*Reviewer 1 (R1, Itiel Dror) liked the manuscript but pointed out that he is not an expert of the specific topic. R1 points to some own work on linear sequential unmasking and how it might be relevant to the present work. In particular, R1 suggests discussing how potential biases can be avoided by presenting information in specific ways. Furthermore, R1 suggests replacing some references with more recent work.*

The own work to which Professor Dror points describes findings that are highly relevant here: individuals appear most affected by “first impressions” created by the first evidence samples experienced. We gave formalised and tested this idea in our “primacy model”, which we now test against other computational models in Experiment 2. Our results, however, suggest that a heuristic delta model can best explain the results we report here.

*Reviewer 2 (R2) thinks the study is fascinating. Nevertheless, R2 believes it suffers from a major shortcoming, namely, that the ideal Bayesian observer model had not been correctly specified. In particular, R2 argues that the assumption of equal priors for the two potential states is unjustified. R2 argues that the priors should be either estimated as a free parameter or derived from the self-reports of the participants. R2 wonders how the results would look like when comparing participants' behavior with the suggested optimal Bayesian model assuming unequal priors. R2 makes some specific suggestions of how the Bayesian observer model could be estimated on an individual basis.*

The inclusion of a model with parameterised priors, we think, makes a considerable theoretical contribution to the paper. We now build, fit to participant data and formally compare four computational models that parameterise the prior in this way.

From our perspective, models that parameterise the prior should play a rather different role in this paper than we intended for the ground truth computation. The former are meant to theoretically explain participants’ ratings in terms of their prior belief, while the latter was meant as a contrast to, not an explanation of, participants’ behaviour.

We agree that a computational theory of participants’ behaviour affords us stronger and more impactful conclusions than a contrast against a ground truth computation. Our resubmission now places the emphasis in the paper and draws our main conclusions from our theoretical model comparison, de-emphasising comparisons to the ground truth computations.

*I also think that the topic of your paper is quite essential and exciting. However, I also think it has various shortcomings:*

*I believe your manuscript could benefit substantially by adopting a more cautious view of which behavior can and should be called a "bias." Just because people make judgments inconsistent with a Bayesian model, does not necessarily mean that this represents a bias because there might be various reasons and different adaptive ways of how people form judgments inconsistent with Bayesian thinking.*

It appears we may not have been clear about the role the ground truth calculations are meant to play in our paper. We never claim that participants form judgments in a way consistent with Bayesian thinking. Quite to the contrary. Indeed, our intention in including the contrast between participants’ behaviour and this Bayesian performance standard was always: to highlight the inconsistency.

To make our conclusions clearer and more direct, we have rebranded this Bayesian conditional probability calculation as “ground truth computation” and excised the word “bias” from the paper to emphasise further that this “computation” is based on “ground truth” facts about the paradigm (e.g., that male and female suspects are equally likely to be guilty) and is not designed to explain or model participants’ psychology or subjective beliefs. Although we include ground truth computation on our plots, we have refocussed the paper’s narrative to emphasise instead our theoretical model comparison, which we believe better enables strong conclusions about why participants differ from a Bayesian standard. Indeed, we propose a simple, prediction error-based heuristic that avoids the full Bayesian computation, which incorporates subjective prior beliefs about suspects, and which better fits participants’ behaviour than three alternative models based on the Bayesian conditional probability formula.

*You derived your predictions and hypothesis from past empirical findings. This approach has the drawback that the potential evidence for the predictions does not provide evidence for a theory of how people form beliefs on the basis of a sample of information. Maybe it would be possible to provide at least a verbal theory of the cognitive process underlying belief formation and when and why different amounts of information could be used.*

The new theoretical computational models and formal model comparison provide such a theory in the form of the delta model that we propose.

*Following R2, it will be necessary to test the Bayesian model much more thoroughly on an individual level. This test should imply an optimal model that takes different priors into account as suggested by R2. However, you could even adjust the Bayesian model to include additional mechanisms such as primacy or recency effects. With these additions, it would be possible to illustrate the cognitive components of how people deviate from a standard Bayesian model.*

Our resubmission is now rewritten to accommodate an extension theoretical model development and comparison for the data in Experiment 2. Our models quantify prior beliefs using free parameters for each suspect, as requested. Moreover, they explore primacy, recency, split (proportion guilt claims) and delta models as explanation of sensitivity to disconfirmatory evidence. Our finding that the delta model fits best and accurately reproduces participants’ behaviour (including suspect and sensitivity to disconfirmatory evidence effects) indeed “illustrates components of how people deviate from a standard Bayesian model”.

*In Experiment 1, you also manipulated the religious beliefs of the suspect. However, no motivation for this manipulation is given in the introduction.*

We now cite research that people have widespread prior beliefs that atheists are prone immoral behaviour.

*In all figures, I recommend using an identical scale for the y-axis for better comparisons.*

We have made this change.

*I wonder whether computing the average adjustments when receiving innocent or guilty claims presented in Figure 1 is useful. The change in the posterior probability of being guilty depends on the differences between the guilty versus innocent claims received. The adjustment should be larger when having little evidence and smaller when having already collected substantial evidence. This difference will be lost when calculating averages.*

Unfortunately, it was not clear to us what plots the editor is requesting. Figures 2 and 4 already plot adjustments by guilt and innocent claims. Moreover, we report linear mixed models where Preceding Context is treated as a continuous variable and report convergent results to the ANOVAs, which are based on the averages. If the editor has something different in mind and it is relevant for directly illustrating our main conclusions, we are of course happy to oblige.

*In the discussion of Experiment 1, you point out a bias regarding disconfirming evidence. However, if I understood you correctly, you classify evidence as being confirming or disconfirming relative to the preceding context. However, according to the optimal model, the preceding context should be ignored, so according to the optimal model, the evidence is neither confirming nor disconfirming, so calling the different responses to the preceding context a "bias" appears to be misleading.*

“Confirming” means that the evidence sample is consistent with the preceding evidence, which is not dependent on the ground truth computation. To avoid any semantic confusion, we no longer refer to this as a bias in the manuscript.

*Experiment 2 provides strong evidence that the assumption of equal priors for the optimal model is not justified.*

Experiment 2 indeed provides strong evidence that participants are not using the ground truth computation as their strategy. This is and was also our conclusion. Perhaps the manuscript might have been clearer that the contrast of participants with ground truth model was indeed meant to highlight this very fact.

We hope that the inclusion of our new computational models provides a more direct and intuitive theory of participants’ behaviour in terms of a delta heuristic with unequal priors.

We also hope that our rebranding of the “ground truth computation” better emphasises that this computation is not intended as a “model” of participant’s behaviour and is instead meant to be an objective conditional probability computation based only on the ground truth facts of the paradigm, in which the prior probability of an atheist/Christian or male/female suspect’s guilt was in fact 50% and not in any way dependent on participants’ prior beliefs.

Our new model comparison allows us to theoretically explain the discrepancy between the ground truth computation and participants’ actual behaviour in terms of participants’ use of the delta heuristic with unequal priors. We hope this provides a clearer demonstration of this “strong evidence”.

Reviewer #1:

*I very much enjoyed reading this manuscript, but have to say that I am not familiar with the beads task, and it is not within my area of expertise (which focuses on forensic science decision making and bias).*

*As somewhat of an 'outsider' I found the manuscript very interesting and important, and think it should be published in JEP:LM&C.*

*I have just two comments, relating to my expertise and perspective:*

We appreciate Professor Dror’s positive comments. Our motivation for this study was (in part) to introduce the beads task to forensic application and, in this sense, an expert on forensic decision making we believe can attest to the value of this introduction.

*First, as the author correctly point out, "the forensic decision maker must update beliefs upon the receipt of new evidence". I think this is very critical, given the vast experimental psychology literature about order effects (e.g., when the same information is presented in different order, people reach different conclusions, due to how different pieces of information create expectation, and then confirmation bias kicks in, etc., etc.).*

*In this regard, I'd be interested to know what the authors think of Linear Sequential Unmasking' (LSU), where biases are minimized by controlling the order of information. Rather than the order being random or accidental, it is controlled to minimize bias and noise, etc. See details at: ​Dror, I. E. & Kukucka, J. (2021). Linear Sequential Unmasking-Expanded (LSU-E): A general approach for improving decision making as well as minimizing noise and bias. Forensic Science International: Synergy, 3, 100161. DOI: 10.1016/j.fsisyn.2021.100161*

*It is open access, so it is available at:* [*https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.sciencedirect.com%2Fscience%2Farticle%2Fpii%2FS2589871X21000310%3Fvia%253Dihub&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=hAWysg09neEcVddvqBKBmhROAOhtE6PA2vATzUsmF10%3D&reserved=0*](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.sciencedirect.com%2Fscience%2Farticle%2Fpii%2FS2589871X21000310%3Fvia%253Dihub&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=hAWysg09neEcVddvqBKBmhROAOhtE6PA2vATzUsmF10%3D&reserved=0)

*I wonder if the authors want to discuss such approaches, as ways to move forward to deal with bias. It will definitely add a nice (& optimistic) perspective, or, at least something to think about: Beyond establishing bias, what can psychologists suggest to do about such biases?*

This work indeed appears relevant. We found especially interesting the finding that individuals are more strongly impacted by information towards the beginning of sequences.

The new manuscript builds new computational models of participant evidence accumulation, fits them to participant behaviour, and then compares them to identify the best theory. We were inspired to develop the primacy model on the basis of the reviewer’s recommended work. We also raise the issue in our revised General Discussion section.

*My second point is minor, just to update some of the forensic science literature. Please replace:*

*Dror, I.E. (2017). Human expert performance in forensic decision making: seven different sources of bias. Australian Journal of Forensic Sciences, 49, 541-547.* [*https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1080%2F00450618.2017.1281348&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=RVRqNJh4zcbGwEla8EYC5xDKQH1%2FVqcqqCIp64dqUIY%3D&reserved=0*](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1080%2F00450618.2017.1281348&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=RVRqNJh4zcbGwEla8EYC5xDKQH1%2FVqcqqCIp64dqUIY%3D&reserved=0)

*With this more updated reference:*

*Dror, I. E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. Analytical Chemistry, 92 (12), 7998-8004. DOI: 10.1021/acs.analchem.0c00704*

*(it is open access, so it is available at:* [*https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubs.acs.org%2Fdoi%2F10.1021%2Facs.analchem.0c00704&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=uFogls3OSRUlspeBMcJl7U48utW%2FXzOBT9b8BDHDBo0%3D&reserved=0*](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubs.acs.org%2Fdoi%2F10.1021%2Facs.analchem.0c00704&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=uFogls3OSRUlspeBMcJl7U48utW%2FXzOBT9b8BDHDBo0%3D&reserved=0)

*--I think the authors will appreciate this newer article and may find some of the new points there interesting and relevant to their work.*

Many thanks for this reference, which we have used accordingly.

*I would very much encourage the authors to continue in their important and interesting research.*

*Itiel Dror*

*Reviewer #2: Thank you for the opportunity to review 'Humans Form Biased Beliefs from Samples of Evidence in Forensic Scenarios'.*

*It is a most fascinating study, but it contains, in my view, a serious error which runs through the study and it needs to be corrected throughout, as it impacts on key results. I have other concerns too, but this is the most serious:*

*The authors base many of their results on comparisions with an ideal Bayesian observer, but they use the equation in p. 9 of the manuscript. This appears to be the equation originally written by Moutoussis et al 2011. However, upon consulting this reference, I note that the authors derived this formula (appropriately for their highly un-naturalistic task) by assuming that participants have equal prior beliefs for the two potential causes (there, jars) of the observations. This is not an ideal Bayesian observer formula for in the current context. The formula should be corrected at least to*

*\frac{1}{1+ \frac{1-P\_0(G)}{P\_0(G)} (\frac{q}{1-q})^{n\_c - 2 n\_g}}*

*(use https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.numberempire.com%2Flatexequationeditor.php&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=5N53JdUhPdN0mFvvgtpxRu6Zl7ew5h5ODIPbtMzOT9Q%3D&reserved=0 to render the LaTeX above). The authors are praiseworthy for actually having obtained a self-report estimate of P\_0(G), the strength of belief in guilt before any samples are seen.*

*My advice here would be to either use this initial self-report for each individual or fit the P\_0(G) as a free parameter for each individual. In the latter case, I would also advise using at least one further parameter per individual, to fit reporting noise. A third parameter should account for reporting bias, but using several parameters would necessiated hierarchical Bayesian fitting.*

We now include four theoretical computational models that include free parameters for the prior probability of each of the suspect categories. We also explore further parameterisations to explain participants’ sensitivity to disconfirmatory information (which also is relatively absent in the ground truth computation). The best fitting of these models appears to be an original delta model, which is a simple heuristic that begins with a prior probability, and then keeps a running average of the probability of guilt, which is updated by each new evidence sample. We show that this model explains both suspect effects and sensitivity to disconfirmatory information, in contrast to the ground truth computation, which cannot.

We have de-emphasised the comparison of participant behaviour against performance derived from the Moutoussis et al formulation throughout the paper. Where it still appears (in the plots), we have re-branded it the “ground truth computation”. The prior probabilities that one or the other suspect categories would be guilty were literally 50% in the ground truth and they were not dependent on participants’ prior beliefs. We also refer to it now as a “computation” rather than a “model” to avoid giving a misleading impression that this computation explains participant behaviour. To the contrary, we include the ground truth computation on purpose to highlight how prior belief can affect participant behaviour, relative to this standard.

We agree, however, that parameterised models do more directly illustrate how prior belief comes to explain participant behaviour and therefore we hope the inclusion of our new computational models is more satisfying in this regard.

*Now let us consider the implications of the above for the disconfirmatory analysis. Using the above formula, the authors should note that the ideal Bayesian observer her/himself over-weighs witnesses providing disconfirmatory evidence ('oddball'). This can be demonstrated using the equation above, but now rather than using the formula for c samples, consider only the last sample. Then the prior can be written simply as the participant's last belief before the current draw, P\_{t-1}(G). Substituting in the above, we can derive that*

*\frac{P(G|g)}{P\_{t-1}(G)} = \frac{1}{P\_{t-1}(G) + \frac{0.4}{0.6}(1-P\_{t-1}(G))}*

*This means that the increase in the belief in guilt will be \*smaller\* when it is already strong and a 'guilt' witness is seen, than if it is weak and the same witness (now disconfirmatory) is seen.*

We now include a new model comparison, in which we formally compare model fits for three models based on the above Bayesian conditional probability formula with parameterised priors against a heuristic delta model. We find the model comparison favours the latter model.

*The third point concerns the literature. This shows that people prone to psychosis show a bias \*against\* disconfirmatory evidence, not \*increased\* swaying by disconfirmatory evidence. A review is*

*<https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.sciencedirect.com%2Fscience%2Farticle%2Fpii%2FS0005791616300891%3Fvia%253Dihub&data=05%7C01%7Cnicholas.furl%40rhul.ac.uk%7C161efc03696e469105c708db1cef0b7e%7C2efd699a19224e69b601108008d28a2e%7C0%7C0%7C638135585259014829%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=cwUblugaPxAvaksbZZTtvis2UQ%2FeqcMrm5QAQ01BEn0%3D&reserved=0>*

*Therefore it could well be that the 'bias' that people show in general is not a distortion similar to psychosis-spectrum, but indeed an approximation to Bayesian reasoning. But this needs an analysis based on the formulae above.*

The reviewer has raised a topic that is indeed of great interest to us! We have now incorporated a mention of this into our revised General Discussion section.

As we have cited already in our submission, there is prolific evidence favouring overadjustment to disconfirmatory evidence when using the beads task – the task that we investigate here. As it happens, the BADE (bias against disconfirmatory evidence) does not involve the beads task. In the BADE task, participants read a scenario, which unfolds via a sequence of clues. Following each clue, participants rate the plausibility of competing interpretations of the scenario. The clues are arranged such that earlier cues suggest a “lure” interpretation, which later cues then undermine the lure in favour of a “true” interpretation. Deluded individuals are more likely to cling to the lure during these later cues than less deluded individuals. The BADE task retains the sequential evidence presentation nature of the beads task but is otherwise quite different.

The discrepancy between the results stemming from the draws to decision and graded estimated version of the beads task on the one hand and the BADE task on the other hand is well-known and of great interest to our team. We have written a recent review on this matter ourselves, where we argue that this discrepancy reflects a wider disagreement between perspectives on psychosis: Are affected individuals are too sensitive to evidence or have too rigid prior beliefs? <https://osf.io/preprints/psyarxiv/e37c2>. We have also contrasted these tasks in our other work e.g. <https://doi.org/10.2147/NDT.S56870>, <https://doi.org/10.1037/abn0000844>.

Our current submission is not aimed at conclusively resolving this complex debate, but rather at identifying how individuals make decisions in mock forensic settings.

*A further consideration concerns how shallow the real persons' curves in Fig 1 are, compared to ideal observers. I wonder if this requires a further refinement of the formula of Moutoussis et al etc., in this naturalistic setting, but I think that this should be thought about more after an ideal Bayesian observer with non-trivial priors has been implemented and fitted to the behaviour.*

This is indeed an interesting feature of participants’ behaviour. The new delta heuristic we propose appears to explain participants’ relatively conservative adjustments.

*Taking good account of prior beliefs in the ideal observer situation would then have serious consequences as to the identity analyses:*

*1. Is the guilty-male bias (etc) simply a matter of \*estimable\* prior beliefs?*

*2. Or maybe people are 'worse Bayesians' depending on identity?*

*3. If it is a matter of prior beliefs, are these 'ideal' in reflecting the statistics of the real world? Arson is a very serious crime, does 'beyond reasonable doubt' work the same for different identities (the criteria -- so no discrimination) as opposed to the priors, where different priors for different groups may be -- I am not sure! -- justifiable?*

We previously concluded on the basis of the discrepancy between participants’ judgements and that of the ground truth computation, that participants must be bringing a prior belief about the suspect categories into an experiment, where the ground truth priors were 50/50. We believe that the addition of the formal model comparison among models with parameterised prior probabilities strengthens this conclusion with a more direct illustration of how these priors lead to behaviour. Estimable prior beliefs can indeed explain higher guilt ratings for male suspects (point 1), though perhaps a delta heuristic, rather than a Bayesian model best explains participants’ strategies (point 2).

Our submission involves experimental studies where the ground truth (and consequently, the ground truth conditional probability computations) are controlled. We agree entirely that an examination of “real world” statistics is a logical and interesting next step for our approach. We bring up these issues in our revised General Discussion (point 3).