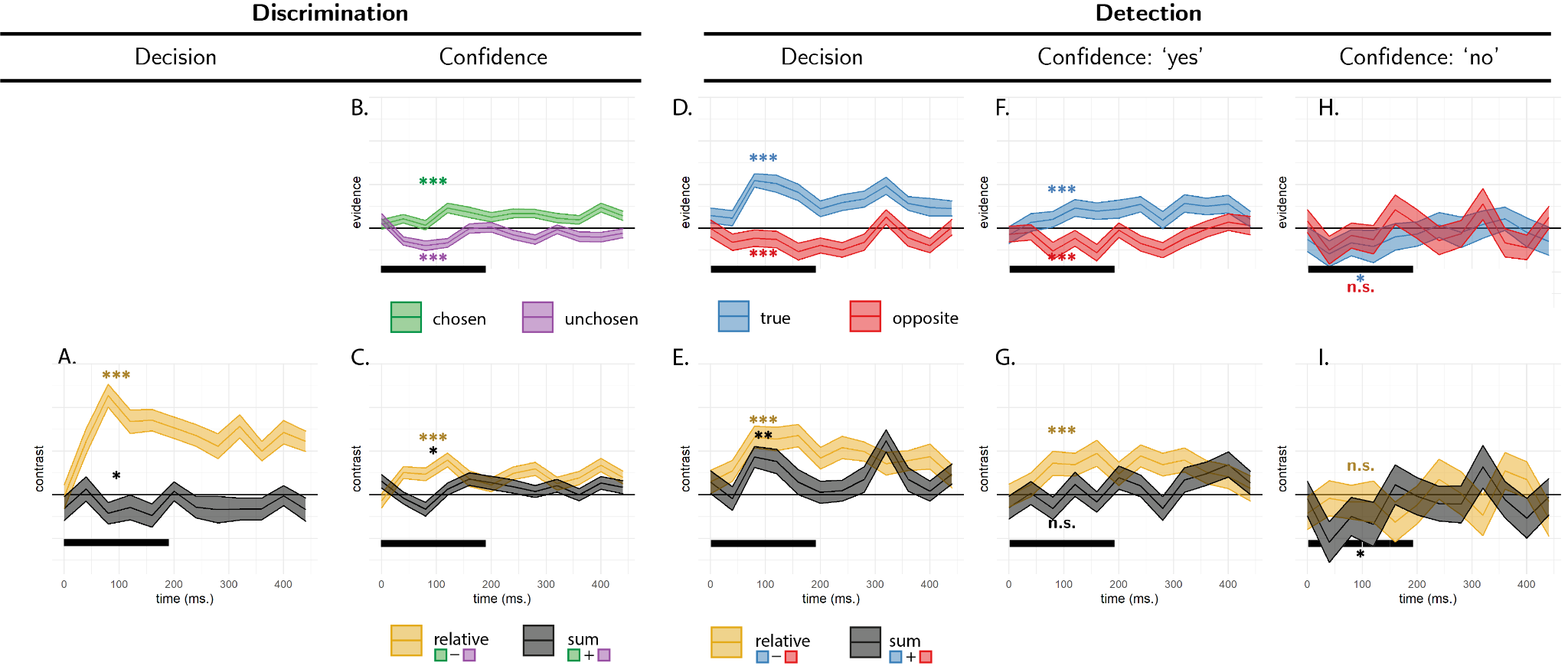
Motion energy vectors in the true and opposite directions both had a mean energy value of 0. The reason for this is that we first mean-centered all vectors as a function of motion direction and coherence.

21/40: The word "intensity" here is odd. You kind of mean d', but presumably don't want to call it "discriminability" (between the two stimulus categories) when you are already contrasting discrimination vs. detection. It might be worth saying explicitly here that you expect a difference of sqrt(2); well, not exactly, since you have stimulus-direction uncertainty in the detection case.

What we meant by stimulus intensity here was coherence rather than d’. The staircasing procedure resulted in lower coherence levels in discrimination, in order to equate accuracy between the two tasks and due to the fact that, for the same coherence levels, detection accuracy is lower.

26/22: Given that I was generally a little unclear reading this paper, here (and in several other places) it would have been kind to point out where in the figure I was supposed to see that decisions were, in this case, more sensitive in the non-signal compared to the signal.

We thank the Reviewer for this comment, which led us to re-plot some of our reverse correlation results. Specifically, panel C shows the effect of sum luminance in black, and corresponds to a stronger effect of positive evidence on decision confidence:



The appendices on zROC curves don't seem to be cited or discussed anywhere in the main text.

The zROC analysis was pre-registered and is therefore included in the appendix. But in order to simplify an already complex paper, we do not mention it in the main text.

# Reviewer 2

Mazor et al., investigated how sensory evidence informs discrimination and detection tasks both in terms of choice behavior and confidence. Using a combination of in-lab and online experiments with pre-registered hypothesis, the authors find support for the "positive evidence bias" - the finding, well documented in the literature, that participants overweight evidence in favor of their choice when reporting confidence in a discrimination task. A novel finding, which received mixed support across the three experiments, is that confidence in a detection decision may also exhibit a positive evidence bias and that the detection decision itself exhibits the positive evidence bias. An account of the positive evidence bias is given based on a Bayesian ideal observer, although this is found to not fully account for existing data.

The paper is well-written, easy to read, and the data are very throughout analyzed and informative with respect to an important issue in the field regarding how confidence in different types of decisions is computed. I have one relatively major concern regarding the two online experiments (experiments 2 and 3) and a few minor points.

We thank the Reviewer for this positive assessment of our work!

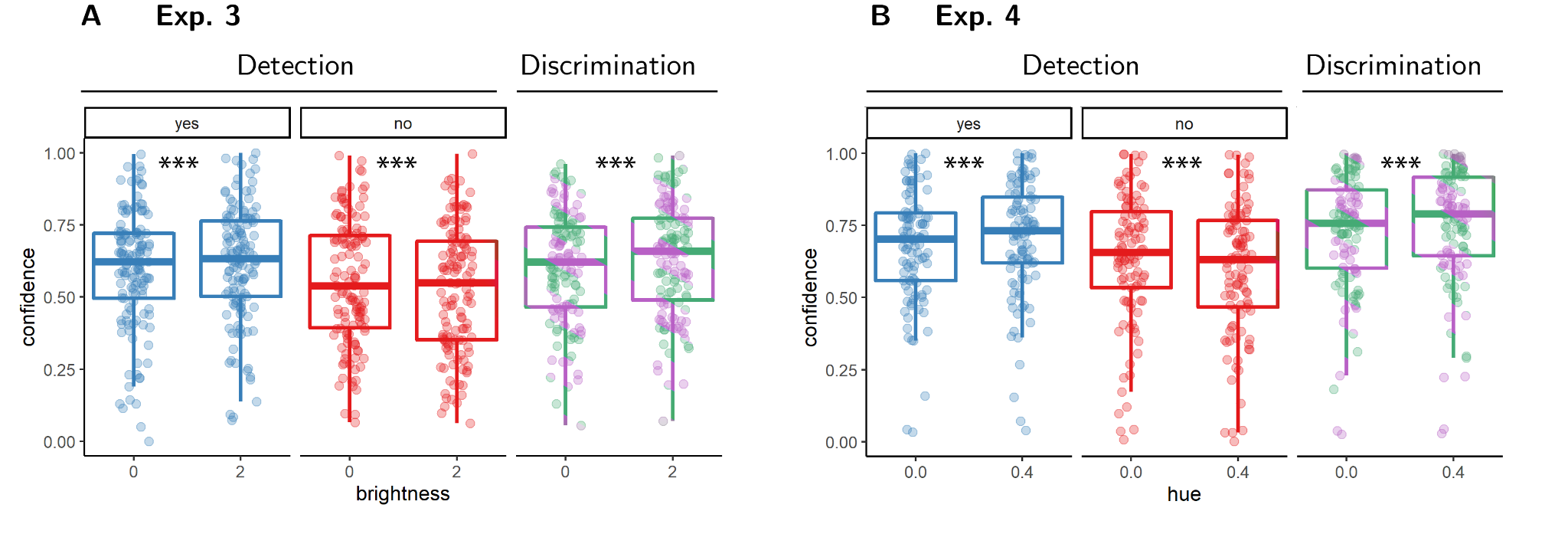
The online experiments required subjects to make luminance judgments based on which of two sets of bars had higher mean luminance (discrimination) or whether either set of bars had higher mean luminace than the background (detection). In a lab context, the monitor used to present the stimuli would be gamma-corrected, which accounts for the fact that actual screen luminance (in cd/m^2, for instance) is not a linear function of RGB values. I didn’t see any mention of this in the methods, and frankly, am not sure how it would be possible in an online setting since it typically requires measuring the luminance profile of each monitor separately to determine that monitor’s unique gamma function. So assuming some kind of correction was not done for the online experiments, the problem this presents is that the actual brightness of the screen is not linearly related to the RGB values used for the reverse correlation analysis (and used by the subject for their deicsions). In other words, a dark bar of say -15 RGB values from the mean does not cancel out a white bar that is +15 RGB values form the mean, so the mean lumiannce across the bars is actually brighter than the background. Assuming a typical gamma shape for an uncorrected monitor (e.g., <https://floyd.lbl.gov/radiance/refer/Notes/gamma.html>), increases in RGB values relative to the middle RGB value of 128 usually lead to larger luminance changes than decreases from 128. This likely means that in the detection task, which sampled RGB valies from a normal distribution centered on the background RGB value, there is actually always a (possibly small) net increase in luminance relative to background (since the dark bars are not “as dark” as the ligth bars are light). This could also potentially complicate the reverse correlation analysis looking at light and dark fluctuations separately since they are not symmetric. If online data collection is not terribly cumbersome, I’d recommend an additional experiment using a stimulus feature that is not so monitor-dependent and non-linear.

We thank the Reviewer for this important point. Indeed, a nonlinear mapping from RGB values to the perception of luminance can explain some results from Exp. 2 and 3. We now include a fourth experiment to address this concern, and replicate our main findings when controlling for such nonlinearities:

“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 address 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.”

The results were a full replication of Exp. 3:



*“Figure* *9.*  Difference in confidence between standard and higher-evidence trials for the three response categories (detection ‘yes’ and ‘no’ responses, and discrimination responses) in Exp. 3 and 4. 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”

The authors may be interested in a related line of work that looks at trial-to-trial variation in detection reports as a function of pre-stimulus brain states. Although positive evidence is not explicitly manipulated, a similar model has been applied to explain detection and visibility judgments. Namely when the brain is in an excitable state, a stimulus appears more visible but is not actually better detected against noise. We recently reviewed these findings and propose a similar model to the PEB (Samaha, Iemi, Haegens, Busch, 2020, TiCS).

Thanks for this highly useful reference. We have now incorporated a paragraph about this idea to the Discussion:

“Alternatively, changes to the global perception of overall stimulus intensity may have had an internal source. For example, slow brain oscillations in the alpha band affect both detection criterion and discrimination confidence, but hardly affect discrimination sensitivity. This nonselective effect on perception has been attributed to a global change in the baseline firing rate of sensory neurons (Samaha et al., 2020). Similar to our evidence-boost manipulation, an overall increase in baseline firing rate increases sum evidence without affecting relative evidence. If agents do not have metacognitive access to the current excitability of their perceptual system but do know that such global effects exist, focusing on relative evidence in detection may be rational as any global effect would cancel out in subtracting negative from positive evidence.”

Could color legends be added to the figures?

We have now added color legends where they were previously missing.

# Reviewer 3

This project investigates the positive evidence bias in decision making and confidence using perceptual decision tasks, including a motion detection task and a brightness detection task. Experiments tested both discrimination (e.g., systematic motion to the left or right?) and detection (e.g., any systematic motion or random motion?), and the inclusion of the detection tasks was the primary advance over previous studies. Results were a bit mixed, but positive evidence bias effects showed up in one way or another in both discrimination and detection tasks.

The experimental and analysis methods were impressively rigorous, and I suspect that decision researchers will be interested in the results. I commend the researchers for using preregistration, and I was happy to see one of the theoretical accounts implemented in a simulation.

We thank the Reviewer for this positive assessment of our work.

However, the current version of the manuscript doesn’t make a clear case as to the theoretical significance of the findings. The Introduction covers a couple of theoretical accounts, and then the experiments go through a complex web of findings without any discussion linking back to these accounts. The General Discussion loops back to the theoretical positions, but they are still only loosely related to the empirical results. For example, there was no clear analysis of whether the alternative accounts make different predictions for what should happen in detection, or how detection results should relate to discrimination results. In other words, the paper didn’t make a convincing case that the novel empirical contribution of the current experiments (the detection tasks) helps us better understand the positive evidence bias at a theoretical level. I also wondered about a potential alternative explanation based on selection effects, as explained below.

We thank the Reviewer for this set of comments, which we have taken seriously. In this new, revised version of the paper, we identify one key finding, which is consistent across all four experiments and is not predicted by four different Bayes-rational models (a negative weighting of evidence in a detection task). We hope the novel empirical contribution of this work is now more immediately apparent.

“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.”

The paper also did not provide a clear picture of the relationship between the different analysis strategies, and it seemed to me that “positive evidence bias” meant different things in different places. These are issues that could be potentially remedied with writing, provided that there is a close link between theory and data that didn’t quite come across in the current version.

I do not do research on the positive evidence bias, and by the end of the paper I felt like I didn’t really know what it means. My initial understanding was that positive evidence was evidence in support of a decision and negative evidence was evidence against a decision. This mostly holds up throughout the paper, although sometimes this is defined based on a true stimulus category (“To test for a potential asymmetry in evidence weighting in discrimination decisions, we contrasted the contribution of motion energy in the true and opposite directions of motion… defined with respect to the stimulus”) and sometimes it is defined based on the participant response (all the chosen vs. unchosen analyses). Moreover, the effect was defined both for actual decisions and for confidence. It would have been helpful if the Introduction had delineated all of these possibilities and explained the theoretical significance of these different ways to define the effect. For example, do the accounts under investigation make the same predictions regardless of how a positive evidence bias is assessed?

We thank the Reviewer for this useful comment. For consistency between the different analyses, we now clarify the relation between the different effects in the Introduction, focusing on the effect of sum evidence in all ‘positive evidence’ effects:

“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).”

Additionally, we include a sub-section about the relation between confidence and decision ‘positive evidence’ effects:

**“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.”

Finally, we now made sure to restrict the use of the terms positive and negative evidence to describing evidence in support or against a decision.

Another thing I found confusing was that the paper switched between different ways of defining the effect without a clear rationale. For example, Experiment 1 seems to focus on the response-defined effect for confidence results versus the stimulus-defined effect for actual decisions. I understand why the response-defined analysis doesn’t make sense for the latter, but why not focus on the stimulus-defined effect for confidence? This analysis could be conditioned on response (as seems to happen in Experiment 2) so that accuracy differences don’t mediate the confidence effect. I found myself wondering why that was less important to evaluate and wondering how the implemented set of analyses was related to the broader goals of the project. The results sections came off as just a list of things that happened. Again, this isn’t my primary field, which might play a role in this, but the authors shouldn’t assume that readers already know why they are doing what they are doing.

We thank the Reviewer for this set of comments. The positive evidence bias effect in discrimination is most easily understood as differential effect of decision-congruent and decision-incongruent evidence on decision confidence. However, since a positive evidence bias equals an effect of sum evidence, it stays the same regardless of how evidence channels are partitioned (e.g., chosen/unchosen or true/opposite).

The discrimination confidence evidence bias effects remain the same in our data when restricting analysis to correct responses only (such that decision- and response- contingent effects collide): Exp. 1: t(9) = 4.45, p = .002, Exp. 2: t(100) = 3.56, p = .001, Exp. 3: t(99) = 0.23, p = .816; Exp. 4: t(108) = 2.32, p = .022.

Positive evidence bias seems to mean subtly different things in other places, such as here: “when deciding whether one of the flickering patches was brighter, participants were more sensitive to positive noise in the brighter patch than to negative noise in the darker patch.” This sounds to me like evidence that supports the decision but happens to be coming from the distractor stimulus (i.e., the non-target looked particularly dark helping them pick the target as the bright one.) Am I reading that right, and if so, is that also considered “negative evidence?” I was also unsure how the following comment related to the idea of evidence that went against the ultimate decision that was made: “For discrimination judgments, participants were also more confident in higher- compared to lower luminance trials (𝑀 = 0.02, 95% CI [0.01, 0.03], 𝑡(99) = 3.20, 𝑝 = .002), replicating a positive evidence bias for discrimination confidence.” Maybe the logic here is that increased overall brightness increases both positive and negative evidence, so higher confidence means that the increased positive evidence won out? Do all the theories really predict the same thing for this sort of comparison as comparing chosen versus non-chosen options?

We hope these subtle points are now better explained in the new version of the Introduction. We now make sure to explain that negative evidence is evidence in support of an alternative to a decision:

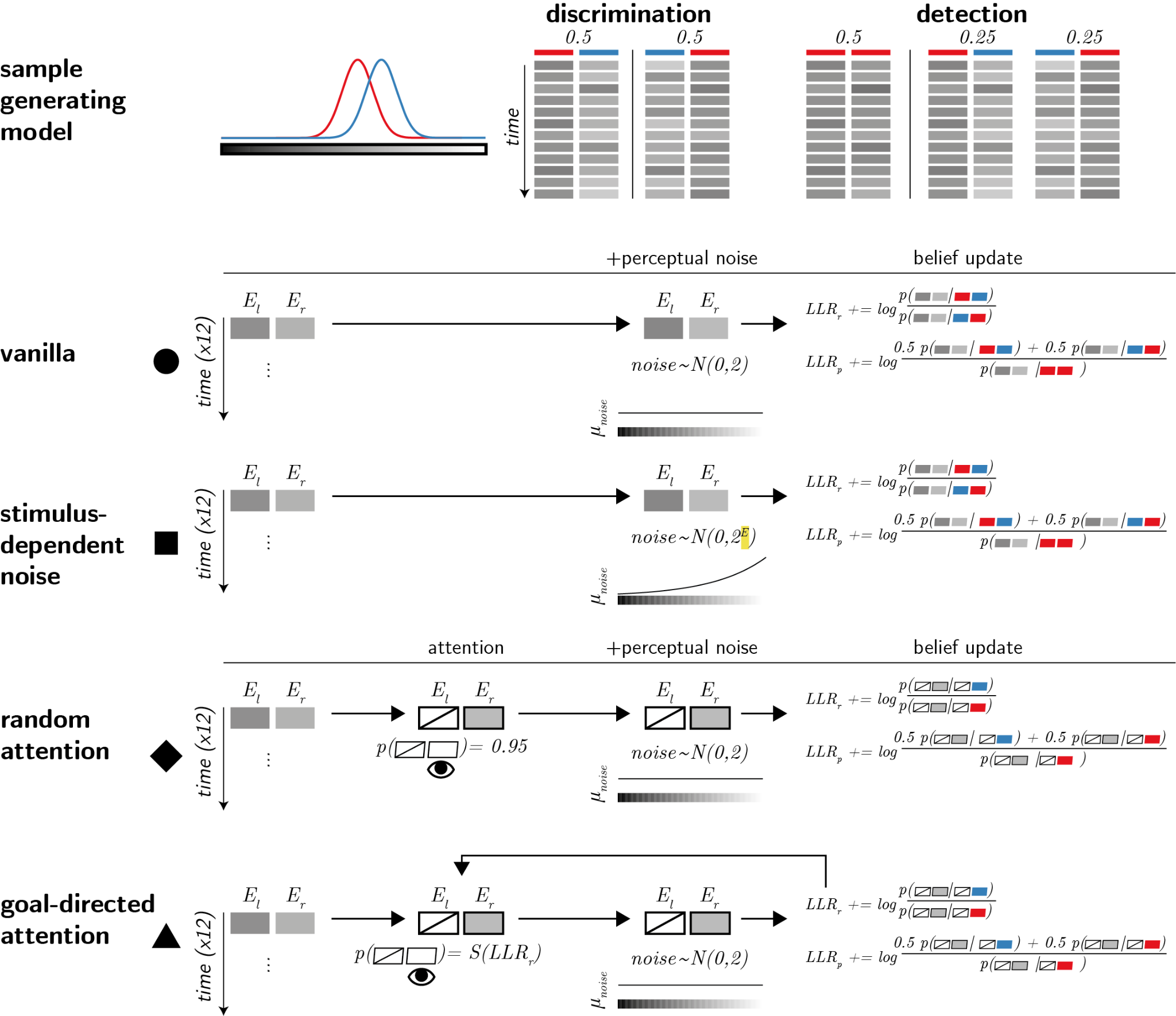
“For example, when deciding that there are more ants in the kitchen than in the living room, confidence should not only positively weigh the number of ants found in the kitchen (positive evidence), but also negatively weigh the number of ants found in the living room (negative evidence).“

We also comment about the equivalence between an effect of sum evidence and a positive evidence bias:

“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.“

I wondered whether there was a potential alternative account based on selection effects. I’m not sure this covers all the different ways of defining a positive evidence bias, but it seems like it should be applicable for the response-defined analyses, at least. Throughout the paper, evidence is defined in terms of the properties of the physical stimulus, but the human decision makers aren’t taking in a perfect representation of the stimulus, of course. In other words, we can imagine some weighting mechanism (call it “attention,” if you like) the determines the extent to which \*potential\* evidence from the stimulus is actually registered in the decision making process. If you just have fluctuation in weighting for different evidence sources and you base confidence on whatever evidence you considered, would you get a positive evidence bias? This could be explored in a simulation. For example, if you pull out trials for which the participant responded “moving to the right,” then you will also select trials that happened to have high weights on the potential sources of evidence (individual dots, say) that provided support for a “right” response. Confidence responses would also be more influenced by potential evidence sources that had a higher weight, so confidence will tend to be more influenced by potential evidence sources that are consistent with the response. To what extent could these sorts of selection mechanisms influence the results? Are there aspects of the results that rule out this sort of mechanism?

As the Reviewer rightly points out, selection effects can give rise to a positive evidence bias in discrimination confidence. This has been documented by Sepulveda and colleagues (2020), in the context of value-based and visual decision-making. We now include a simulation of a similar model, and show that it produces the expected asymmetry in evidence weighting (the goal-directed attention model below). This model fails to explain the negative weighting of evidence in detection decisions and confidence ratings.



“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.”

Again, I liked the simulation exploring the effect of unequal variance. Could this explain effects in a discrimination task, though?

The stimulus-dependent noise (unequal variance) model perfectly accounts for the discrimination data, but fails to account for the negative weighting of evidence in detection decisions:

“The same two models [stimulus-dependent noise and goal-directed attention] also made predictions for the detection task: when inferring presence, decisions and confidence ratings should positively weigh evidence for both alternatives (e.g., motion energy to the right and to the left). Paradoxically, however, in the detection task subjects adopted a discrimination-like disposition, and negatively weighted evidence in the non-signal channel. In Exp. 1 and 2, this negative weighting of evidence in the non-signal channel was strong enough to bring to the total effect of sum evidence on detection confidence down to zero. Although experiments 3 and 4 did show an effect of sum evidence on detection confidence, the negative weighting of evidence in the non-signal channel remained. Overall, subjects incorporated detection-relevant evidence into their confidence in discrimination judgments, and discrimination-relevant evidence to their detection judgments and confidence ratings”.

The paper gives terse descriptions of some pretty sophisticated analyses. Based on that, I’d recommend posting the data and analysis code online, if it isn’t already. (Kudos again on the preregistration.)

We agree. A fully reproducible data-to-paper code is available on the project’s GitHub: <https://github.com/matanmazor/reverseCorrelation/blob/main/docs/reverseCorrelationPaper.Rmd>. Raw data is available on GitHub and on the project’s OSF page: <https://osf.io/7a4fm/>

# Reviewer 4

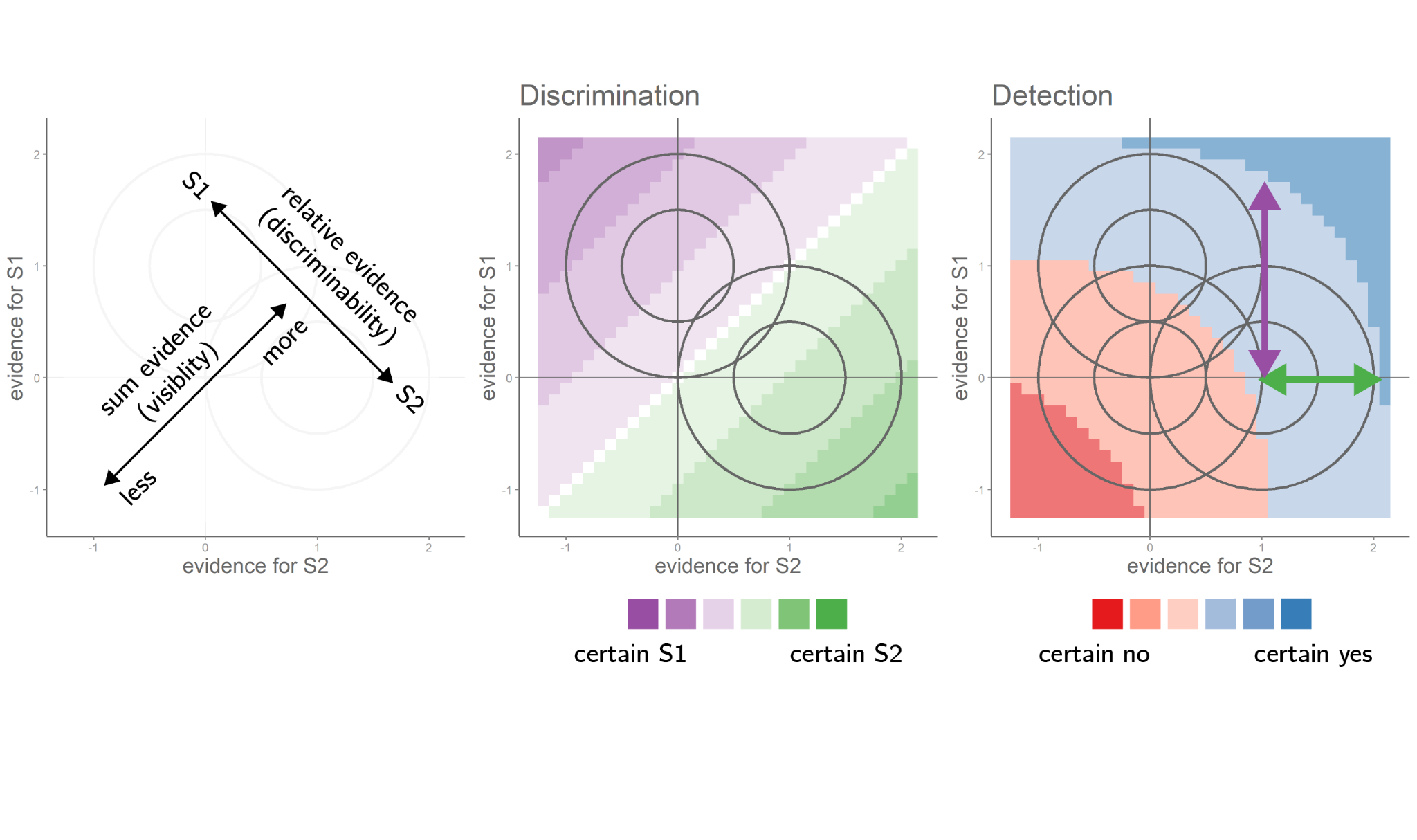
The positive-evidence bias (PEB) for confidence has been observed for many binary discrimination decisions. The PEB refers to the observation that confidence is more strongly influenced by the evidence in favor of the chosen option than by the evidence for the unchosen option. The present paper investigates if the PEB extends to decision about the presence or absence of a stimulus (i.e., ’detection’ judgments). The main result is that detection decisions also show a PEB, supporting recent proposals that a PEB is due to participants adopting a detection-like strategy even in discrimination judgements.

I found the paper timely, interesting and clearly written. My congratulations to the authors. Please see my comments below.

We thank the Reviewer for this positive assessment our work.

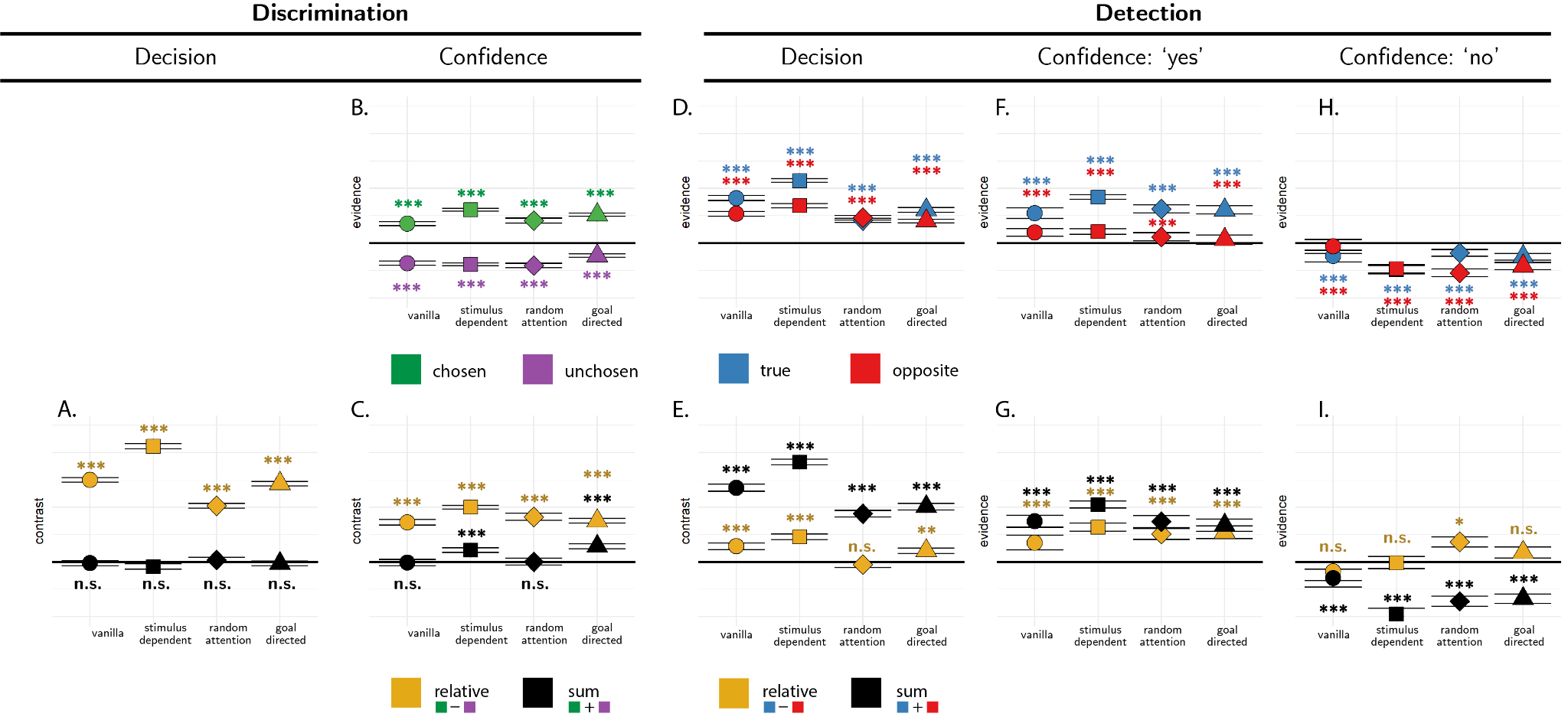
1. One page 5, it is written that “…the likelihood ratio between stimulus presence and absence is more sensitive to evidence for the detected stimulus (positive evidence) compared to evidence for the absence of other, undetected stimuli (negative evidence; see Fig. 1, right panel).”. Can the authors provide some intuition for this statement? Fig. 1 is references, but I do not know what to look for in this figure in support for the statement. Could the authors clarify, or annotate the figure somehow?

Our intention was to draw readers’ attention to the fact that the LLR gradient is steeper in the direction of the true stimulus, compared to the direction of the other stimulus. This can be seen by measuring the distance from the center of the evidence distribution to the dark blue area in on the x and on the y axis (see green and purple arrows):



We have now removed this sentence to avoid confusion. Instead, applying reverse correlation to our model simulations provides evidence for increased sensitivity to evidence for the true stimulus in detection decision and confidence, resulting in a net effect of relative evidence (the difference between the two quantities):

“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).”



*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.

2. On page 15, it is indicated that ‘… we constructed the contribution of motion energy in the true and opposite directions of motion.. independently of decision’. I do not understand what is being done here. If the kernels are independent of the decision, how are the kernels informative about the decision? It might help to include the equations for the computation of the kernels, and the de-meaning procedure.

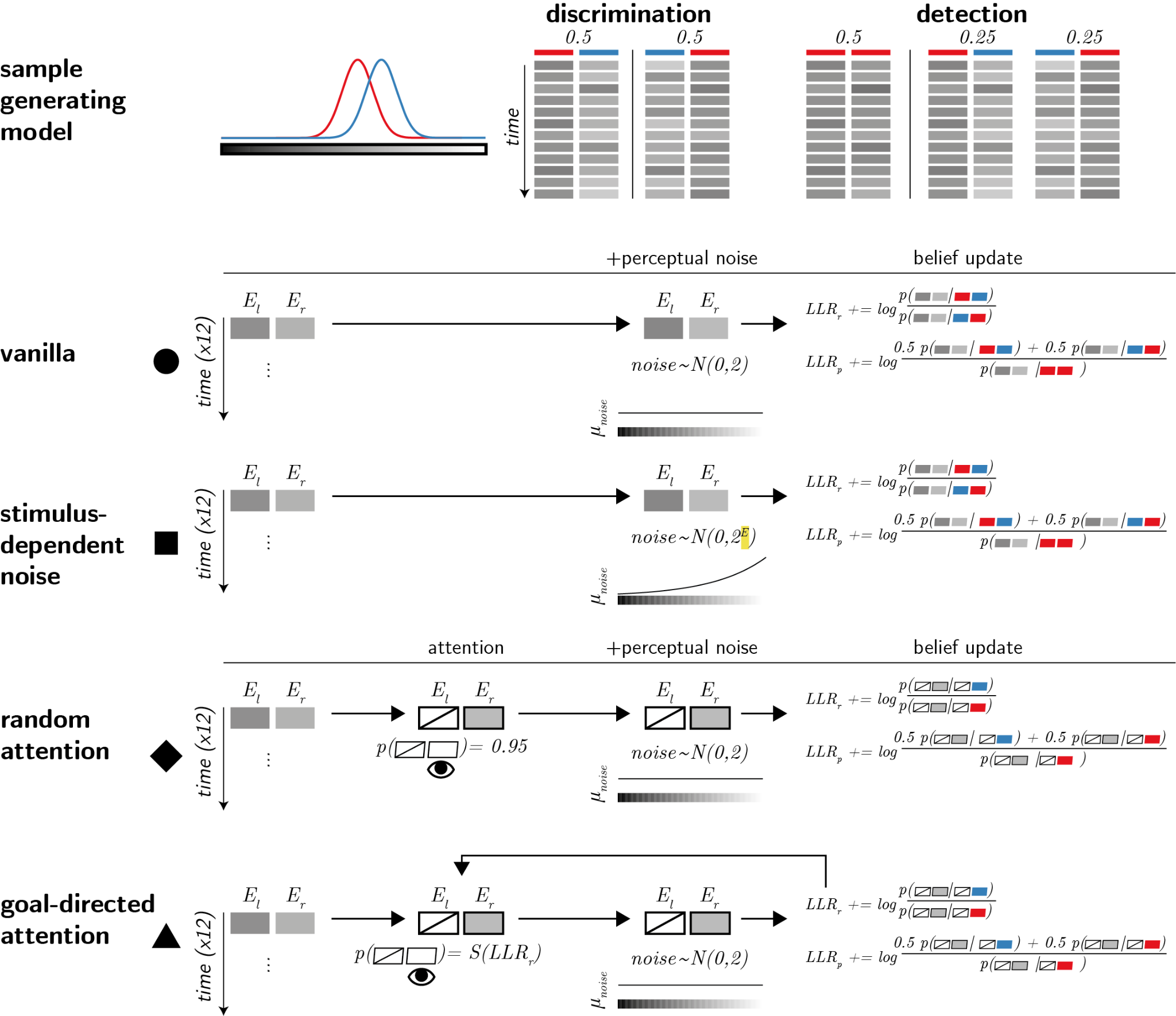
We now include the equations for the kernels in the “Compuatational Models” section (pasted here, but see paper for rendered equations):

“From each trial (tr) we extracted random fluctuations in perceptual evidence in the signal E\_s^tr (t) and non-signal E\_n^tr (t) sensory channels. To make sure we are measuring true random fluctuations and not systematic differences between noise and signal channel, we mean centered the signal channels to 0. For simplicity, in analyzing model simulations we averaged all timepoints in a trial to obtain trial-level noise estimates E\_s^tr and non-signal E\_n^tr. Human data was analyzed in a similar fashion, but separately for each timepoint. ‘Relative evidence’ was defined as the difference in noise terms between the signal and non-signal channels (E\_relative^tr=E\_s^tr-E\_n^tr). To obtain a decision kernel, we took the difference between the average relative evidence in trials where agents chose the signal and non-signal channels E\_relative=⟨E\_relative^tr ⟩\_CORRECT-⟨E\_relative^tr ⟩\_INCORRECT. This was done separately for each simulated agent, and the resulting values were tested against zero in a t-test. In all four models, relative evidence was higher on trials in which the agent correctly identified the signal channel (Fig. 3A, orange markers).”

3. The authors show that the detection confidence kernels in experiments 1 & 2 do not show a positive-evidence bias. Can this effect be explained by the authors not knowing which of the two patches was perceived as the brightest? I am worried that the symmetry of the detection confidence kernels is explained by the experimenters not knowing which target was covertly detected. It would help to conduct some modeling to provide some intuition for what is expected. For example, the authors could simulate a simple signal-detection theory model, in which detection is determined by comparing the noisy evidence to a threshold, and confidence is a function of the distance to the criterion (or probability correct). Will the confidence on “yes” trials show a positive evidence bias if the experimenter does not know the side that was perceived as the brightest?

Since a positive evidence bias (pos > -neg) is mathematically equivalent to the effect of sum evidence (pos+neg>0) , sensitivity to this effect is fully independent of how we partition evidence into positive and evidence categories. As a result, not knowing which was the detected direction of motion does not affect our ability to detect a positive evidence bias. We have now simulated the behaviour of four models, similar in nature to the one proposed by the Reviewer, and show that all four models produce a strong effect of sum evidence on detection confidence:

“We model a setting in which agents are presented with a sequence of samples from two sensory channels: *E1* and *E2*. 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 *N(0.5,1)*, and when a signal is absent evidence is sampled from *N(0,1)* (see Fig. 2, upper panel). In all four models agents only have access to a noisy version of these samples *E'*, 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.

“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.”

“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).”

4. The simulations of the Bayesian-rational model with scaled noise are very interesting. The authors argue that if stronger stimuli are also noisier, then sum evidence should have a negative effect on confidence, which is not what is observed in the data. Have the authors explored models in which the noise is correlated for the two patches? In this case, even if noise increases with stimulus strength, participants would still be able to determine which of the two patches was brighter. I think that assuming some degree of noise correlation is very reasonable (e.g., on some trials bright stimuli may seem to be brighter than they really are).

Although theoretically stronger stimuli should produce a negative effect on task accuracy in a scaled noise model, we now include model simulations which show that this effect is relatively mild (see Fig. 3A, black markers above). Instead, we now focus on a much more pronounced miss of the predictions made by these models: they fail to account for a negative weighting of some evidence in detection confidence and decisions.

We agree that the effects of correlated noise are interesting and can potentially account for some aspects of the data, including the focus on relative evidence in detection even when sum evidence is more informative. We now refer to this possibility in the General Discussion:

“Alternatively, changes to the global perception of overall stimulus intensity may have had an internal source. For example, slow brain oscillations in the alpha band affect both detection criterion and discrimination confidence, but hardly affect discrimination sensitivity. This nonselective effect on perception has been attributed to a global change in the baseline firing rate of sensory neurons (Samaha et al., 2020). Similar to our evidence-boost manipulation, an overall increase in baseline firing rate increases sum evidence without affecting relative evidence. If agents do not have metacognitive access to the current excitability of their perceptual system but do know that such global effects exist, focusing on relative evidence in detection may be rational as any global effect would cancel out in subtracting negative from positive evidence.”

Please indicate the instruction given to the participants in the detection task. Specifically, would be useful to know if the participants knew that the target was present in only one of the patches, or if they may think that the target could be present in both at the same time.

We now clarify this important point in the text:

“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. Participants were told that only one of the two patches could be bright, but never both. Decisions in Detection trials were reported using the ‘Y’ and ‘N’ keys. Confidence ratings and feedback were as in the discrimination task.”