When is one confession better than two? Rational and observed belief updating under competing explanations

Greta Arancia Sanna (greta.sanna.23@ucl.ac.uk)[1], Jens Koed Madsen (J.Madsen2@lse.ac.uk)[2], David Young (dy286@cam.ac.uk)[3] & David Lagnado (d.lagnado@ucl.ac.uk)[1]

[1]Department of Experimental Psychology, University College London
[2]Department of Psychological and Behavioural Science, London School of Economics
[3]Department of Psychology, Cambridge University

Abstract

A central challenge in belief updating is how individuals evaluate evidence when multiple competing hypotheses are plausible. Although prior work shows that belief updating can align with Bayesian principles, this evidence typically involves intuitively directional cases. It remains unclear whether similar rational trajectories are observed in counter-intuitive settings, where accumulating supportive evidence can increase the plausibility of an alternative explanation. In the present study, participants evaluated a scenario in which five sequentially presented confessions were diagnostic either of collective guilt or of coercive interrogation. Early confessions led to increases in belief in both guilt and force, indicating non-zero sum updating. However, contrary to a priori model predictions, participants did not exhibit the non-monotonic pattern whereby belief in force overtakes belief in guilt. Participant-specific Bayesian Network models showed closer alignment with observed trajectories, yet revealed systematic under-weighting of force and over-weighting of guilt. Overall, belief updating reflected sequential evidence integration rather than full Bayesian re-computation.

Keywords: Belief updating; Competing hypotheses; Bayesian updating; Testimony evaluation

Introduction

When presented with evidence, a common problem we might encounter is that of competing hypotheses, where the same evidence could count in favour of multiple explanations. Consider a product with an overwhelming number of five-star reviews (Bhui & Gershman, 2020): while this could be genuine evidence of product quality, it could equally suggest the presence of fabricated reviews. Thus the same evidence, an abundance of positive reviews, can now be explained by two competing hypotheses, and accumulating evidence may paradoxically undermine rather than reinforce confidence in the product's quality. How individuals choose to engage with such cases of competing explanations is central to navigating informational landscapes, as systematic errors in this process may have substantial consequences for belief formation and decision making (Meegan, 2010; Nisbett & Ross, 1980; Pilditch et al., 2019; Smithson & Shou, 2016).

In this vein, Bayesian Network models have been used to benchmark the rationality of human updating. People tend to update their beliefs in the same direction as predicted by rational Bayesian models (Shengelia & Lagnado, 2021); although quantitatively their updating tends to be conservative (Kovach, 2021) and rarely reaches the correct magnitude (Gershman, 2019; Haselton et al., 2009). Moreover, people rationally integrate factors such as source reliability and evidence diagnosticity (Connor Desai et al., 2020; Pilgrim et al., 2024). Importantly, even seemingly irrational updating can often be justified relative to individuals' own rational models of evidence (Benoît & Dubra, 2019; Bullock, 2009; A. J. L. Harris & Hahn, 2009; Jern et al., 2014; Levy, 2021; Mandelbaum, 2019; Olsson, 2013; Pallavicini et al., 2021; Young et al., 2025). However, these studies typically employ relatively intuitive scenarios where rational updating is straightforward; less is known about cases where normatively correct updating is counter-intuitive, such as when accumulating evidence paradoxically increases the probability of a competing hypothesis (Bhui & Gershman, 2020; A. Harris et al., 2009; Orchinik et al., 2024).

These frameworks are especially useful in legal contexts (e.g. Fenton et al., 2013; Lagnado, 2021) where decision makers must evaluate the probative value of competing evidence, integrate testimony of varying reliability, and draw inferences under uncertainty. In an interrogation context, for example, a confession is typically treated as strong evidence of guilt. However, when multiple members of a group confess, people should weigh the relative likelihood of this pattern of evidence under two competing hypotheses: that the group is guilty and confessions are voluntary, or that the confessions are the result of coercion. Crucially, a large number of confessions may be more probable when the suspects are innocent than when they are guilty, depending on assumptions about the evidence-generating process. The present research tests whether individuals are sensitive to these competing explanations and whether their belief updating reflects not only the diagnosticity of confessions, but also their expected frequency under each hypothesis.

The Study

Participants

We recruited 150 UK participants from Prolific (75 female, 75 male: mean age 44, range 21–79), and paid them £0.75 for a 5 minute survey. In line with pre-registration criteria, no participants were excluded from the sample. The study was pre-registered with supplementary material provided on OSF.

Design

Participants were presented with a scenario in which five members of a group were questioned about their involvement

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

in a crime. They were informed that the group could be held in one of two prisons: one known for routinely using coercive interrogation methods that lead to "a very high probability of eliciting a confession regardless of whether the accused are guilty", and one known for good conduct, "where the use of coercive interrogation methods is very unlikely". Participants were then informed of each of the five confessions sequentially, with beliefs about guilt and the use of force measured before and after each confession.

Procedure

The study was administered on Prolific using a Qualtrics survey and took approximately five minutes to complete. After providing informed consent, participants read background information describing five group members accused of planning a terrorist attack, and were told that all the members were either collectively guilty or innocent, and equally likely to be held in either of the two prisons described. Participants then completed three manipulation check questions to assess their understanding of these vignette assumptions: whether force applied to one member increased its likelihood for others, whether the group was equally likely to be held in either prison, and whether guilt applied collectively. Participants were not excluded on the basis of these checks, though subgroup analyses omitted those who did not rate the prior probability of each prison as approximately 5 out of 10.

Participants were then told that the five individuals had been taken simultaneously to separate interrogation rooms, unaware of each other's outcomes. Before any outcome was revealed, participants reported baseline beliefs about the probability of group guilt and the use of coercion on slider scales ranging from 0 (*very unlikely*) to 10 (*very likely*). Interrogation outcomes were presented sequentially, each resulting in a confession, with belief ratings in guilt and force repeated after each confession.

Following the fifth confession, participants judged the likelihood of observing five confessions under two scenarios: when the group was guilty and coercion was used, and when the group was guilty and coercion was not used. Finally, participants completed four conditional probability questions assessing the likelihood of a confession under all possible combinations of guilt and coercion, framed as: *"How likely is a member of the group to confess if the group is (not) guilty and coercion tactics are (not) used?"*. All four questions were presented simultaneously on a single page, with free navigation between them, using the same 0–10 slider scale.

Bayesian Network model

To formalise this intuition, we used an *a priori* mathematical proof (Dawid & Lagnado, 2007) demonstrating that under certain assumptions a rational agent faced with multiple items of corroborating evidence should, rather than further increasing belief in the focal hypothesis, shift belief towards an alternative explanation of the evidence (the full proof can be accessed on OSF). One assumption of the model is that force and guilt are unconditionally independent (but because of the common effect structure, they become dependent conditional on evidence about a confession). Another key assumption is that the probability of a confession when force is used is the same irrespective of whether or not the group is guilty. That is:

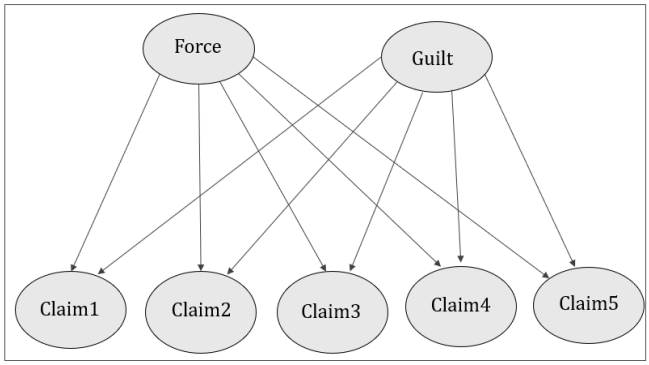

Figure 1: Bayesian Network of the repeated confession vignette used in our study. Guilt = whether or not the whole group is guilty; Force = whether or not coercion was used to extract a confession from every suspect; Claim1-5 = whether or not each interrogation resulted in a confession

$$P(\text{confess} \mid \text{force} \wedge \text{guilt}) = P(\text{confess} \mid \text{force} \wedge \neg\text{guilt}) \quad (1)$$

Finally, the model assumes that the probability of a confession given force and not guilty is higher than given no force and guilty. That is:

$$P(\text{confess} \mid \text{force} \wedge \neg\text{guilt}) > P(\text{confess} \mid \neg\text{force} \wedge \text{guilt}) \quad (2)$$

To assess whether human belief updating follows this normative pattern, we used the Bayesian Network framework (Pearl, 1988). Bayesian networks (BNs) consist of two components: (1) a directed acyclic graph encoding the probabilistic dependencies between variables, and (2) conditional probability tables associated with each variable that specify the strength of these dependencies and support quantitative belief updating via Bayes' rule. The Bayesian network corresponding to the vignette used in our study is illustrated in Figure 1. The model assumes a scenario in which five suspects each give the same testimony (confession). At the core of the network are Claims 1-5 which can be explained by two competing accounts: the claimants are all guilty (Guilt) or they are all the victims of coercion (Force). Note that although these hypotheses are competing, they are not mutually exclusive (both could be true). The prior probabilities for *Force* and *Guilt*, as well as the conditional probability tables governing *Claims 1-5*, are derived from participants' individual judgments. Parameterizing the network with these values allows us to compare participants' observed belief updating in response to the claims with the normative predictions generated by the Bayesian network (see Methods for details).

The core of the model lies in a trade-off between two competing inferences that accumulate as repeated claims are observed. On the one hand, repeated claims may be taken as support for guilt: if multiple sources admit to the group's crime, belief in the claim that the group is in fact guilty should increase. On the other hand, if one expects that guilty suspects would be reluctant to confess, but that this reluctance can be overcome by force, then as each additional confession is presented, the likelihood of force having been used should also increase. Importantly, the use of force can *explain away* the confessions as even innocent people might confess to crimes under torture, simply to make it stop. Therefore as the likelihood of force having been used increases, people might retroactively discount the confessions they have already heard as evidence for guilt. So, while the probabilities of both guilt and force are expected to increase after the first confession, as additional confessions accumulate the inferred likelihood of force should continue to rise, surpassing the likelihood of guilt. At this point, coercion becomes a more probable explanation for the observed pattern of confessions than genuine guilt, causing perceptions of guilt to decrease back to the level of the person's prior belief before any confessions were made.

Given the model outlined above, the empirical aim of this study was to examine the extent to which participants approximate rational inference when evaluating repeated claims in this scenario. Prior to data collection, we conducted a predictive analysis to characterise expected belief trajectories in guilt and the use of force as identical claims accumulated. To parameterise these predictions, we specified a set of indicative conditional probability values capturing the assumption that, under coercive interrogation, confessions are highly likely regardless of guilt (see Table 1). Under this assumption, the model predicts that although both guilt and force increase following the first confession, additional confessions increasingly favour force over guilt, ultimately driving belief in guilt back towards baseline (Figure 2). We note, however, that this prediction depends critically on the assumption that participants treat confessions under force as non diagnostic of guilt, an assumption we did not presuppose would necessarily be shared by all participants.

Table 1: Conditional probability table for confession given guilt and force

Guilt (G)	Force (F)	$P(C = 1 \mid G, F)$
1	1	0.9
0	1	0.9
1	0	0.3
0	0	0.1

Hypotheses

In line with predictions from the Bayesian Network model, our pre-registered hypotheses were:

H1: After the first confession, both a) belief in the likelihood that the group is guilty and b) belief in the likelihood that force was used will increase relative to baseline.

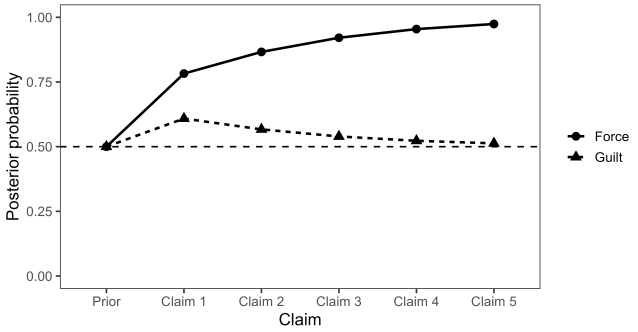

Figure 2: Posterior probability of likelihood of guilt and force across five independent confessions. Representative values were chosen to populate the conditional probability table (CPT) in line with the proof based on our a priori model

H2: Belief in the likelihood of guilt will be lower after five confessions than after one confession.

H3: Belief in the likelihood that force was used will be higher after five confessions than after one confession.

H4: Participants' belief trajectories will directionally align with the Bayesian Network predictions, such that empirical belief updates will positively correlate with model-predicted posteriors for both guilt and force.

Note that although the model formally predicts a decrease following the second piece of testimony for H2 and H3, we evaluated this prediction at the final time point, where the cumulative pattern is expected to be more robust to response noise.

Results

Ahead of the main analyses, we examined the responses to the manipulation check questions[1]. In line with the vignette assumptions, participants indicated that if one group member was guilty, the others were also likely to be guilty (M = 8.92, SD = 1.82). Participants likewise judged that if force was used on one group member, it was likely to have been used on the others (M = 8.11, SD = 2.18). Participants also correctly indicated that the group was equally likely to be held in either of the two prisons (M = 5.03, SD = 1.75). Finally, consistent with the assumptions of the normative proof, participants judged that when the group was guilty, confessions were more likely under force (M = 8.54, SD = 1.82) than in the absence of force (M = 2.94, SD = 5.68). Together, these responses indicate that participants engaged with the task and broadly understood the key assumptions of the vignette.

[1]The analysis was also conducted on a subset of 114 participants who, in line with the vignette introduction, judged the likelihood of the group being in either prison to be approximately equal (ratings of 4–6 on a 10 point scale). This subset analysis yielded qualitatively similar results; therefore, all participants were retained in accordance with the pre registered criteria.

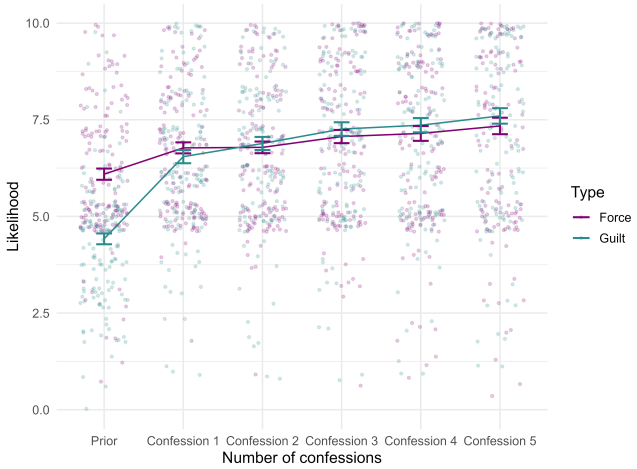

Figure 3: Likelihood of force and guilt across confessions. Note. Ratings ranged from 0 (very unlikely) to 10 (very likely) with error bars representing ±1 standard error of the mean. Individual data points have been jittered along the x-axis to improve clarity and visibility.

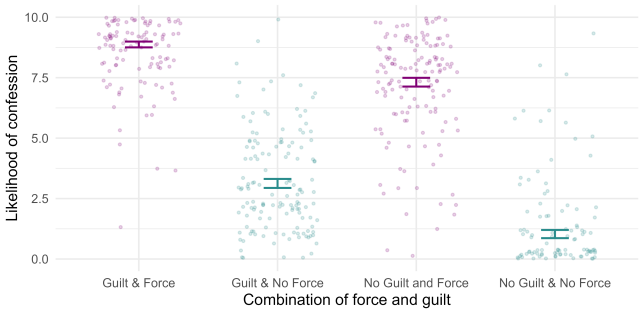

Figure 4: Conditional probability questions posed to participants at the end of the vignette. Participants rated the likelihood of the confession under four different conditions: guilty vs. not guilty, and force vs. no use of force. Note. Ratings ranged from 0 (very unlikely) to 10 (very likely) with error bars representing ±1 standard error of the mean. Individual data points have been jittered along the x-axis to improve clarity and visibility.

Hypotheses Tests

Consistent with H1, belief in the likelihood of guilt was significantly higher after the first confession than at baseline, $t(149) = 12.35$, $p < .001$, Cohen's $d = 1.01$, 95% CI $[0.81, 1.20]$. Belief in the likelihood that force was used to extract a confession was also significantly higher after the first confession than at baseline, $t(149) = 5.05$, $p < .001$, Cohen's $d = 0.41$, 95% CI $[0.25, 0.58]$ (see Figure 3). It should be noted that prior ratings of force (M=6.09, SD=1.78) were significantly higher than those for guilt (M=4.42, SD=1.71).

Regarding H2, contrary to our expectations perceived likelihood of guilt after five confessions was significantly higher than after the first confession, $t(149) = 6.20$, $p < .001$, Cohen's $d = 0.51$, 95% CI $[0.34, 0.68]$.

Finally, consistent with H3, belief in the likelihood that force was used to extract confessions was significantly higher after five confessions than after the first confession, $t(149) = 3.14$, $p = .002$, Cohen's $d = 0.26$, 95% CI $[0.09, 0.42]$.

Looking at individual level variability, 55 out of 150 participants reported the same belief in the likelihood of force after five confessions as after one confession, 57 increased their belief in the use of force, and only 16 decreased it. A similar pattern was observed for guilt judgments: 29 participants reported the same belief in guilt after five confessions as after one confession, 80 increased their belief in guilt, and only 19 decreased it. Overall, these distributions suggest that increases in perceived guilt and perceived use of force were common responses across participants.

Consistent with H4, overall, participants' belief updating qualitatively aligned with the a priori Bayesian Network model with respect to the inferred use of force. However, with regard to the likelihood of guilt updating qualitatively contradicted model expectations, increasing rather than decreasing across confessions. One plausible explanation concerns participants' assumptions about the evidence generation process encoded in the vignette. Participants' conditional probability judgments (see Figure 4) revealed systematic deviations from the assumptions underlying the normative proof. Whereas the proof assumed that confessions obtained under force were equally likely regardless of guilt status, many participants judged confessions under force to be more likely when the group was guilty (M = 8.87, SD=1.47) than when it was innocent (M = 7.31, SD=2.21). Similarly, in the absence of force, confessions were judged to be more likely if the group was guilty (M = 3.13, SD=2.28) than if it was innocent (M = 1.03, SD=2.07). While this latter asymmetry is consistent with the premises of the proof, the former directly violates them. To provide a more appropriate rational benchmark for belief updating, we therefore constructed individual-specific Bayesian Networks parameterized using each participant's own conditional probability judgments. This approach allowed us to make a more principled comparison between observed belief updates and normative predictions, and provides a clearer characterization of the reasoning processes underlying participants' judgements.

Bayesian model comparison

Using the R package gRain (Højsgaard, 2012), individual models were created for each participant. Conditional probability judgments collected from participants were used to parameterize each model. As shown in Figure 5, for perceived force, participants' beliefs directionally aligned with the predictions of their individualised models, albeit updating more conservatively than the model. This pattern was supported by a modest positive correlation between predicted and observed force estimates across confessions, $r = .21$, 95% CI $[.14, .28]$, $p < .001$.

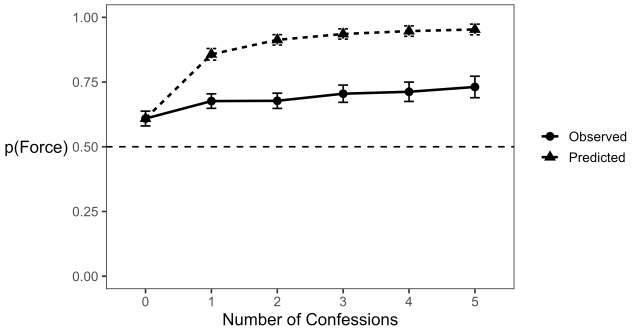

Figure 5: Averaged predicted vs observed trajectories across participants for the likelihood of force across confessions. Note. Ratings ranged from 0 (very unlikely) to 10 (very likely) with error bars representing ±1 standard error of the mean.

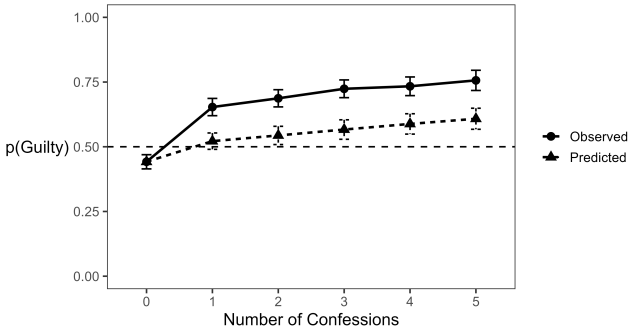

Figure 6: Averaged predicted vs observed trajectories across participants for the likelihood of guilt across confessions. Note. Ratings ranged from 0 (very unlikely) to 10 (very likely) with error bars representing ±1 standard error of the mean.

By contrast, for guilt (Figure 6), participants' observed trajectories diverged from the predictions of the a priori model but nonetheless showed directional alignment with predictions derived from participants' own diagnosticity judgements. This was reflected in a moderate positive correlation between predicted and observed guilt estimates across confessions, r = .40, 95% CI [.33, .45], $p < .001$. Compared to the updating trajectory for force, which was more conservative than model predictions, participants *overestimated* the diagnosticity of the confessions relative to their conditional probability judgment.

Overall, while individual-level predictions for both guilt and force seem to more accurately capture the qualitative direction of belief updating, participants appear to have systematically underestimated the diagnosticity of the evidence with regards to force and systematically overestimated the diagnosticity of confessions with regards to guilt. To assess whether participant-specific Bayesian Network models captured the non-monotonic guilt trajectory implied by the rational account, we focused on model predictions that fall within the parameter regime identified by the proof, namely cases in which confessions under force are judged equivalently irrespective of guilt,

$$P(C \mid G, F) = P(C \mid \neg G, F).$$

Under this assumption, and given appropriate values of the remaining conditional probabilities, the rational model predicts an initial increase in guilt following the first confession, followed by a decrease as additional confessions accumulate and force becomes the more plausible generative explanation. Rather than classifying participants directly on the basis of their conditional probability tables, we operationalised non-monotonicity at the level of the model's predicted posteriors. Specifically, a non-monotonic pattern was defined as an increase in predicted guilt from zero to one confession, followed by a decrease from one to five confessions. Using this criterion, the Bayesian Network predicted non-monotonic guilt trajectories for 31 participants. However, only five of these participants both exhibited non-monotonic predictions and showed the corresponding curvature in their observed belief updates.

Overall, H4 was supported in terms of directional alignment between observed belief updates and model-predicted posteriors. However, participants tended to underestimate the use of force and overestimate the likelihood of guilt relative to model assumptions. Moreover, although a subset of participants' inferred models satisfied the normative conditions for non-monotonic belief updating, only a small fraction expressed this pattern behaviourally. This suggests that the counter-intuitive implications of competing generative explanations, while formally captured by the rational model, were difficult for most participants to follow in their belief updates.

Discussion

The present study explored how people update their beliefs in competing hypotheses (guilt or force) given multiple pieces of evidence (repeated confessions) that appear to support both hypotheses. It also aimed to assess the extent to which people's belief updating fit with an a priori Bayesian model derived from first principles. Following the first confession, participants on average increased both their belief in the group's guilt and their belief that force may have been used to elicit a confession. This pattern qualitatively aligns with the *a priori* "assumed" BN model predictions and contrasts with previous findings suggesting that individuals engage in zero sum reasoning, treating competing hypotheses as mutually exclusive (e.g. Pilditch et al., 2019). With respect to belief updating, between the first and fifth confession, participants increased their belief in the use of force in line with the assumed BN model. However, they also increased their belief in guilt, diverging from model predictions.

One possible explanation for this discrepancy concerns the assumed conditional probability values employed in the model. Subsequent model fits using participants' individual conditional probability judgements revealed closer

alignment between predicted and observed belief updating trajectories; in effect, our initial prediction that participants' guilt perceptions would decrease with an increasing number of confessions turned out to be incorrect because our assumptions about their conditional probability estimates turned out to be incorrect; when participants' *own* conditional probability estimates are inputted to the model, their belief updating trajectory qualitatively agreed with the model's predictions. Nonetheless, while the qualitative trajectories for both force and guilt broadly aligned with the post hoc model using participants' own CPTs (Empirical BN model), participants tended to underestimate the diagnosticity of each confession with respect to force and overestimate its diagnosticity with respect to guilt. Even among participants whose assumptions matched the model, only a small proportion (5 out of 31) exhibited the non-monotonic trajectory predicted by the proof. This belief updating pattern may be explained by two mechanisms discussed in the literature.

First, although participants' priors initially reflected a substantially higher likelihood of force than guilt before any confession was made, this difference was eliminated after the first confession, irrespective of their conditional probability judgments. This pattern is consistent with base rate neglect, whereby individuals undervalue the prior probabilities when integrating new evidence (e.g. Ashinoff et al., 2022; Welsh & Navarro, 2012).

Second, following the first confession, participants' updating appears more consistent with non-normative sequential models (Hogarth & Einhorn, 1990), whereby individuals update an internal evaluation state incrementally rather than recomputing a posterior over all available evidence. This distinction is consequential: for later confessions to decrease perceptions of guilt, individuals must recognise that accumulating confessions raise the likelihood of coercion, which retroactively weakens the evidential force of earlier confessions. This requires revisiting and down-weighting prior evidence, a step participants appear to have largely omitted. Under this account, while the Bayesian Network is insensitive to the chronology of evidence, individuals may not be. Although people evaluate new incoming evidence, they may fail to revise the evidential weight of information received earlier. Once beliefs about guilt and force are anchored, additional confessions may be processed primarily in terms of whether they confirm or disconfirm existing hypotheses, rather than prompting a reevaluation of the full body of evidence. In this context, early confessions might be diagnostic of both guilt and force. However, as confessions accumulate, force becomes the more plausible explanation, given the improbability that all group members would confess voluntarily. Individuals may nonetheless fail to retrospectively apply this revised interpretation to earlier judgements, continuing to increase belief in guilt with each additional confession rather than re-weighting prior evidence in favour of force. This process would ultimately lead participants to judge guilt as substantially more likely than predicted by the Bayesian Network model.

This also aligns with literature suggesting that humans do not engage in explaining away in cases where there are multiple possible causes for the same evidence to the degree that a normative Bayesian model would predict (Liefgreen et al., 2018; Tešić et al., 2020). Future work could directly test this assumption to further characterise these dynamics. It is worth adding, however, that in line with the assumed BN Model, the present results provide evidence of a diminishing returns effect for additional confessions. Specifically, the first confession exerted the largest marginal impact on belief updating, with subsequent confessions producing progressively smaller changes in beliefs. This pattern mirrors paradoxical effects observed in related work on persuasive messages Bhui and Gershman (e.g. 2020) and Orchinik et al. (2024), and, while less counter-intuitive than a fully non-monotonic trajectory, nonetheless reflects a departure from simple additive belief updating. Importantly, this diminishing sensitivity to repeated confessions is predicted by the Bayesian Network model and underscores the relevance of competing generative explanations in shaping belief revision.

Limitations

In terms of limitations, the present study examines a specific instance of competing hypotheses within a legal context, which may not readily generalise to other domains involving consensus or repeated testimony. Moreover, only a subset of participants appeared to engage with the counter-intuitive implications of the normative proof, limiting inferences about the full range of belief updating strategies. Additionally, conditional probabilities were elicited post-task, which may introduce retrospective bias; future work should examine whether similar patterns emerge with pre-task elicitation. Finally, hypotheses were provided to participants rather than self-generated; future work should examine whether participant-generated hypotheses differentially affect belief updating.

Notwithstanding these limitations, the present study advances our understanding of how individuals evaluate accumulating evidence under competing hypotheses, and how cognitive constraints may shape this process. It further demonstrates the utility of participant-specific Bayesian Network models for characterising individual differences in belief updating. These findings carry applied implications for legal decision making, where the sequencing and accumulation of testimony may systematically influence judgments of guilt and coercion.

References

Ashinoff, B. K., Buck, J., Woodford, M., & Horga, G. (2022). The effects of base rate neglect on sequential belief updating and real-world beliefs. *PLOS Computational Biology*, *18*(12), e1010796.

Benoît, J.-P., & Dubra, J. (2019). Apparent Bias: What Does Attitude Polarization Show? [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12400]. *International Economic Review*, *60*(4), 1675–1703. https://doi.org/10.1111/iere.12400

Bhui, R., & Gershman, S. J. (2020). Paradoxical effects of persuasive messages. *Decision*, *7*(4), 239.

Bullock, J. G. (2009). Partisan Bias and the Bayesian Ideal in the Study of Public Opinion. *The Journal of Politics*, *71*(3), 1109–1124. https://doi.org/10.1017/S0022381609090914

Connor Desai, S. A., Pilditch, T. D., & Madsen, J. K. (2020). The rational continued influence of misinformation. *Cognition*, *205*, 104453. https://doi.org/10.1016/j.cognition.2020.104453

Dawid, P., & Lagnado, D. (2007, May). *Gang confessions* [Unpublished manuscript, University College London].

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, *37*(1), 61–102. https://doi.org/10.1111/cogs.12004

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8

Harris, A., Corner, A., & Hahn, U. (2009). Damned by faint praise": A bayesian account. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 292–297.

Harris, A. J. L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1366–1373. https://doi.org/10.1037/a0016567

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias [Place: US]. *Social Cognition*, *27*(5), 733–763. https://doi.org/10.1521/soco.2009.27.5.733

Hogarth, R. M., & Einhorn, H. J. (1990). Venture theory: A model of decision weights. *Management science*, *36*(7), 780–803.

Højsgaard, S. (2012). Graphical independence networks with the grain package for r. *Journal of Statistical Software*, *46*, 1–26.

Jern, A., Chang, K.-m. K., & Kemp, C. (2014). Belief polarization is not always irrational [Place: US]. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

Kovach, M. (2021, January). Conservative Updating [arXiv:2102.00152 [econ]]. https://doi.org/10.48550/arXiv.2102.00152

Lagnado, D. A. (2021, October). *Explaining the Evidence: How the Mind Investigates the World* [Google-Books-ID: OiJJEAAAQBAJ]. Cambridge University Press.

Levy, N. (2021). *Bad Beliefs: Why They Happen to Good People* [Accepted: 2022-02-03T10:29:20Z]. Oxford University Press. https://doi.org/10.1093/oso/9780192895325.001.0001

Liefgreen, A., Tesic, M., & Lagnado, D. (2018). Explaining away: Significance of priors, diagnostic reasoning, and structuralcomplexity. *Proceedings of the annual meeting of the cognitive science society*, *40*.

Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mila.12205]. *Mind & Language*, *34*(2), 141–157. https://doi.org/10.1111/mila.12205

Meegan, D. V. (2010). Zero-sum bias: Perceived competition despite unlimited resources. *Frontiers in psychology*, *1*, 191.

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment.

Olsson, E. J. (2013). A Bayesian Simulation Model of Group Deliberation and Polarization. In F. Zenker (Ed.), *Bayesian Argumentation: The practical side of probability* (pp. 113–133). Springer Netherlands. https://doi.org/10.1007/978-94-007-5357-0_6

Orchinik, R., Dubey, R., Gershman, S. J., Powell, D. M., & Bhui, R. (2024). Learning from and about scientists: Consensus messaging shapes perceptions of climate change and climate scientists. *PNAS nexus*, *3*(11), pgae485.

Pallavicini, J., Hallsson, B., & Kappel, K. (2021). Polarization in groups of Bayesian agents. *Synthese*, *198*(1), 1–55. https://doi.org/10.1007/s11229-018-01978-w

Pearl, J. (1988). Embracing causality in default reasoning. *Artificial Intelligence*, *35*(2), 259–271.

Pilditch, T. D., Liefgreen, A., & Lagnado, D. (2019). Zero-sum reasoning in information selection. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *41*.

Pilgrim, C., Sanborn, A., Malthouse, E., & Hills, T. T. (2024). Confirmation bias emerges from an approximation to Bayesian reasoning. *Cognition*, *245*, 105693. https://doi.org/10.1016/j.cognition.2023.105693

Shengelia, T., & Lagnado, D. (2021). Are jurors intuitive statisticians? bayesian causal reasoning in legal contexts. *Frontiers in Psychology*, *11*, 519262.

Smithson, M., & Shou, Y. (2016). Asymmetries in responses to attitude statements: The example of "zero-sum" beliefs. *Frontiers in psychology*, *7*, 984.

Tešić, M., Liefgreen, A., & Lagnado, D. (2020). The propensity interpretation of probability and

diagnostic split in explaining away. *Cognitive psychology*, *121*, 101293.

Welsh, M. B., & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, *119*(1), 1–14.

Young, D. J., Madsen, J. K., & De-Wit, L. H. (2025). Belief polarization can be caused by disagreements over source independence: Computational modelling, experimental evidence, and applicability to real-world politics. *Cognition*, *259*, 106126.