**Heuristic Belief Updating in Mock Forensic Scenarios**

Didrika S. van de Wouw, Natalia Bogowicz, Sunaina Kanda, Rujiba Akthar, Matteo Lisi, Ryan T. McKay and Nicholas Furl

Department of Psychology, Royal Holloway, University of London

**Author Note**

Data and code are available at <https://github.com/nicholasfurl/Forensic-beads-paper-1>. Study materials (stimuli) will be uploaded to a public archive following acceptance of the manuscript for publication and will be available for sharing by email request from the corresponding author in the meantime.

Word count:

We have no conflicts of interest to disclose

Correspondence concerning this article should be addressed to: Nicholas Furl, Department of Psychology, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, United Kingdom, nicholas.furl@rhul.ac.uk

**Abstract**

A ubiquitous challenge concerns how to form accurate beliefs and choose rewarding actions, despite access to only limited and noisy evidence. Forensic scenarios pose a notable example, where evidence favouring guilt or innocence can arise from only a few potentially unreliable or even conflicting eyewitnesses. We adapted the beads task and computational modelling to study probabilistic reasoning strategies that might arise in forensic scenarios. For one such strategy, participants rated male (versus female) suspects as more guilty, incorporating their prior beliefs about males, though we found negligible effect for suspects varying in religious or mental health status. For the second strategy, participants gave more weight to “oddball” witness claims (which conflicted with preceding witness claims), and gave less weight to claims that were more consistent with preceding witnesses. We demonstrate a heuristic which can explain both of these effects and which better fits participants’ behaviour than the three Bayes rule based models we tested. In this heuristic, participants’ estimates of the probability of guilt depends on a suspect-specific prior probability, which participants then update based on the prediction error associated with each new witness claim. Our new behavioural paradigm and associated computational approach could be extended to predict choices in real forensic settings.

**Introduction**

Beliefs and decisions undertaken in the face of uncertainty or ambiguity are an integral part of everyday life and therefore have been researched by psychologists, economists and sociologists over many decades. In real-world settings, people must often form beliefs and choose actions using limited samples of unreliable evidence. This is especially challenging when the evidence needed to decide is not delivered all at once. In real-world policing/judicial scenarios, eyewitnesses who report uncertain information are typically interviewed sequentially. The forensic decision maker must update beliefs upon the receipt of new evidence accurately enough to motivate appropriate action (e.g., an arrest or conviction).

The beads task has become prevalent for experimentally measuring how participants make beliefs and decisions from sequences of discrete, unreliable evidence samples (Ross, et al., 2015), making it a prime candidate for application to forensic evidence evaluation. There are two versions, both of which we will investigate herein. In the “draws to decision” version (Huq et al., 1988), participants infer from which of two jars coloured beads are being drawn. One possible jar contains a majority of pink beads and a minority of green beads, and the other has a reversed ratio. Following each draw, participants either infer which jar or seek more evidence (draw another bead). In the “graded estimates” version, participants experience the entire sequence of bead colours and, following the presentation of each drawn bead, they rate the probability it comes from one jar or the other. Both versions of the beads task are associated with well-established Bayesian “solutions”, which use the ground truth facts about the experimental paradigm to prescribe which action participants should take (Averbeck, 2015; Furl & Averbeck, 2011; Hauser et al., 2017; Hauser, et al., 2018; Moutoussis et al., 2011; Romero-Ferreiro, et al., 2022; Sonnemans and van Dijk, 2012; van der Leer, et al., 2015). The extent to which such formulations also can explain participants’ behaviour is a question we take up herein.

Our study will be among the first to use realistic scenarios for the beads task. Some previous frames were not so realistic: colours behind tiles (Balzan et al., 2017), fish colours in ponds (Pytlik et al., 2021; Speechley et al., 2005; van der Leer et al., 2014; van der Leer et al., 2015), personality trait names associated with a person (Dudley et al., 1997; Fraser et al., 2006; Menon et al., 2006; Young & Bental, 1997), happy or angry faces from different villages (Romero-Ferreiro, et al., 2022) and commodities manufactured in a factory (Globig et al., 2021). Other, somewhat more realistic attempts included choosing whether a co-worker is having an affair with her boss (Lincoln et al., 2011), whether one’s boss is observing them and whether a person has a drinking problem (Westermann et al., 2012). Participants might draw less evidence for more social scenarios (Dudley et al., 1997; Warman & Martin, 2006), but not always (Balzan et al., 2017; Fine et al., 2007; Menon et al., 2006; Romero-Ferreiro, et al., 2022). Our variant resembles that of Wilkinson and Caulfield (2017), where participants read crime vignettes and then decide whether defendants did “the right thing”, by drawing new statements about the crime. Similarly, in Sonnemans and van Dijk’s study (2012), participants acted as a fictitious judges, requesting evidence about a defendant.

Following the lead of such studies, we modified the beads task to simulate realistic forensic scenarios such as crime investigations and courtrooms, in which participants are given evidence samples about the guilt or innocence of a suspect in the form of witness claims. This framework allows us to use an experimental and controlled setting to investigate and detect strategies and behaviours participants might use in forensic settings “in the wild”. Moreover, the experimental setting allows us to implement computational models to theoretically explain any participant strategies we might observe. We used this opportunity to investigate and model two types of decision strategy that we hypothesised might manifest in forensic scenarios.

For the first strategy, we investigated how prior beliefs about various demographic categories (atheist versus Christian, male versus female, mentally ill versus healthy) would alter how participants respond to evidence about the suspects’ guilt or innocence. Indeed, prejudice against certain categories of suspect can manifest in real-world decisions (Ahola, Hellström, & Christianson, 2010) and influence outcomes of forensic decision-making scenarios (Dror, 2020). Thus, we hypothesised that prior beliefs would sway belief updating about suspects from demographic groups about which are commonly held stereotypes of criminal behaviour and that this influence would be detectable using our novel “forensic beads” framework. We also present computational models which incorporate prior probabilities of guilt as a parameter, which can be estimated from participants’ behaviour, allowing us to quantify the influence participant prior belief exerts over participants’ judgements of guilt for different suspect demographics.

The second strategy we will investigate concerns sensitivity to disconfirmatory versus confirmatory evidence. In the clinical literature (Huq et al., 1988), individuals with psychosis “jump to conclusions” – they make fewer draws to decision than healthy individuals. Jumping to conclusions is relatively well-replicated and bolstered by meta-analyses (Dudley et al., 2016; Fine et al., 2007; McLean et al., 2017; Ross et al., 2015; So et al., 2016) and large-sample studies (Henquet et al., 2020), although relationships between this phenomenon and delusion severity (Baker et al., 2019) and potentially mediating / confounding variables (Pytlik et al., 2021; Ross et al., 2016; Tripoli et al., 2021) remain unclear. Jumping to conclusions also extends to the general population to some degree. Subclinical jumping to conclusions is associated with individuals at risk of psychosis (Broome et al., 2007), asymptomatic first-degree relatives of individuals with psychosis (Henquet et al., 2020; Van Dael et al., 2006) and healthy individuals who score highly on questionnaires measuring prone-ness to delusion or anomalous belief (Colbert & Peters, 2002; Lincoln et al., 2011; Henquet et al., 2020; McKay et al., 2006; Tripoli et al., 2021; Van Dael et al., 2006; Zawadzki et al., 2012), although findings do not always replicate (So & Kwok, 2015; Warman et al., 2007). For present purposes, it is important to note that jumping to conclusions may be related to a parallel finding from the graded estimates task: Individuals with psychotic traits adjust beliefs more than healthy individuals when newly-sampled evidence is disconfirmatory; that is, newly-drawn bead colours mismatch colours from previous samples of beads (Adams, et al., 2018; Fear & Healy, 1997; Fine et al., 2007; Garety et al., 1991; Langdon et al., 2010; Moritz & Woodward, 2005; Strube et al., 2021; Stuke et al., 2017; Peters & Garety, 2016; Young & Bentall, 1997). Herein, we will use the graded estimates version of the beads task to test the hypothesis that participants (to some degree) show greater sensitivity to disconfirmatory evidence than to confirmatory evidence. Moreover, having established this effect in our sample, we will use computational modelling to test theories about the origin of this effect. As discussed in detail in the Methods section of Experiment 2, we designed a relatively simple heuristic “delta model” and compared it against three models based on the Bayesian conditional probability formula, adding free parameters to this formula that might explain how sensitivity to disconfirmatory evidence comes about.

Experiments 1 and 2 use the graded estimates version of the beads task to test for influences of prior beliefs about suspects (atheist versus Christian in Experiment 1 and male versus female in Experiment 2) and to test for special sensitivity to disconfirmatory evidence. In Experiment 2, we will fit and test computational models, which will attempt to explain these two effects. Experiment 3 will use the draws to decision version of the beads task to test whether participants sample evidence differently, depending on suspect mental health (mentally ill versus healthy). To anticipate, participants will differentially rate males (versus females) as guilty and we will replicate sensitivity to disconfirmatory evidence in Experiments 1 and 2. Our model comparison suggests that the delta heuristic model, which incorporates prior belief about the guilt of male suspects, and is based on a prediction error computation, can better explain both effects, compared to any of the three models based on Bayes rule that we tested.

**Experiment 1**

Experiment 1 used the graded estimates version of the beads task, where participants rate their beliefs about the guilt of a crime suspect following witness claims of guilt or innocence. findings17),we tested if the guilt beliefs of participants from the United States (from a majority Christian background) would be most swayed by witness claims about the guilt of atheist suspects and witness claims about the innocence of Christian suspects. We also tested whether participants’ belief updating would be especially sensitive to disconfirmatory evidence.

**Method**

*Participants*

We enrolled 599 participants (ages 18-75 yo, M = 37.09 yo, SD = 13.34; 319 men, 274 women, six participants who selected “other”) from the United States through the online recruitment service Prolific ([www.prolific.co](https://www.prolific.co/)).  The local ethics committee approved the experiment.

*Procedures*

Gorilla Experiment Builder [www.gorilla.sc](http://www.gorilla.sc) (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) hosted the experiment. Each participant underwent one trial, which included a crime scenario (See Supplementary Materials for an example) followed by a sequence of witness claims. The religious affiliation of the suspect mentioned in the scenario was manipulated, such that each participant was randomly-assigned either “American Atheists” or “the Christian Church”. We sampled participants from the United States, who might be familiar with American Atheists. Religious affiliation appeared within a list of suspect characteristics.

Before viewing witness claims, participants rated prior beliefs about the suspects’ guilt using a slider scale from 0 (completely innocent) to 100 (completely guilty). The slider appeared when participants clicked on the scale, with the rating number appearing above selected locations on the slider. Participants indicated suspect guilt beliefs using the same slider scale. Witness appeared as two silhouette avatars, resembling a man or woman, randomly assigned to each witness with 50% probability, with a dialogue bubble declaring “*It was him*” (Guilty claim) or “*It wasn’t him*” (Innocent claim). Trials could be guilt (six guilty, four innocent claims) or innocent (six innocent, four guilty) sequences (randomly assigned to sequences with 50% probability). This proportion of claims substantiated the instructions that “*60% of the witnesses to a crime are accurate about what they report, and 40% are mistaken*”. Innocent and guilty witness claims were randomised. All ratings were self-paced. A blue bar across the top of the screen indicated progression through each sequence.

After completing the sequence, participants answered a demographics questionnaire which queried age, gender (Male, female, other- please specify) and religious affiliation (Christian (Catholic), Christian (Baptist), Christian (Other), Hindu, Buddhist, Muslim, Jewish, Sikh, None, Atheist, Agnostic, Other (Please specify)). Participants also used a slider to answer “*How strongly do you believe in God or gods (from 0-100)? To clarify, if you are certain that God (or gods) does not exist, please put “0” and if you are certain that God (or gods) does exist, then put “100”*.

*Ground truth computation*

Following previous work, , we used the ground truth details in the study to compute the mathematical probability of guilt , conditional on the current number of guilty claims and the total number of claims so far The probability *q* of the majority claim was fixed to 0.6 as per the instructions. The conditional probability of innocence was ,.

We computed ground truth probability for each sequence that was seen by a participant. We did not include any prior probability term and instead used a formula that reflected the flat, 50% ground truth prior probability using in the study and controlled by the experimenter.

Nevertheless, we parameterised the prior and other aspects of this conditional probability formula, fit these models to choices of individual participants and performed model comparisons to identify the best-fitting model parameterisation. Unfortunately, we found parameter recovery of the best model to be unacceptable, presumably because the single sequence per participant provided insufficient data for accurate fitting. Nevertheless, we will report results from the same fitting and model comparison procedure in Experiment 2.

*Design and analysis*

Data and analysis code are available at <https://github.com/nicholasfurl/Forensic-beads-paper-1>. Using MATLAB R2022b (The MathWorks, Nattick, MA) and JASP (version 0.14; JASP Team, 2020), we conducted three “base” analyses and some variations further explained below. They tested hypotheses preregistered at aspredicted.org #25182 <https://aspredicted.org/blind.php?x=5qr89i>, and some basic data features and auxiliary hypotheses.

In the first base analysis, the *Probability Analysis*, *Probability* (ratings) was the dependent variable. The goals were to (1) confirm a basic expected data feature: that guilt probability increased or decreased as sequences progressed, consistent with the majority evidence, and (2) test effects of religious status of the suspect. This ANOVA implemented between-participant factors *Suspect* (atheist versus Christian) and *Sequence Guilt* (guilty versus innocent sequence) and the repeated measures factor *Sequence Position* (positions 0-10).

Preregistered hypothesis #2 predicted a Suspect effect at sequence position 0. Additionally, pre-registered hypothesis #3 predicted, in part, that this Suspect effect would be modulated by participants’ religious affiliation. We therefore computed a between-participants binary variable Affiliation by coding responses on our demographic questionnaire as “religious” if participants responded Christian, Buddhist, Muslim, Hindu, Sikh or Jewish (*N* = 268) and “non-religious” if participants responded none or atheist (*N* = 187) and excluding the 144 participants with other responses.

In the second base analysis, *Adjustment Analysis*, *Adjustment* was the dependent variable - the subtraction of probability ratings before and after viewing each witness claim. Its positive or negative sign indicates whether witness claims increased or decreased guilt belief. This analysis’s goals were to (1) confirm basic expected data features: that guilt and innocent witness claims sway adjustments respectively towards guilt or innocence, and (2) test pre-registered hypothesis #1: participants positively adjust belief most to guilty claims about atheist suspects. This ANOVA implemented a between-participant factor *Suspect* (atheist versus Christian) and a repeated-measures factor *Witness Claim* (guilty versus innocent). We further addressed preregistered hypothesis #3’s prediction of larger Suspect effects for religiously affiliated participants by adding Affiliation as another factor to the Adjustment Analysis.

Further modifications of the Adjustment Analysis tested for heightened sensitivity to disconfirmatory evidence. We created a between-participants factor *Preceding Context,* defined bycategorising claims as preceding guilty or innocent context, depending on if they were preceded by >50% guilty or innocent claims respectively (and removing data with 50/50 context). We then implemented an ANOVA with factors *Preceding Context* and *Witness* *Claim* using only data from participants. In a parallel analysis, we implemented a linear mixed model with the covariate *Proportion Guilt Context* (i.e., proportion of preceding guilty claims), while treating participant as a random effect.

**Results**

*Basic data features of probability and adjustment*

We first orient the reader to basic data features, important for contextualising the reports that follow. Figure 1 overviews evolution of participants’ Probability (0 = innocent, 100 = guilty) as sequences progressed for guilty (60% guilty claims) and innocent (60% innocent) sequences. Participants estimated somewhat less than 50% probability before viewing witnesses (position 0). Then, for guilty sequences, participants adjusted probability positively and, for innocent sequences, adjusted negatively. Participants updated their probability more conservatively than the ground truth computation. This overall pattern of participant performance was confirmed by the Probability Analysis. We observed a Sequence Guilt × Sequence Position interaction *F*(10, 5950) = 11.12, *p* < .001, *ηp2* < .02 and main effects of Sequence Guilt *F*(1, 595) = 41.47, *p* < .001, *ηp2* = .07 and Sequence Position *F*(10, 5950) = 4.79, *p* < .001, *ηp2* = .008. Suspect did not produce any higher order interactions with Sequence Guilt or Sequence Position (*p*s ≥ .13). We will continue our coverage of the Probability Analysis with a report of the main effect of Suspect in a later section.

**Figure 1.**

*Participant and ground truth probability for Experiment 1.A group of graphs showing the same number of objects

Description automatically generated with medium confidenceNote.* Upper panels (in blue) show conditional probability of guilt, computed mathematically using ground truth facts about the paradigm. Lower panels show participant estimates of guilt probability. Position 0 refers to the rating before any witness claim.

The Adjustment Analysis confirmed our expected data feature that adjustments would be positive or negative on average depending on whether guilt claims were positive or negative, respectively (Figure 2). Indeed, there was a significant main effect of Witness Claim *F*(1, 597) = 210.39, *p* < .001, *ηp2* = .26, with no Witness Claim × Suspect interaction (p = .83). Generally, participants adjusted more conservatively, compared to the ground truth probability. The next section reports the Suspect main effect.

**Figure 2.**

*Adjustment in Experiment 1.*

A screenshot of a graph

Description automatically generated

*Note:* Upper panel shows Suspect effects. Lower right panel shows Preceding Context effects. Error bars are 95% confidence intervals. Blue horizontal lines indicate ground truth adjustment.

*Null effects of Suspect, Affiliation on Participants’ Probability and Adjustment*

Pre-registered hypothesis #1, in part, states “*participants, when confronted with positive (but uncertain) testimony of a suspect’s guilt, will make larger adjustments to their probability estimates about that suspect’s guilt (being more inclined to take on board that positive evidence) if the suspect is an atheist than if the suspect is a Christian. Conversely, when confronted with positive (but uncertain) testimony of a suspect’s innocence, they will make smaller adjustments to their probability estimates about that suspect’s guilt (being less inclined to take on board the evidence of innocence) if the suspect is an atheist than if the suspect is a Christian*.” In spite of this hypothesis, there were no Suspect main effects in the Probability Analysis *F*(1, 595) = 1.61, *p* = .21, *ηp2* = .003 (Figure 1) or Adjustment Analysis *F*(1, 597) = 1.31, *p* = .25, *ηp2* = .002 (Figure 2).

The second pre-registered hypothesis states “*participants will rate an atheist suspect as more likely than a Christian suspect to be guilty before they are provided with any witness testimony.*” In spite of this hypothesis, there was no significant effect of Suspect on participants’ probability at sequence position 0 *t*(587) = -1.6, uncorrected *p* = .11(Figure 1).

Our third pre-registered hypothesis was that Suspect effects “*would be moderated by religiosity*.” The Suspect × Affiliation interaction was non-significant for Adjustment *F*(1, 451) = .37, *p* = .55, *ηp2* = <.001 and Bias *F*(1, 451) = 0.004, *p* = .97, *ηp2* < .001. Nor were there significant main effects or interactions involving Affiliation (*p*s ≥ .39).

*Participants adjust more to disconfirmatory claims*

Participants showed a main effect of Preceding Context *F*(1, 95) = 9.11, *p* = .003, *ηp2* < .09, in the absence of any Witness Claim × Preceding Context interaction *F*(1, 95) = 0.90, *p* = .59, *ηp2* = .003, confirming that adjustments in innocent contexts yielded more positively-deflected adjustments than did guilty contexts. In contrast, tground truth computationFigure 2perform similarly to the ground truth computation when evidence is disconfirmatory, but adjust less than the ground truth computation when evidence is confirmatory.

**Discussion**

We did not find any evidence for our hypotheses that participants would differentially interpret evidence related to an atheist suspect compared to a Christian suspect. Nevertheless, more experiments are needed to test for differential evidence evaluation for other groups, about which participants may hold prior beliefs. In Experiment 2, we will investigate gender and in Experiment 3 we will investigate mental health status. Given that the information about religious affiliation was not presented especially prominently within the paradigm (i.e., it was only in the opening scenario vignette, listed in the midst of other suspect information), future iterations of the paradigm may benefit from modifications to make the relevant demographic information more salient, such as leaving it on the screen throughout the task (See Experiments 2 and 3) or conveying the information using more engaging materials like images (See Experiment 2).

Participants were more swayed by “oddball” samples of disconfirmatory evidence than they were by samples of confirmatory evidence. It remains to be seen whether this effect is replicable (although this will be confirmed in Experiment 2) and whether it manifests in real-world forensic scenarios. This strategy may reflect a less extreme version of the “overadjustment to disconfirmatory evidence” that appears especially acute in the delusion-prone and those with psychosis (Mortiz & Woodward, 2005). Interestingly, the ground truth computation did not show the same behaviour, but instead was less sensitive to preceding context than participants, suggesting that participants’ behaviour arises from an influence beyond Bayesian conditional probability computation. We will address the source of this influence directly using formal theoretical modelling in Experiment 2.

**Experiment 2**

Experiment 2 tested whether there exists differential evidence evaluation for suspects of different genders. Prior beliefs about genders can influence the decisions made during legal insanity evaluations and courtroom proceedings (Lindholm & Cederwall, 2010; Norton, et al., 2006; Rogers & Davies, 2007; Yourstone et al., 2008) and judicial proceedings (Lindholm & Cederwall, 2010), and visible attributes of a victim, witness or defendant (such as gender) can influence individuals making decisions as a juror, judge or expert witness (Devine & Caughlin, 2014; Dror, 2013). Also, with the manipulation of suspect gender, we used a more sensitive within-participants manipulation of Suspect and we used facial photographs to render the demographic category of the suspect more salient to participants without also making the hypothesis of the study obvious. Experiment 2 also attempted to replicate our findings that participants are especially sensitive to disconfirmatory evidence when they update their beliefs. Last, we built and formally compared four theoretical computational models (described below in the Methods), designed to explain our (hypothesised) findings of Suspect and any sensitivity to disconfirmatory evidence.

**Method**

*Participants*

At consent, 104 participants provided age (18-64 yo, *M* = 31.05, *SD* = 13.19) and gender (men = 39, women= 63, other = 2). Participants accessed an online link, disseminated via social media (e.g., Facebook) and email. The experiment was approved by the local ethics committee.

*Procedures*

Gorilla Experiment Builder (Anwyl-Irvine, et al., 2020) hosted the experiment, which was a variation of Experiment 1. Witnesses comprised 30 male and 30 female neutral-expression faces, sampled from the 3DSK set (DeBruine & Jones, 2020). Suspect faces comprised four male and four female neutral-expression faces, sampled from the KDEF set (Lundqvist, Flykt, & Ӧhman, 1998).

Example instructions are presented in the Supplementary Materials. Every participant encountered the same eight crime scenarios. Each of the eight sequences that followed began by describing its scenario, alongside the suspect’s face. This suspect’s image remained throughout each sequence. Like Experiment 1, participants rated prior belief of guilt using a sliding scale ranging from 1 (completely innocent) to 100 (completely guilty). When participants clicked on the scale, the slider would appear with the rating number just above. Participants then used this scale to rate each of the ten witness claims. All ratings were self-paced. The witness’s face appeared alongside the suspect’s face, together with a speech bubble stating either “*I’m 70% certain that the subject is innocent!*” or “*I’m 70% certain that the suspect is guilty!*” A counter indicated how many witnesses remained. After each sequence, a screen indicated the number of remaining sequences. Each of the eight suspects was associated with a corresponding scenario, and fixed orders of guilt and innocent claims and witness photos (Supplementary Table 1). The first five witnesses were female and the last five male. Four scenarios were “guilt sequences”, in which 70% claims were guilty.

*Theoretical Models*

We tested four computational models, designed to jointly explain effects of suspect (in terms of prior expectations of guilt) and enhanced adjustment to disconfirmatory evidence. The stimuli input to these models were the same eight sequences as encountered by each individual participant (Supplementary Table 1). Free parameters were estimated by minimising a sum squared error loss function using fminsearchbnd.m function in Matlab R2022B. We then compared the model loss across the four models by computing the Bayesian Information Criterion (BIC)(Supplementary Materials include analysis of model residuals to show this loss function satisfies assumptions of BIC). Successful parameter recovery for the winning model, the delta model, is described in the Supplementary Materials.

Three of the four models were based on the aforementioned conditional probability formula, but with an adjustment included for the prior probability of guilt .

Two free parameters were estimated for , one for each suspect gender. During fitting, these parameters were initialised to their ground truth values of 50% and bounded between 0 and 100). In the *Primacy Model*, is computed using only the first *N* samples and the current sample under consideration, with *N* treated as a free parameter (initialised to include the whole sequence and bounded between 1 and 10 evidence samples). Such a model would embody a strategy where participants are disproportionately influenced by the earliest evidence encountered, a strategy observed in forensic settings (Dror & Kukucka, 2021). In the *Recency Model*, is computed using only the most recent *N* samples and the current sample under consideration, with *N* treated as a free parameter (initialised to include the whole sequence and bounded between 1 and 10 evidence samples). This model embodies a strategy where a limited number of evidence samples can be held in working memory. In the *Split Model*, we treated *q* (i.e., the probability of the majority claim or the guilt / innocence split) as a free parameter, in case participants in practice use a higher probability of guilt / innocence than instructed. Our simulations show that larger *q* leads to greater sensitivity to disconfirmatory evidence. Compare blue lines for the ground truth computations in lower panels of Figures 2 and 4, where the .6 split in Experiment 1 visibly shows less sensitivity to disconfirmatory evidence than the .7 split in Experiment 2. The free parameter *q* was initialised to its ground truth value of .7 and bounded between 0 and 1. Decision noise for these three models was modelled using a regression with free parameters intercept (initialised to 0 and bounded between 0 and 1) and slope (initialised to 1 and bounded between 0 and 1).

The *Delta Model* captures how an observer might incrementally update their estimate of the probability of drawing a bead of a certain colour (or of a guilt/innocent testimony), following each new observation. The probability update for the evidence sample *t* is given by:

Where indicates (with a 1) whether the *t*th claim is guilty (and 0 if innocent). Then:

The free parameter *α* is the learning rate, which controls how responsive is the participant to new information (initialised to .5 and bounded between 0 and 1). The free parameter, *β* is the slope of the Softmax function, which controls how decisively the participants differentiate between the guilty/innocent choice options (initialised to 1 and bounded between 0 and Inf).

Thus, in the delta model, probability updates are made proportionally to the prediction error, that is the difference between what was expected and what was observed. Unlike the Bayesian ground truth computation, which requires more complex calculations and memory of the entire sequence, the delta rule model only requires updating a single estimate, so it is computationally simpler and might be seen as more biologically plausible. Like the other three models, the delta model’s free parameters include separate prior probability parameters for each suspect gender (i.e., the initial values of before any claims are processed).



*Design and analyses*

We implemented the same base analyses as Experiment 1 using dependent measures Probability and Adjustment to characterise basic data features and test hypotheses about Suspect, Preceding Context and Proportion Guilt Context. Data and analysis code are available at <https://github.com/nicholasfurl/Forensic-beads-paper-1>.

**Results**

*Basic data features of Probability and Adjustment*

As in Experiment 1 (Figure 1), participant probability (Figure 3) approximated 50% on average at position 0. Probabilities in guilt sequences climbed modestly towards guilt while, in innocent sequences, they fell. Consistent with this, the Probability Analysis showed a Sequence Position × Guilt Sequence interaction *F*(1, 1030) = 34.89, *p* < .001, *ηp2* = .25, and main effects of Guilt Sequence *F*(1, 103) = 69.8, *p* < .001, *ηp2* = .4 and Sequence Position *F*(10, 1030) = 16.63, *p* < .001, *ηp2* = .14. Unlike Experiment 1, higher-order interactions with Suspect (reported in next section) qualified these effects. Ground truth probability resembled participant trends, though (as in Experiment 1), participants were more conservative in their use of the probability scale, showing only about 10% probability change across the sequence, compared to the ground truth computation, which changed about 50%.

*A group of graphs showing different types of women's

Description automatically generated***Figure 3.**

*Behavioural performance for Experiment 2.*

*Note.* The upper row (blue) shows probability estimates derived from the ground truth computation. The middle row show (black) participants’ probability ratings. The lower row (green) shows probability estimates simulated by the delta model after fitting to participant ratings. Position 0 refers to the prior probability, before any witness claim is experienced, where a higher prior beliefs about the guilt of males is already visible. Asterisks show significant suspect gender effects at *p* < .05, Bonferroni corrected for 22 tests.

As in Experiment 1, the interpretation of statistical effects of adjustment in Experiment 2 hinges on the notion that guilt witness claims on average give rise to positive-valued adjustment (i.e., towards guilty) while innocent witness claims give rise on average to negative-valued adjustment (i.e., towards innocent), consistent with the significant main effect of Witness Claim that we observed *F*(1,103) = 63.08, *p* < .001, *ηp2* = .38, with no interaction with Suspect *F*(1,103) = .01, *p* < .91, *ηp2* < .001 in the Participant Adjustment Analysis. The next section reports the Suspect main effect.

*Effects of Suspect on Probability and Adjustment*

Participants rated male suspects as more guilty than female before viewing any witness claims (Figure 3) and at many later sequence positions. This differential rating for different suspects and sequence positions gave rise in the Participant Probability Analysis to significant Suspect × Sequence Position × Guilt Sequence *F*(10, 1030) = 7.7, *p* < .001, *ηp2* = .07 and Suspect × Sequence Position interactions *F*(10, 1030) = 19.49, *p* < .001, *ηp2* = .16 and a Suspect main effect *F*(1, 103) = 31.92, *p* < .001, *ηp2* = .24. Surprisingly, we obtained the opposite effect in the Participant Adjustment Analysis (lower left panel, Figure 2): Participants adjusted more towards guilty for *female* than for male suspects, resulting in a significant main effect of Suspect *F*(1, 103) = 30.83, *p* < .001, *ηp2* = .23 in the Participant Adjustment Analysis. We suspect that participants began sequences biased against male suspects but then corrected belief towards innocent to align with experienced witness claims.

**Figure 4**

*Adjustment in Experiment 2*

*A screenshot of a graph

Description automatically generated*

*Note:* The upper panel shows Suspect effects on adjustment. The lower right panel shows Preceding Context effects on adjustment. Blue horizontal lines indicate adjustment from the ground truth computation. Green lines with error bars indicate adjustment simulated by the delta model after fitting to participant ratings. Error bars are 95% confidence intervals.

*Participants adjust more to disconfirmatory claims*

Like Experiment 1, participants adjusted more in response to disconfirmatory claims than to confirmatory ones. Indeed, there was a main effect of Preceding Context on participants’ adjustment *F*(1,103) = 32.83, *p* < .001, *ηp2* = .24, with a non-significant Preceding Context × Claim interaction *F*(1,103) = 3.87, *p* = .05, *ηp2* = .04. A disconfirmatory evidence bias was confirmed by a significant effect of Proportion Guilt Context in a linear mixed model with participant as a random effect. The negative slope of -5.27 suggests that participants’ adjustments became more innocent-deflected, the more guilty was the context.

While the ground truth computation in Experiment 1 showed little sensitivity to disconfirmatory versus confirmatory evidence samples (Figure 2), the ground truth computation in Experiment 2 showed a small trend towards greater sensitivity to disconfirmatory evidence. This difference appears accounted for by the greater probability of majority claim *q* (i.e., guilt / innocence split) used in Experiment 2 (.7/.3) compared to Experiment 1 (.6/.4). For this reason, we tested whether participants’ perceived *q* might account for their sensitivity to disconfirmatory evidence by building a “Split” model where *q* was a free parameter, as we report in the next section.

*The delta model best explains participants’ behaviour*

The delta model shows significantly less (better) BIC compared to the other models (Figure 5). Moreover, the delta model appears to accurately approximate participants’ probability ratings (Figure 3) and adjustment (Figure 4), including both the effects of Suspect and the higher sensitivity to disconfirmatory evidence. The delta model’s mean prior probability for male guilt (.53) was significantly greater than the mean prior probability for female guilt (.48), *t*(103) = 2.88, *P* = .005. The average learning rate was .14 and the average beta was 242.36.

**Figure 5**

*Model comparison*

A graph of data with black dots

Description automatically generated with medium confidence

*Note:* Delta model fitted parameter values, with points representing individual participant parameter estimates and bars representing mean values. Horizontal lines connect bars if mean values differ significantly after Bonferroni correction for the six pairs of mean values.

**Discussion**

We hypothesised that we would observe differential prior beliefs about male and female suspects. Participants indeed rated male suspects as more probably guilty than female suspects. This occurred before any witness claims were experienced and to some extent at other sequence positions. Our findings of a gender bias on Probability and Adjustment in Experiment 2 stand in contrast to those of Experiment 1, in which we did not detect evidence for differential probability rating of guilt directed at atheists versus Christians. More research, including a replication, is needed to determine whether this occurred because participants hold stronger prior beliefs about gender than about religiosity, or because of methodological issues. Indeed, Experiment 2 introduced a more sensitive within-participants manipulation of Suspect and conveyed the Suspect information via images visible throughout the sequence rather than text visible only at the beginning of sequences. If further research vindicates visual image presentation as key for inducing differential guilt beliefs about demographic categories, then this may have important implications for presentation of evidence about suspects using images in real-world forensic settings. We also hypothesised that we would replicate the enhanced sensitivity of participants to disconfirmatory evidence. Indeed, participants were swayed more by “oddball” witness claims that were inconsistent with the context of the rest of the sequence.

Both effects of Suspect and Context appear explainable by a heuristic delta rule, where participants keep a running estimate of the probability of guilt, updated by the prediction error arising from each evidence sample. This model outperformed other models that were derived from the Bayesian condition probability computation.

**Experiment 3**

Real world contexts such as forensic scenarios do not often involve explicit probability rating, as in Experiments 1 and 2. The draws to decision task better resembles deliberative judicial procedures, in which evidence from witnesses is used to motivate guilty or innocent verdicts. Here, we tested whether participants make differential decisions about guilt to some suspects, but using the draws to decision task to test how much evidence is needed to render guilty or innocent verdicts, as well as the accuracy of such verdicts. Participants posed as a jury, which must choose how many witness claims to review before deciding on a verdict of guilty or innocent. We manipulated whether the defendant was healthy or had a diagnosis of schizophrenia. We hypothesised that the stereotype that individuals with schizophrenia can be dangerous might lead participants to sample less evidence before rendering guilty verdicts for such defendants than for defendants without a diagnosis of schizophrenia.

**Method**

*Participants*

We enrolled248 participants via dissemination of online links on social media (e.g., Facebook). Participants used a computer or tablet of choice. Participants answered questions about age, gender, ethnicity, highest level of education achieved and whether they had come into contact with the criminal justice system. Ages ranged from 18 to 63 years (*M* = 25.32 yo, *SD* = 5.83; 186 women, 61 men, one replied “prefer not to say”). Informed consent was obtained by local ethics committee approval.

*Procedures*

Gorilla Experiment Builder (Anwyl-Irvine, et al., 2020) hosted the study. Participants were first informed they were mock jurors who were evaluating witness claims so they could render a Guilt or Innocent verdict on the suspect (See Supplementary Materials). Our main hypothesis concerns effects of suspect mental health conditions on participants’ draws to decision. So that we could compute the number of draws to decision associated with the ground truth facts of the paradigm, we quantified choice costs. Before beginning each witness sequence, participants read: “*Each time you choose to see another witness it will cost the court £10*” and “*If you choose the wrong verdict it will cost the court £1000. Try to minimise the cost*”.

To proceed, participants needed to correctly pass an instructions comprehension test (Balzan, et al., 2011; Balzan et al., 2012). They answered the following multiple-choice questions, with options shown in brackets with correct option in bold. Text was not bold in study. “*How many witness statements will you be able to see during each task?*” (5, **10**, 15, 20), “*The probability of seeing a guilty statement is either 60% or 40%*” (**True**, false), “*How much will each witness statement cost the court?*” (**£10**, £100, £1000, £10,000) and “*How much will the wrong verdict cost the court?*” (£10, £100, **£1000**, £10,000). After each answer, green ticks indicated correct answers and red crosses indicated wrong answers. For wrong answers, participants could choose a different answer. Choice of the correct answer advanced the study to the next question or task.

Next, participants proceeded to the two witness sequences. Each began with a crime scenario (Supplementary Materials). Participants were randomly assigned a mental health condition (health or schizophrenia diagnosis), which was the same in both of a given participant’s scenarios. Therefore, the scenarios read either “*The defendant is a white British male in his early 30s and has no mental health conditions*” or “*The defendant is a white British male in his early 30s who is receiving treatment for schizophrenia. A clinical psychologist has provided some of the characteristics of schizophrenia in the top left corner*”. In the schizophrenia condition, this screen displayed the list of “*Schizophrenia characteristics: Hallucinations, delusions, social withdrawal (asociality), lack of emotional and facial expression (affective flattening)*”.

We used two fixed orders of witness claims, which were counterbalanced across the two sequences. In these orders, two or three draws were from the majority claim, followed by the first disconfirmatory witness claim. Every sequence was randomly assigned to be a guilt sequence (GGIGIIGGIG or GGGIIGGIGI) or an innocent sequence (IIGIGGIIGI or IIIGGIIGIG). Guilt and innocent verdicts were scored as correct for guilt and innocent sequences respectively. Before viewing a witness claim, participants viewed coins and “*cost of next witness: £10*. Witness claims appeared as blank avatars as in Experiment 1. Visible during the witness claims were defendant demographics (top-left of screen), previously-shown witness claims (top right) and numbers of witness claims remaining (bottom right).

Participants who chose “*more information*”, then viewed the cost to sample screen and another witness claim. Participant who chose “*verdict*” were told “*please choose a verdict*” of “*guilty*” or “*innocent*”. When participants reached the final witness claim, they viewed all previous witness claims and two response buttons labelled “*guilty*” and “*innocent*”, with no option to sample again. After verdict choice, a feedback screen displayed whether their chosen verdict was correct or incorrect and total costs incurred by the court.

After the feedback screens, participants rated on a ten-point scale from 1 (Not at all) to 10 (Entirely): “*How certain were you about your decision before you saw the feedback screen?*” (i.e., Confidence); and on a five-point scale from 1 (Not at all) to 5 (Entirely) the following questions: “*How responsible do you believe the defendant was for his behaviour?*” (i.e., Responsibility), “*Do you think the defendant is a danger to others?*” (i.e., Danger Others) and “*Do you think the defendant is a danger to himself?*” (i.e., Danger Self). Participants also completed the 21-item Peters et al. Delusions Inventory (PDI; Peters, Joseph, Day & Garety, 2004), a questionnaire that assesses the delusion-proneness trait in the general population. Exploratory analyses of these measures is found in the Supplementary Materials.

*Design and Analyses*

Data and analysis code are available at <https://github.com/nicholasfurl/Forensic-beads-paper-1>. We conducted analyses using JASP version 0.14 and MATLAB R2022b. One set of analyses used three dependent variables: *Draws to decision*, *Accuracy* (proportion correct verdicts) and *Court Costs* (in GBP). For each dependent variable, a linear mixed model tested effects of the between-participants factor *Defendant* (Health or Schizophrenia), treating *Participant* and *Guilt Sequence* (Guilty or innocent sequences) as nuisance random effects.

We also tested how much evidence participants used to render a guilty verdict and whether this varied by mental health state of the defendant. We employed a generalised linear mixed model with a logit link function to predict values of the binary dependent variable *Verdict, using Draws to Decision*, *Defendant* and their interaction as fixed effects and *Participant* and *Guilt Sequence* as nuisance random effects.

**Results**

As shown in Figure 3, we found null effects of our main factor of interest Defendant on Draws to Decision *F*(1,4.83) = 0.02, *p =* .90, Accuracy *F*(1,1.74) = 0.02, *p =* .90 and Court Costs *F*(1,1.66) = 0.02, *p =* .90. When we attempted to predict guilty verdicts, we found null effects for the Draws to Decision × Defendant interaction *Χ2*(1) = .80, *p* = .37 and for the main effects of Draws to Decision *Χ2*(1) = .38, *p* = .54 and Defendant *Χ2*(1) = .38, *p* = .54.

**Figure 3.**

*Performance on optimal stopping version of forensic-based beads task where defendant either has schizophrenia or no mental health diagnosis.*

A group of black and white bars

Description automatically generated with medium confidence

*Note.*  Error bars indicate 95% confidence intervals.

**Discussion**

Using a draws to decision task instead of a graded estimates task, we could not find statistical evidence for the hypothesis that there is prejudice based on a suspect’s schizophrenia diagnosis when weighing evidence about fictitious courtroom defendants. We tentatively recommend probability estimates on the graded estimates task as the best measure for detecting prejudice, compared to draws to decision. However, future studies need to test whether suspect gender effects also manifest on the draws to decision measure.

Our findings here have clear implications for the types of criminal justice settings (e.g., police investigations, courtrooms) we staged in our paradigms, in which potentially decision makers with prior beliefs about demographic groups must evaluate samples of unreliable evidence. Our paradigm provides a proof of concept for how influences of prior belief in such contexts, if they exist, could be measured.

**General Discussion**

Experiments 1 and 3 respectively did not find evidence that participants differentially rate guilt for different suspect categories. In contrast, in Experiment 2, before viewing any witness claim, participants believed that male suspects were more guilty than female suspects. At the same time, heightened sensitivity toversus confirmatory Computational modelling from Experiment 2 showed that a heuristic that learns by prediction error and employs a suspect-specific prior belief about guilt could reproduce participants’ behaviour and better fit participant behaviour than three alternative models, designed based on the Bayesian conditional probability formula. These findings demonstrate that our newly-developed forensic-orientated beads task, along with its associated modelling techniques, can be used to detect and explain participant strategies in realistic forensic scenarios.

We hypothesise that disconfirmatory evidence bias could be a generalisation to the general population of phenomena observed in psychosis, where affected individuals especially adjust beliefs when evidence is disconfirmatory (e.g., Adams, et al., 2018). it is tempting to speculate that disconfirmatory evidence bias and jumping to conclusions in the draws to decision task are two sides to the same coin. Indeed, it is already well established that participants in the general population, without psychosis diagnosis, sample fewer beads than mathematically optimal (Furl & Averbeck, 2011; Hauser et al., 2017; Hauser, et al., 2018; Sonnemans and van Dijk, 2012; van der Leer, et al., 2015). If these biases arise from a common mechanism, then perhaps it relates to the heuristic probability calculation we identified in Experiment 2. We further hypothesise that variability in the general population of traits may mediate or be directly responsible for one or both biases. Jumping to conclusions, in some form, is related to not just to delusion-proneness in the general population (Henquet et al., 2020) but also to intelligence (Tripoli et al., 2021) and analytic cognitive style (Pytlik et al., 2021). Less is known about individual differences in sensitivity to disconfirmatory evidence in the graded estimates task or whether it follows similar patterns as jumping to conclusions.

We note some limitations. We tried to vary our scenarios, using a range of crimes. However, scenario is a variable that could be manipulated experimentally in many ways in future studies. For example, had we used crimes stereotypical for male or female perpetrators, we hypothesise that participants’ prior beliefs about men or women may have had differential influence. There are other ways the paradigm might be developed to explore more real-world elements. For example, real witnesses would rarely simply declare a suspect as innocent or guilty without providing more concrete details. Other attempts at framing realistic beads tasks offered detailed statements as evidence (Westermann et al., 2012). Further, despite the fine experimental control the beads task gives us, simulated forensic scenarios are not “real”. We believe that this paradigm and its associated modelling techniques might be further developed for fieldwork in forensic settings, where the biases we show hypothetically should manifest. We also did not always fully randomise witness claim order. Many studies of the classic beads task used only one sequence with the same evidence order for every participant, and this fixed sequence was often recycled from previous work (Howe et al., 2018; Lincoln et al., 2011; Ross et al., 2016; Pytlik et al., 2021; Tripoli et al., 2021), which limits the generalisability of their results. When studies use only one sequence in the graded estimates task (Howe at al., 2018; Moritz et al., 2012), the disconfirmatory nature of the evidence sample can sometimes be confounded with sequence position. Although we used a limited number of fixed-order sequences in Experiments 2 and 3, we recommend using large numbers of random sequences as in Experiment 1. In Experiment 1, we illustrate methods suitable for measuring the preceding context and thereby identifying disconfirmatory evidence samples across the positions of randomised sequences.

Beads tasks are subsumed by a broader class of paradigms, which concern how individuals make decisions based on evidence, and which could be used to simulate real-world (forensic) situations, as we did here. Many can be theorised about within the “accumulation-to-bound” framework (Forstmann et al., 2016). Here, exposure to evidence adds noisy increments to a decision variable in the possible direction of at least one “bound” or decision threshold. Eventually, this evidence accumulation will exceed one of these decision thresholds and thereby trigger a response. The beads task offers some advantages when working within this framework, as it partitions evidence into discrete samples, allowing experimenter control of the accumulation process (Globig et al., 2021). It also allows designs such as the graded estimates task, which quantifies the influence of each individual evidence sample. However, the accumulation-to-bound framework is commonly applied to perceptual decision making (O’Connell et al., 2021; Shadlen & Kiani, 2013), using the random dot motion or similar task (e.g., Drugowitz et al., 2012). Here, the drift diffusion modelling framework is often applied to decompose reaction times and choices into drift rate (speed of accumulation) and decision bound (how much evidence is needed to respond). Perceptual decisions are also prevalent in the real-world contexts and can have serious consequences, depending on whether the individual acts now or waits for more evidence. The speed accuracy trade-off, for example, is ubiquitous among reaction time tasks, and involves decisions about when to stop sampling evidence (Ratcliff et al., 2015). We propose that one might frame perceptual decision-making tasks as realistic (e.g., forensic) decision making scenarios and apply associated models, like the drift diffusion model. Some work has already been done in this area detecting prejudicial bias in evidence accumulation using a shooter task (Plesak, et al., 2018).

There are more evidence-evaluation tasks that might fruitfully be applied to discover decision strategies in realistic (e.g., forensic) contexts. In one task (Coenen & Gureckis, 2016), participants do not know the proportion (e.g., 60%) of the majority bead colour in the hidden jar and (as with the draws to decision task) they sample evidence until they either attempt to name the majority colour or estimate the majority percentage. In the “ultimate sampling dilemma” (Fiedler, 2008), positive and negative samples (e.g., customer reviews) are drawn from two or more sources (e.g., companies / service providers) and participants sample until they can name the more positive source (e.g., the company reviewers liked better). In a similar “expanded judgement” paradigm, participants play a video game where they sample noisy (70% accurate) cues about which of two paths leads to a reward (Malhotra et al., 2017). In yet another task, participants sample oriented Gabor gratings, attempting to infer from which of two or more distributions (with different mean grating orientations) the gratings derive (Tickle et al., 2020). Last, “best-choice problems” ask participants to sample choice new options and attempt to stop and choose the best possible option (Costa & Averbeck, 2015; Furl et al., 2019; van de Wouw, et al., 2022). These tasks are nearly always framed using unrealistic scenarios and any of them could be reframed to detect potential real-world decision bias, as we have done here for the beads task. Some of these tasks come with associated computational frameworks, which can be used for theoretically explaining the strategies participants use in forensic contexts.

One specific task that might be of special interest when generalising experimental tasks to forensic contexts is the “Bias against disconfirmatory evidence” task or BADE task (Woodward, Moritz, Cuttler & Whitman, 2006; Eisenacher & Zink, 2017). 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. More deluded individuals tend to cling to the lure during these later cues, compared to less deluded individuals, a phenomenon known as BADE. BADE is often assumed to embody the rigidity associated with delusions and insensitivity to new evidence evinced by the delusion-prone. In contrast, “jumping to conclusions” on the beads task is often explained by recourse to the idea that the delusion prone adopt bizarre beliefs too easily due to oversensitivity to evidence (or oversensitivity to disconfirmatory evidence, in the case of the graded estimates version). Findings of both strategies are well-documented and the problem of how to resolve them remains inconclusive in clinical research (Furl et al., 2024).

The BADE also loosely resembles the types of primacy effects observed in forensic contexts, in which judgements are unduly influenced by the first evidence samples presented in a sequence (Dror & Kukucka, 2021). We explicitly modelled participants’ probability judgements using our primacy model, in which guilt judgements depend mainly on early sample samples, only yet the delta heuristic was found to better fit participants in our study, using the graded estimates beads task. Under the delta heuristic, larger belief adjustments would occur when prediction error is largest. This makes the delta heuristic quite flexible: Large belief adjustment could occur early in sequences insofar as those evidence samples are not predictable from participants’ prior beliefs. And relatively large belief adjustment could occur later in sequences whenever evidence is disconfirmatory. This type of prediction error learning is also biologically plausible, with a venerable association with dopaminergic reward learning (See Furl et al., 2024 for a review in the context of jumping to conclusions and psychosis).

The potentially context-dependent nature of sensitivity to evidence makes it all the more urgent to better understand the statistical nature of evidence sequences in real-world forensic contexts. To what extent does the use of heuristics leads to accurate performance in such settings? Is it always adaptive or can it lead to systematic errors? The current paper is among the first to examine responses to confirmatory and disconfirmatory evidence in the context of the general population and more work needs to be done in this population, and with forensic scenarios in general, to determine when one or the other strategy might manifest.

**References**

Adams, R.A., Napier, G., Roiser, J.P., Mathys, C. & Gilleen, J. (2018). Attractor-like Dynamics in Belief Updating in Schizophrenia. *Journal of Neuroscience, 38,* 9471-9485. <https://doi.org/10.1523/JNEUROSCI.3163-17.2018>

Ahola, A. S., Christianson, S. Å., & Hellström, Å. (2009). Justice needs a blindfold: Effects of gender and attractiveness on prison sentences and attributions of personal characteristics in a judicial process. *Psychiatry, Psychology and Law, 16(Suppl 1),* S90–S100. <https://doi.org/10.1080/13218710802242011>

Anwyl-Irvine, A.L., Massonnié J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52,* 388-407. <https://doi.org/10.3758/s13428-019-01237-x>

Averbeck, B.B. (2015). Theory of choice in bandit, information sampling and foraging tasks. *PLoS Computational Biology, 11,* e1004164. <https://doi.org/10.1371/journal.pcbi.1004164>

Baker, S.C., Konova, A.B., Daw, N.D. & Horga, G. (2019). A distinct inferential mechanism for delusions in schizophrenia. *Brain, 142,* 1797-1812. <https://doi.org/10.1093/brain/awz051>

Balzan, R., Delfabbro, P. & Galletly, C. (2011). Delusion‐proneness or miscomprehension? A re‐examination of the jumping‐to‐conclusions bias. *Australian Journal of Psychology, 64,* 101-107.  <https://doi.org/10.1111/j.1742-9536.2011.00032.x>

Balzan, R. P., Delfabbro, P. H., Galletly, C. A., & Woodward, T. S. (2012). Over-adjustment or miscomprehension? A re-examination of the jumping to conclusions bias. *Australian and New Zealand Journal of Psychiatry, 46,* 532e540. <http://dx.doi.org/10.1177/0004867411435291>

Balzan, R.P., Ephraums, R., Delfabbro, P. & Andreou, C. (2017). Beads task vs. box task: The specificity of the jumping to conclusions bias. *Journal of Behavior Therapy and Experimental Psychiatry, 56,* 42-50. <https://doi.org/10.1016/j.jbtep.2016.07.017>

Broome, M. R., et al. (2007). Delusion formation and reasoning biases in those at clinical high risk for psychosis. *The British Journal of Psychiatry, 191,* s38–s42. <https://doi.org/10.1192/bjp.191.51.s38>

Coenen, A., & Gureckis, T. M. (2016). The distorting effect of deciding to stop sampling. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2819–2824). Austin, TX: Cognitive Science Society.

Colbert, S. M., & Peters, E. R. (2002). Need for Closure and Jumping-to-Conclusions in Delusion-Prone Individuals. *The Journal of Nervous and Mental Disease, 190,* 27–31. <https://doi.org/10.1097/00005053-200201000-00007>.

Costa, V.D. & Averbeck, B.B. (2015). Frontal-parietal and limbic-striatal activity underlies information sampling in the best choice problem. *Cerebral Cortex, 25,* 972-982. <https://doi.org/10.1093/cercor/bht286>

DeBruine, L., & Jones, B. C. (2020). *3DSK face set with webmorph templates.* <https://doi.org/10.17605/OSF.IO/A3947>

Devine, D. J., & Caughlin, D. E. (2014). Do they matter? A meta-analytic investigation of individual characteristics and guilt judgments*. Psychology, Public Policy, and Law, 20,* 109–134. <https://doi.org/10.1037/law0000006>

Dror, I.E. (2013). Practical solutions to cognitive and human factor challenges in forensic science. *Forensic Science Policy & Management: An International Journal, 4,* 105-113. <https://doi.org/10.1080/19409044.2014.901437>

Dror, I.E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical Chemistry, 92,* 7998-8004. [https://doi.org/10.1021/acs.analchem.0c00704](https://doi.org/10.1021/acs.analchem.0c00704" \o "DOI URL)

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. <https://doi.org/10.1016/j.fsisyn.2021.100161>.

Drugowitsch, J., Moreno-Bote, R., Churchland, A.K., Shadlen, M.N. & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience, 32,* 3612-3628. <https://doi.org/10.1523/JNEUROSCI.4010-11.2012>

Dudley, R. E., John, C. H., Young, A. W., & Over, D. E. (1997). The effect of self- referent material on the reasoning of people with delusions. *British Journal of Clinical Psychology, 36,* 575-584. <http://dx.doi.org/10.1111/j.2044-8260.1997.tb01262.x>.

Dudley, R., Taylor, P., Wickham, S. & Sutton, P. (2016). Psychosis, Delusions and the "Jumping to Conclusions" Reasoning Bias: A Systematic Review and Meta-analysis. *Schizophrenia Bulletin, 42,* 652-65. <https://doi.org/10.1093/schbul/sbv150>

Eisenacher, S. & Zink, M. (2017). Holding on to false beliefs: The bias against disconfirmatory evidence over the course of psychosis. *Journal of Behavior Therapy and Experimental Psychiatry, 56,* 79-89. <https://doi.org/10.1016/j.jbtep.2016.08.015>

Evans, S.L., Averbeck, B.B. & Furl, N. (2015). Jumping to conclusions in schizophrenia. *Neuropsychiatric Disease and Treatment, 11,* 1615-1624. <https://doi.org/10.2147/NDT.S56870>

Fear, C.F. & Healy, D. (1997). Probabilistic reasoning in obsessive-compulsive and delusional disorders*. Psychological Medicine, 27,* 199 –208. <https://doi.org/10.1017/S0033291796004175>

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(1), 186–203. <https://doi.org/10.1037/0278-7393.34.1.186>

Fine, C., Gardner, M., Craigie, J. & Gold, l. (2007). Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry, 12,* 46–77. <https://doi.org/10.1080/13546800600750597>

Forstmann, B.U., Ratcliff, R. & Wagenmakers, E.J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology, 67,* 641-666. <https://doi.org/10.1146/annurev-psych-122414-033645>

Fraser, J., Morrison, A. & Wells, A. (2006). Cognitive processes, reasoning biases and persecutory delusions: A comparative study. *Behavioural and Cognitive Psychotherapy, 34,* 421-435. <https://doi.org/10.1017/S1352465806002852>

Furl, N. & Averbeck, B. (2011). Parietal cortex and insula relate to evidence seeking relevant to reward-related decisions. *Journal of Neuroscience, 31,* 17572-82. <https://doi.org/10.1523/JNEUROSCI.4236-11.2011>

Furl, N., Averbeck, B.B. & McKay, R. (2019). Looking for Mr(s) Right: Decision bias can prevent us from finding the most attractive face. *Cognitive Psychology, 111,* 1-14. <https://doi.org/10.1016/j.cogpsych.2019.02.002>

Furl, N., Coltheart, M., & McKay, R. (2024). The paradox of delusions: Are deluded individuals resistant to evidence? In E. Sullivan-Bissett (Ed.), *Belief, Imagination, and Delusion* (pp. 240-265). Oxford University Press; Oxford. <https://doi.org/10.1093/oso/9780198872221.003.0012>

Garety, P.A., Hemsley, D.R. & Wessely, S. (1991). Reasoning in Deluded Schizophrenic and Paranoid Patients. T*he Journal of Nervous and Mental Disease, 179,* 194. <https://doi.ord/10.1097/00005053-199104000-00003>

Garety, P.A., et al., (2013). Neuropsychological functioning and jumping to conclusions in delusions. *Schizophrenia Research, 150,* 570-574. <https://doi.org/10.1016/j.schres.2013.08.035>

Gervais, W.M., Xygalatas, D., McKay, R., van Elk, M., Buchtel, E.E., Averyard, M., Schiavone, S.R., Dar-Nimrod, I., Svedholm-Häkkinen, A.M., Riekki, T., Klocová, K., Ramsay, J.E. & Bulbulia, J. (2017). Global evidence of extreme intuitive moral prejudice against atheists. *Nature Human Behaviour, 1,* 0151. <https://doi.org/10.1038/s41562-017-0151>

Globig, L.K., White, K., Feng, G. & Sharot, T. (2021). Under Threat, Weaker Evidence Is Required to Reach Undesirable Conclusions. Journal of Neuroscience, 41, 6502-6510. <https://doi.org/10.1523/JNEUROSCI.3194-20.2021>

Hauser, T.U., Moutoussis, M., NSPN Consortium, Dayan, P. & Dolan, R.J. (2017). Increased decision thresholds trigger extended information gathering across the compulsivity spectrum. *Translational Psychiatry, 7,* 1296. <https://doi.org/10.1038/s41398-017-0040-3>

Hauser, T.U., Moutoussis, M., Purg, N., Dayan, P. & Dolan R. (2018). Beta-blocker propranolol modulates urgency during sequential information gathering. *Journal of Neuroscience, 38,* 7170-7178. <https://doi.org/10.1523/JNEUROSCI.0192-18.2018>

Henquet, C. et al. (2020). A replication study of JTC bias, genetic liability for psychosis and delusional ideation. *Psychological Medicine* 1–7. <https://doi.org/10.1017/S0033291720003578>

Howe, J., Ross, R., McKay, R. & Balzan, R.P. (2018). How do delusion-prone individuals respond to disconfirmatory evidence? Improving comprehension of the beads task may help. *Zeitschrift für Psychologie, 226,* 182-190. <https://doi.org/10.1027/2151-2604/a000333>

Huq, S.F., Garety, P.A. & Hemsley, D.R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *Quarterly Journal of Experimental Psychology A, 40,* 801. <https://doi.org/10.1080/14640748808402300>

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. American Journal of Psychiatry, 160, 13-23. <https://doi.org/10.1176/appi.ajp.160.1.13>

Katthagen, T., et al. (2018). Modeling subjective relevance in schizophrenia and its relation to aberrant salience. PLoS Computational Biology, 14, e1006319. <https://doi.org/10.1371/journal.pcbi.1006319>

Langdon, R., Ward, P.B. & Coltheart, M. (2010). Reasoning anomalies associated with delusions in schizophrenia. *Schizophrenia Bulletin, 36,* 321–330. <https://doi.org/10.1093/schbul/sbn069>

Lincoln, T., Salzmann, S., Ziegler M. & Westermann S. (2011). When does jumping-to-conclusions reach its peak? The interaction of vulnerability and situation-characteristics in social reasoning. *Journal of Behavior Therapy and Experimental Psychiatry, 42,* 185-191. <https://doi.org/10.1016/j.jbtep.2010.09.005>

Lindholm, T., & Cederwall, J. Y. (2010). Ethnicity and gender biases in the courtroom. In P. A. Granhag (Ed.), *Forensic psychology in context: Nordic and international approaches* (pp. 228–246). Willan Publishing.

Lundqvist, D., Flykt, A. &  Öhman, A. (1998). *The Karolinska Directed Emotional Faces - KDEF*, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet. Retrieved from <https://kdef.se/>.

Malhotra, G., Leslie, D.S., Ludwig, C,J,H. & Bogacz, R. (2017). Overcoming indecision by changing the decision boundary. *Journal of Experimental Psychology: General, 140,* 776-803. <http://dx.doi.org/10.1037/xge0000286>

McKay, R., Langdon, R., & Coltheart, M. (2006). Need for closure, jumping to conclusions, and decisiveness in delusion-prone individuals. *Journal of Nervous and Mental Disease, 194,* 422–426. <https://doi.org/10.1097/01.nmd.0000221353.44132.25>

McLean, B.F., Mattiske, J.K. & Balzan, R.P. (2017). Association of the Jumping to Conclusions and Evidence Integration Biases With Delusions in Psychosis: A Detailed Meta-analysis. *Schizophrenia Bulletin, 43,* 344-354. <https://doi.org/10.1093/schbul/sbw056>

Menon, M., Pomarol‐Clotet, E., McKenna, P. J., & McCarthy, R. A. (2006). Probabilistic reasoning in schizophrenia: a comparison of the performance of deluded and nondeluded schizophrenic patients and exploration of possible cognitive underpinnings. *Cognitive Neuropsychiatry, 11,* 521-536. <https://doi.org/10.1080/13546800544000046>

Moritz, S., Van Quaquebeke, N., Lincoln, T. (2012). Jumping to conclusions is associated with paranoia but not general suspiciousness: A comparison of two versions of the probabilistic reasoning paradigm. *Schizophrenia Research and Treatment.* 384039 <https://doi.org/10.1155/2012/384039>

Moritz, S., Pfuhl, G., Ludtke, T., Menon, M., Balzan, R.P., & Andreou, C. (2017). A two-stage cognitive theory of the positive symptoms of psychosis. Highlighting the role of lowered decision thresholds. *Journal of Behavior Therapy and Experimental Psychiatry, 56,* 12–20. <https://doi.org/10.1016/j.jbtep.2016.07.004>

Moritz, S. & Woodward, T.S. (2005). Jumping to conclusions in delusional and non-delusional schizophrenic patients. *British Journal of Clinical Psychology, 44,* 193-207. <https://doi.org/10.1348/014466505X35678>

Moritz, S., Woodward, T.S. & Lambert, M. (2007). Under what circumstances do patients with schizophrenia jump to conclusions? A liberal acceptance account. *British Journal of Clinical Psychology, 46(Pt 2),* 127-37. <https://doi.org/10.1348/014466506X129862>

Moutoussis, M., Bentall, R.P., El-Deredy, W & Dolan, R. (2011). Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. *Cognitive Neuropsychiatry, 16,* 422-447. <https://doi.org/10.1080/13546805.2010.548678>

Norton, M. I., Sommers, S. R., Vandello, J. A., & Darley, J. M. (2006). Mixed motives and racial bias: The impact of legitimate and illegitimate criteria on decision making. *Psychology, Public Policy, and Law, 12,* 36–55. <https://doi.org/10.1037/1076-8971.12.1.36>

O'Connell, R.G. & Kelly, S.P. (2021). Neurophysiology of human perceptual decision-making. *Annual Review of Neuroscience, 44,* 495-516. <https://doi.org/10.1146/annurev-neuro-092019-100200>

Peters, E., Joseph, S., Day, S. & Garety, P. (2004). Measuring delusional ideation: the 21-item Peters et al. Delusions Inventory (PDI). *Schizophrenia Bulletin, 30,* 1005-1022. <https://doi.org/10.1093/oxfordjournals.schbul.a007116>

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72,* 346–354. <https://doi.org/10.1037/h0023653>

Plesak, T.J., Cesario, J. & Johnson, D.J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin and Review, 25,* 1301-1330. <https://doi.org/10.3758/s13423-017-1369-6>

Pytlik, N., Soll, D. & Mehl, S. (2021). Thinking preferences and conspiracy belief: Intuitive thinking and the jumping to conclusions bias as a basis for the belief in conspiracy theories. *Frontiers in Psychiatry, 11,* 568942. <https://doi.org/10.3389/fpsyt.2020.568942>

Ratcliff, R., Smith, P.L. & McKoon, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Current Directions in Psychological Science, 24,* 458-470. <https://doi.org/10.1177/0963721415596228>

Rogers, P & Davies, M. (2007). Perceptions of victims and perpetrators in a depicted child sexual abuse case: gender and age factors. *Journal of Interpersonal Violence, 22,* 566-584. <https://doi.org/10.1177/0886260506298827>

Romero-Ferreiro, V., Susi, R., Sánchez-Morla, E.M., Marí-Beffa, P., Rodríguez-Gómez, P., Amador, J., Moreno, E.M., Romero, C., Martínez-García, N. & Rodríguez-Jimenez, R. (2022). Bayesian reasoning with emotional material in patients with schizophrenia. *Frontiers in Psychology, 13,* 827037. <https://doi.org/10.3389/fpsyg.2022.827037>

Ross, R.M., McKay, R., Coltheart, M. & Langdon, R. (2015). Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. *Schizophrenia Bulletin, 41,* 1183-1191. <https://doi.org/10.1093/schbul/sbu187>

Ross, R. M., Pennycook, G., McKay, R., Gervais, W. M., Langdon, R., & Coltheart, M. (2016). Analytic cognitive style, not delusional ideation, predicts data gathering in a large beads task study. *Cognitive Neuropsychiatry, 21,* 300–314. <https://doi.org/10.1080/13546805.2016.1192025>

Rubio, J.L., et al., (2011). Jumping to conclusions in psychosis: a faulty appraisal. *Schizophrenia Research,* 133, 199–204. <https://doi.org/10.1016/j.schres.2011.08.008>

Shadlen, M.N. & Kiani, R. (2013). Decision making as a window on cognition. *Neuron, 80,* 791-806. <https://doi.org/10.1016/j.neuron.2013.10.047>

Speechley, W.J., Whitman, J.C. & Woodward, T.S. (2010). The contribution of hypersalience to the “jumping to conclusions” bias associated with delusions in schizophrenia. *Journal of Psychiatry and Neuroscience, 35,* 7–17. <https://doi.org/10.1503/jpn.090025>

Strube, W., et al. (2021). Unstable belief formation and slowed decision-making: Evidence that the jumping-to-conclusions bias in schizophrenia is not linked to impulsive decision-making. *Schizophrenia Bulletin,* sbab108. <https://doi.org/10.1093/schbul/sbab108>

So, S.H. & Kwok, N.T-k. (2015). Jumping to Conclusions Style along the Continuum of Delusions: Delusion-Prone Individuals Are Not Hastier in Decision Making than Healthy Individuals. *PLoS One, 10,* e0121347. <https://doi.org/10.1371/journal.pone.0121347>

So, S.H., Siu, N.Y., Wong, H.L., Chan, W. & Garety, P.A. (2016). ‘Jumping to conclusions’ data-gathering bias in psychosis and other psychiatric disorders—two meta-analyses of comparisons between patients and healthy individuals. *Clinical Psychology Review, 46,* 151–167. <https://doi.org/10.1016/j.cpr.2016.05.001>

Sonnemans, J. & van Dijk, F. (2012). Errors in judicial decisions: Experimental results. *The Journal of Law, Economics and Organization, 28,* 687-716. <https://doi.org/10.1093/jleo/ewq019>

Tickle, H., Tsetsos, K., Speekenbrink, M. & Summerfield, C. (2020). Human optional stopping in a heteroskedastic world. *PsyArxiv*. <https://doi.org/10.31234/osf.io/t7dn2>

Tripoli G et al (2021). Jumping to conclusions, general intelligence, and psychosis liability: findings from the multi-centre EU-GEI case-control study. *Psychological Medicine 51,* 623–633. <https://doi.org/10.1017/S003329171900357X>

Van Dael, F., Versmissen, D., Janssen, I., Myin-Germeys, I., van Os, J. & Krabbendam L. (2006). Data gathering: biased in psychosis? *Schizophrenia Bulletin, 32,* 341–351. <https://doi.org/10.1093/schbul/sbj021>

van de Wouw, D.S., McKay, R.T., Averbeck, B.B. & Furl, N. (2022). Explaining human sampling rates across different decision domains. Judgment and Decision Making, 17, 487-512.

van der Leer, L. & McKay, R. (2014). ‘Jumping to conclusions’ in delusion-prone participants: an experimental economics approach. *Cognitive Neuropsychiatry 19,* 257–267. <https://doi.org/10.1080/13546805.2013.861350>

van der Leer, L., Hartig, B., Goldmanis, M. & McKay, R. (2015). Delusion proneness and ‘jumping to conclusions’: relative and absolute effects. *Psychological Medicine, 45,* 1253–1262. <https://doi.org/10.1017/S0033291714002359>

Warman, D. M., Lysaker, P. H., Martin, J. M., Davis, L., & Haudenschield, S. L. (2007). Jumping to conclusions and the continuum of delusional beliefs. *Behavior Research and Therapy, 45,* 1255e1269. <https://doi.org/10.1016/j.brat.2006.09.002>

Warman, D. M., & Martin, J. M. (2006). Jumping to conclusions and delusion proneness: The impact of emotionally salient stimuli. *The Journal of Nervous and Mental Disease, 194,* 760. <http://dx.doi.org/10.1097/01.nmd.0000239907.83668.aa>

Westermann, S., Salzmann, S., Fuchs, X. & Lincoln, T. (2012). Introducing a social beads task. *Journal of Experimental Psychopathology, 3,* 594-611. <https://doi.org/10.5127/jep.017111>

Wilkinson, D.J. & Caulfield, L.S. (2017). Delusional ideation, cognitive processes and crime-based reasoning. *European Journal of Psychology, 13,* 503-518. <https://doi.org/10.5964/ejop.v13i3.1181>

White, L.O. & Mansell, W. (2009). Failing to ponder? Delusion-prone individuals rush to conclusions. *Clinical Psychology and Psychotherapy, 16,* 111–124. <https://doi.org/10.1002/cpp.607>

Woodward, T.S., Moritz, S., Cuttler, C. & Whitman, J.C. (2006). The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology, 28,* 605-617. <https://doi.org/10.1080/13803390590949511>

Young, H. F., & Bentall, R. P. (1997). Probabilistic reasoning in deluded, depressed and normal subjects: effects of task difficulty and meaningful versus non-meaningful material. *Psychological Medicine, 27,* 455–465. <https://doi.org/10.1017/S0033291796004540>

Yourstone,J., Lindholm, T., Grann, M. & Svenson, O. (2008). Evidence of gender bias in legal insanity evaluations: a case vignette study of clinicians, judges and students. *Nordic Journal of Psychiatry, 62,* 273-278. <https://doi.org/10.1080/08039480801963135>

Zawadzki, J. A., Woodward, T. S., Sokolowski, H. M., Boon, H. S., Wong, A. H. C., & Menon, M. (2012). Cognitive factors associated with subclinical delusional ideation in the general population*. Psychiatry Research, 197,* 345–349. <https://doi.org/10.1016/j.psychres.2012.01.004>