*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 the ground truth computation. The former are meant to theoretically explain participants’ ratings in terms of their prior belief, while the latter was only ever meant to provide a standard by which to evaluate the “accuracy” of participant decisions, considering what the ground truth facts of the paradigm are. We agree of course that a computational theory of participants’ behaviour affords us stronger and more impactful conclusions than a simple comparison 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.*

Given the Editor’s comments, we believe that in-depth coverage of theoretical approaches in the Discussion section appears to be in order. It appears we may not have been clear about the role the ideal observer calculations are meant to play in our paper.

In sum: Although our submission has some computational elements, it was never intended to propose a new theory to theoretically explain evidence accumulation on the beads task in general. Instead, we report purely empirical claims about how participants make probabilistic judgements in realistic tasks, which are interesting BECAUSE they stand in need of theoretical explanation. The use of our ideal observer is just meant to highlight how participants’ estimations can differ from “ground truth” probabilities. The ideal observer is not meant as a computational explanation of participants’ beliefs, any more than the objectively-measured line distances in the Mueller-Lyer illusion are meant to “explain” why participants experience the illusion: It is merely an objective standard that places the behaviour in relief.

Our manuscript, at present, aims to characterise how participants’ beliefs would respond to evidence accumulation in realistic settings, with a view to identifying systematic inaccuracies. For this purpose, we have offered a definition of accuracy (i.e., conditional probabilities, computed using information available to participants in the paradigm). Neither the editor nor either of the reviewers have offered any specific alternatives to this calculation and Bayes rule is indeed the mathematical means of computing conditional probabilities. We have been transparent in our use of this computation.

Our data shows that participants differ from this standard in multiple ways and, for the present paper, those are our main claims. The reasons that participants might differ from this standard is an obvious new research direction that our results suggest. We therefore, have explored these possibilities in our new Discussion, while trying to make clear that, at this stage we can only speculate.

We imagine the editor might have in mind, for example, heuristics that, while leading to some degree of “bias” (as we define it), can produce high accuracy while minimising intrinsic (non-measured) costs, like the need to implement complex calculations. In our view, there is no reason that Bayesian frameworks cannot model such intrinsic processes. Using the same model as Moutoussis et al., for example, we have used an intrinsic sample cost to explain suboptimally low samples to decision (Furl & Averbeck, 2011). Indeed, we are keen to explore this model further and develop other explanatory models within a model comparison framework in future work, to explain why participants deviate from mathematically-computed conditional probabilities. Certainly, we would be very interested to discover that such parameters might have intrinsic adaptive purpose for participants and we do not necessarily agree that our approach exclude this possibility (as we have noted in our new Discussion). Indeed, we note that visual illusions (which involve a deviation between participants’ subjective belief and an objective measurement) are often explained in terms of adaptive mechanisms, such as efficient coding or range adaptation at the neural level. This does not invalidate visual illusions as phenomena worthy of being published or studied, even if (as the Editor rightly encourages here) it should motivate further work into theoretically explaining this “bias”.

However, we regard this formal model comparison of theoretical models to explain this bias as an obvious future research direction, based on our current report. The submitted set of studies is aimed at establishing those biases that would stand in need of explanation, perhaps in terms of adaptive mechanisms.

Our claims are that participants respond to evidence samples in forensic settings in some surprising ways, and that the behaviour differs from an objective standard. It is not clear to us why we need extensive extra cognitive modelling and the proposal of new theories of evidence accumulation. It seems the empirical findings we report are substantiated purely on empirical grounds and that the theoretical work amounts to the expected next step. Indeed our labs are working on precisely this theoretical work at present and we have in mind a much more thorough model specification and formal comparison process than is proposed by the editor and would need to stand as a separate publication (if not publications).

*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.*

Our new Discussion at length speculates on possible theoretical explanations of our findings, which we regard as future directions that stem from the findings we report here. Our submission certainly was purposely intended to serve primarily as an empirical report, which presents new empirical findings that might challenge existing theoretical views. Why individuals jump to conclusions, undersample and overadjust their beliefs on any version of the beads task continues to lack any single theoretical explanation, after many years of publications, a topic about which we have already reviewed (Evans et al., 2013; Furl & McKay, 2023). Our submission does not aim to conclusively resolve this issue. Nevertheless, we expect that our reports of new biases can spur new theory aimed to explain the biases we report, an endeavour on which our labs are already at work, and which we now describe in the Discussion section.

*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.*

It appears that perhaps our procedure is not as clear as it could be. We hope that we have better frontloaded our description of our approach in the Introduction and that our extended Discussion section better explicates the how theoretical modelling approaches might differ from detection of bias based on an objective standard and the differing merits of both of these approaches. Indeed, future directions of our work, we think, should focus on theoretical formal model comparisons that would be aimed at explaining sources of the biases we report here. These sound like fascinating ideas for such models. A computational theory of participants’ belief formation in this task no doubt would need to explain the bias in participants’ prior belief that we report here. And parameters could be added to attempt to explain behaviour in terms of biases towards earlier or later sequence positions, if any exist.

We note that we never claim that the conditional probabilities we compute and compare to participants’ behaviour are suitable to explain participants’ biases. In the paradigm at it was objectively implemented, all suspect categories were in reality equally likely to be guilty or innocent (i.e., the ground truth was that there was no difference in the prior probabilities), the scenarios presented did not state or imply that one suspect category would be more or less likely to be guilty, and thus is would not be normative for an agent to adopt a biased prior and the conditional probabilities that we computed would be incorrect if they were computed from participants’ subjective priors. Moreover, we were especially interested in whether participants’ subjective beliefs would differ from the objective probabilities, and so altering the objectively-computed probabilities to reflect participants’ subjective beliefs would render the measure inappropriate as a normative standard.

We certainly agree that there would be an important place for introducing biased priors into a full model of a Bayesian agent for this task, but that would be as a theoretical computational model to explain participants’ biased behaviour, rather than as the objective standard by which bias is detected. For our objective measure to maintain its objectivity, it is necessary that the computations are based on the paradigm’s ground truth, rather than participants’ subjective impressions about it. As the editor pointed out in the above comment, there certainly might be good reasons for participants to follow their own impressions over the objective ground truth, but our paper focuses on the identification of biases, with building models to explain those biases as the obvious next step, which our labs are actively pursuing now.

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

Our Intro says …

*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.*

The average adjustments stem were proposed in a pre-registered analysis and therefore must be reported. We appreciate the editor’s idea for trying to salvage an effect of suspect type in Study 1 by adding another post hoc analysis that breaks the adjustments down by sequence position. Unfortunately (unless we misunderstand the editor’s suggestion) there is not sufficient data per participant to identify average adjustments to guilt and innocent claims for every sequence position. And, as can be seen by comparing the average probability ratings from one sequence position to the next in the top panels of the figure, there is no effect of suspect when not breaking the data down by guilt and innocent claims.

*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.*

To be frank, we are unsure we understand this comment. The beads task is a probabilistic reasoning task and so we define participants as biased when their probabilistic reasoning does not match the same probabilities when they are mathematically and objectively calculated based on the ground truth of the paradigm. If the objective probabilities are not sensitive to context and participants are, then there is some reason they have not adopted the mathematical solution that stands in need of explanation and for that reason, we are reporting it in this paper. Perhaps it would be helpful if the editor can clarify the logic here for us to better respond to this point?

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

We hope that our revamped discussion of the role of the optimal model, and how in our view it differs from a fitted theoretical computational model helps to clarify things. We reiterate that the aim of the current study is to report empirical findings about how participants form beliefs in response to evidence samples in realistic scenarios, not to propose a computational theory for these beliefs. The terminology “optimal model” here refers to a model most accurately (optimally) reflects the ground truth probability, not the theoretical model that most optimally fits human participants.

Experiment 2 in fact provides strong evidence that *participants* do not use a flat prior on the guilt of different suspect categories. Such an inference should also be true for any plausible computational model that one would propose as a theory of participants’ behaviour, because that model would need to explain participants’ apparent use different priors for different genders in Experiment 2. Such a model would be “optimal” in the sense that it optimally explains human’s choice. That is not the sense in which we mean optimal.

Our ideal observer calculation does not need to explain participants’ biased prior, as it is not an explanatory theory of participants’ behaviour. It is a calculation of the mathematical conditional probabilities of guilt, based on the ground truth of the paradigm. We do not “assume” equal priors. As experimenters, we programmed the paradigm to have equal probability of guilt or innocence for each suspect category and we already know that is the ground truth by which an objective calculation should be based and therefore an agent *should* use a flat prior when calculating probabilities if beliefs are to match the ground truth.

The role of the comparison of participants’ probability estimates to this objective standard is to learn how participants derive their beliefs. **It is because participants deviate from an observer that uses equal priors, that we can claim strong evidence that participants are not using equal priors like that observer is**.

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:

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

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?

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

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

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.

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.

The third point concerns the litterature. 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.

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.

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?