Rethinking delusions: a selective review of delusion research through a computational lens

Brandon K. Ashinoff [a b], Nicholas M. Singletary [a b], Seth C. Baker [a b], and Guillermo Horga [a b *]

[a] Department of Psychiatry, Columbia University, 1051 Riverside Drive, New York, NY, USA
[b] New York State Psychiatric Institute, 1051 Riverside Drive, New York, NY, USA

[*] Corresponding author: HorgaG@nyspi.columbia.edu

Abstract

Delusions are rigid beliefs held with high certainty despite contradictory evidence. Notwithstanding decades of research, we still have a limited understanding of the computational and neurobiological alterations giving rise to delusions. In this review, we highlight a selection of recent work in computational psychiatry aimed at developing quantitative models of inference and its alterations, with the goal of providing an explanatory account for the form of delusional beliefs in psychosis. First, we assess and evaluate the experimental paradigms most often used to study inferential alterations in delusions. Based on our review of the literature and theoretical considerations, we contend that classic draws-to-decision paradigms are not well-suited to isolate inferential processes, further arguing that the commonly cited 'jumping-to-conclusion' bias may reflect neither delusion-specific nor inferential alterations. Second, we discuss several enhancements to standard paradigms that show promise in more effectively isolating inferential processes and delusion-related alterations therein. We further draw on our recent work to build an argument for a specific failure mode for delusions consisting of prior overweighting in high-level causal inferences about partially observable hidden states. Finally, we assess plausible neurobiological implementations for this candidate failure mode of delusional beliefs and outline promising future directions in this area.

Delusions are classically defined as false beliefs held with high certainty despite contradictory evidence. They are one of two defining symptoms of schizophrenia, the other being hallucinations. Delusions typically accompany schizophrenia and are common in other psychotic disorders, often producing immense disruption in the lives of the patients who suffer from them (Heinze et al., 2018; Upthegrove, 2018).

In one famous example, a bright and well-regarded young mathematician became increasingly convinced that he had the unique ability to decipher a secret code embedded in newspapers. He gradually developed an unyielding belief that solving this code was necessary to save humanity and that a vast conspiracy had formed to stop him. Ultimately, this belief consumed much of his life, in spite of persistent efforts from relatives, friends, and others to convince him that his belief was unfounded. Afraid for his life, he left behind his job, family, and country (Nasar, 1998).

This case illustrates the tragic, real-life consequences of delusional beliefs as well as their classic features: falsity, certainty, and rigidity. Of these, the necessity of belief falsity for the operationalization of delusions was questioned from its conception by Karl Jaspers (Jaspers, 1913), who emphasized the clinical value of the *form* over the content of psychotic experiences such as delusions. Jaspers made this point describing the memorable case of a delusion of jealousy in which the patient's partner was actually unfaithful. Difficulties ascertaining belief falsity are now broadly recognized to limit its clinical value. Additionally, challenges associated with the interpretation of beliefs in different cultural or experiential contexts, which are also key determinants of delusional themes, further call the definitional value of delusion content into question (Aschebrock et al., 2003; Gaines, 1995; Gold and Gold, 2012; Spitzer, 1990; Stompe et al., 2003). The variability and intractability of belief content is reflected by current operationalizations of delusions, which exclusively focus on belief form. The DSM-5 defines delusions as: "*fixed beliefs that are not amenable to change* in light of conflicting evidence […]. The distinction between a delusion and a strongly held idea […] depends in part on *the degree of conviction with which the belief is held* despite clear or reasonable contradictory evidence regarding its veracity" [italics added by authors; (American Psychological Association, 2013)]. Therefore, two essential formal features are necessary for beliefs to be considered delusional: (1) high subjective certainty (i.e., beliefs held with high conviction) and (2) belief rigidity (i.e., fixed beliefs resistant to change).

In this review, we will highlight recent work in computational psychiatry aimed at developing quantitative inference models describing the form of delusional beliefs in psychotic disorders, with special attention to those that might capture their two core features—high certainty and rigidity. Other reviews provide a broader review of the neurocognitive literature on delusions (Corlett et al., 2010). Here, we focus more narrowly on inference for two reasons. First, it bears historical relevance to the definition of delusions; e.g., the DSM-III defined delusion as "a false personal belief based on *incorrect inference* about external reality […]" [italics added by the authors; APA, 1980 (American Psychological Association, 1980)]. Second, and more importantly, inferential models deal with the formation of beliefs on the basis of observed evidence and past knowledge, a process that has been long theorized to be central to the genesis of delusions and one that is experimentally tractable. To begin, we first describe the mathematical foundations for models of inference.

A primer on Bayesian inference

Inference is generally defined as a method of logical reasoning in which one draws conclusions based on a set of premises. In abductive inference, a particular type of inference presumed to be relevant to delusions, one produces a best-guess explanation for a phenomenon based on available information (Coltheart et al., 2010). Statistically, inference similarly refers to the estimation of the amount of evidence in support of an explanatory hypothesis based on samples of information.

Bayesian inference is a method for probabilistic computation that optimally combines prior knowledge with new information. The resulting estimates are statistically optimal in that, on average, they maximize prediction accuracy. Estimates in Bayesian inference are framed in probabilistic terms as *beliefs* reflecting the intuited probabilities of different hypotheses under consideration, which are updated through the incorporation of new samples of information. This process of *belief updating* is summarized in Bayes' theorem (Eq. 1). Here, the *prior* belief represents previously acquired knowledge, the *likelihood* refers to the evidence provided by a new piece of information, and the *posterior* belief refers to the new or updated belief. In this formula, the posterior belief, $P(A|s)$, the probability of hypothesis A after observing a sample of information s, is estimated as a function of the prior belief, $P(A)$, or the probability of hypothesis A before observing s, and the likelihood, $P(s|A)$, the probability of s if hypothesis A were true (the strength of the evidence of sample s in support of hypothesis A), divided by a normalization factor.

$$P(A|s) = \frac{P(A) \cdot P(s|A)}{P(s)} = \frac{P(A) \cdot P(s|A)}{\big(P(A) \cdot P(s|A)\big) + \big(P(B) \cdot P(s|B)\big)} \qquad \text{Eq. 1}$$

To illustrate the intuition behind this equation, consider a hypothetical scenario where John, unable to find an important document he saved in a shared computer, suspects that a co-worker may have intentionally deleted it to sabotage his work. John knows of previous similar events in their company, which promotes fierce competition between co-workers. Given this document loss (s), should John conclude his co-worker intentionally sabotaged him (hypothesis A) or that it was an accident (hypothesis B)? Based on his prior knowledge, John considers the *a priori* probability of a co-worker trying to sabotage him $[P(A)]$ to be moderately low, about 0.2. But his meticulous bookkeeping makes this document loss a very rare event, so he considers it strong evidence for sabotage, with a likelihood $[P(s|A)]$ of about 0.75. Applying Bayes' theorem to optimally combine the prior beliefs $[P(A) = 0.2;\ P(B) = 0.8]$ and likelihoods $[P(s|A) = 0.75;\ P(s|B) = 0.25]$ would lead John to reach the posterior belief that the probability he was sabotaged is: $P(A|s) = (0.2\ \cdot\ 0.75)/\big((0.2\ \cdot\ 0.75) + (0.8 \cdot 0.25)\big) = 0.43$.

Bayesian inference over two complementary hypotheses can be reframed as the computation of their log odds (Eq. 2), rather than in terms of the raw probabilities. A formulation of Bayes' theorem in this *logit* space (Eq. 3) shows that inference reduces to an additive process, akin to that observed in the activity of neuronal populations involved in perceptual decisions (Gold and Shadlen, 2007).

$$log\left(\frac{P(A|s)}{P(B|s)}\right) = \ log\left(\frac{P(A)}{P(B)}\right)\ +\ log\left(\frac{P(s|A)}{P(s|B)}\right) \qquad \text{Eq. 2}$$

$$logit(posterior_A) = \ logit\ (prior_A)\ +\ logit\ (likelihood_A) \qquad \text{Eq. 3}$$

$$logit(posterior_A) = \ \omega_1 \cdot logit\ (prior_A)\ +\ \omega_2 \cdot logit\ (likelihood_A) \qquad \text{Eq. 4}$$

Parameterizing this *logit* formulation via a prior weight ω_1 and a likelihood weight ω_2 (weighted Bayesian model; Eq. 4) makes apparent that the Bayesian recipe for optimally combining prior beliefs and likelihoods consists of giving them an equal weight of 1 ($\omega_1 = \omega_2 = 1$). This common parameterization (Ambuehl and Li, 2018; Benjamin et al., 2019) also conveniently captures specific classes of deviations from optimality, since either the prior or the likelihood terms could theoretically be over- or under-weighted with respect to the ideal Bayesian benchmark. In the example above, for instance, John could have partially discounted his prior knowledge ($\omega_1 < 1$), which would have led him to erroneously overestimate the posterior probability that he was being sabotaged (e.g., an $\omega_1 = 0.5$ would produce a posterior belief $P(A|s) = 0.60$ for sabotage).

In sum, Bayesian inference can be used as a formal framework to quantify inference in terms of probabilistic beliefs. Critically, this framework provides an objective benchmark that empirical data can be measured against in order to examine deviations from optimality and interindividual variability in different elements of the inference process.

Brief summary of inferential theories of delusions

Although the general notion that delusions stem from alterations in reasoning was inherent to early clinical conceptualizations, it was Hemsley and Garety who proposed framing delusional beliefs as deviations in specific aspects of optimal Bayesian inference (Hemsley and Garety, 1986). They did not hypothesize a single alteration at the core of delusion formation and maintenance. Rather, they catalogued a bounty of potential deviations at the level of the different variables comprising the Bayesian algorithm that, mostly based on clinical intuition, could be reasonable candidates for explaining some aspects of delusional ideation. Their seminal proposal built on prior work (Fischhoff and Beyth-Marom, 1983) which similarly catalogued deviations from optimal inference as candidate mechanisms for explaining a variety of biases in judgment and decision-making that are commonly observed in the general, healthy population. They argued that variations in these biases could explain the characteristic resistance of delusional beliefs to disconfirmatory evidence, or their *rigidity*, as well as the characteristic *high certainty* with which the beliefs are held. Among the list of possible deviations that Hemsley and Garety (1986) considered was an alteration in the weighting of prior beliefs—captured by the parameter ω_1 in Eq. 4—noting that "deluded patients frequently tell interviewers that they have never considered the possibility of the falsity of their beliefs." As another candidate, they suggested a 'confirmation bias' whereby beliefs might be more responsive to new information consistent with prior beliefs relative to information inconsistent with them (or, equivalently, disproportionate weighting of the numerator in the likelihood ratio in Eq. 2 if A corresponded to the more likely *a priori* hypothesis). By focusing on deviations in specific parameters weighting the variables comprising a relevant algorithm and linking them to clinical phenomena, a concept commonly termed 'failure modes' in the burgeoning field of computational psychiatry (Redish et al., 2008; Walters and Redish, 2018), this work provided an influential framework for understanding delusions in terms of concrete alterations in Bayesian inference.

Crucially, the notion of delusion-related alterations in inference does not imply that healthy individuals are unbiased Bayesians (e.g., exhibiting $\omega_1 = \omega_2 = 1$) and only delusional patients exhibit some distinct biases (e.g., $\omega_1 \neq \omega_2 \neq 1$). That is, "normal" inference in the healthy population does not necessarily correspond to optimal inference. Indeed, this notion built upon research showing common biases among healthy individuals that suggest deviations from optimal Bayesian inference (Fischhoff and Beyth-Marom, 1983), including the underweighting of prior information ($\omega_1 < 1$; (Bar-Hillel, 1980; Benjamin, 2019; Kahneman and Tversky, 1973)) and distortions in the incorporation of likelihoods (Gonzalez and Wu, 1999). Hemsley and Garety instead adopted a more dimensional view under which delusions could be driven by quantitative differences in the same kinds of deviations from optimality exhibited by healthy individuals (Hemsley and Garety, 1986).

Motivated by the known hierarchical organization of the brain and the hierarchical nesting of information in the environment, modern theories of information processing in the brain tend to conceptualize inference as a hierarchical process. Accordingly, modern theories of delusions focus on alterations in hierarchical inference (Adams et al., 2013; Fletcher and Frith, 2009; Friston, 2008; Sterzer et al., 2018). Hierarchical-inference models comprise multiple, interdependent levels of processing, with lower levels supporting inferences on less abstract processes, like perception of the low-level features of sensory stimuli (e.g., the color of a tree leaf), and higher levels supporting inferences on increasingly abstract concepts, such as estimation of the underlying—hidden—states generating the observed stimuli and the processes that govern the variability in these hidden states (e.g., the seasons of the year). Similar to the existing feedforward and feedback connections between brain regions, levels are interconnected through bottom-up connections sending information from lower to higher levels and top-down connections sending information from higher to lower levels. Critically, this message-passing between levels allows hierarchical inference to combine information across levels (e.g., predicting that tree leaves will turn red by incorporating higher-level, contextual prior knowledge that the Fall has arrived). Although different hierarchical-inference models exist that vary in the exact implementation of message-passing between levels and in their overall architecture, these models are conceptually and algorithmically similar. Of these, two are most relevant to delusions and schizophrenia: generalized predictive coding, here understood broadly to encompass active inference and related models (Adams et al., 2013; Friston et al., 2016; Smith et al., 2020), and belief propagation (Jardri and Denève, 2013). We present a simplified explanation of their differences below.

Generalized predictive-coding models posit that the key signal for belief updating at each level of the hierarchy is a weighted prediction error (PE). The level-specific prediction error reflects the difference between a top-down signal encoding a prior expectation conveyed from the level above and the bottom-up input from the level below. Importantly, this prediction error is scaled based on the relative uncertainties of the top-down prior expectation and the bottom-up signal to favor the less uncertain—or the more *reliable*—of these two sources of information. This relates to the concept of Bayesian cue combination (Daw, 2014; Knill and Pouget, 2004), which is apparent when examining Bayesian inference on the mean, μ, of an underlying continuous variable based on an observed stimulus s (representing a sample of the underlying variable corrupted by Gaussian noise):

$$\mu_{posterior} = \omega_1' \cdot \mu_{prior} + \omega_2' \cdot s \qquad \text{Eq. 5}$$

Here, the prior weight $\omega_1{}'$ and the weight on the sensory observation $\omega_2{}'$ reflect the optimal weighting, which here is not fixed for each individual variable but instead depends on their relative uncertainties or variances σ_{prior}^2 and σ_s^2, such that the two weights add up to 1.

$$\omega_1' = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{prior}^2} \quad \text{and} \quad \omega_2' = \frac{\sigma_{prior}^2}{\sigma_s^2 + \sigma_{prior}^2}, \qquad \text{where } \omega_1' + \omega_2' = 1\#\text{Eqs. 6 and 7}$$

Given that the magnitude of a belief update is the difference between the new, updated belief and the previous one ($\mu_{posterior} - \mu_{prior}$), we can rearrange[1] Eq. 5 to show that this Bayesian belief update is driven by weighted prediction errors ($\omega_2' \cdot PE$), or the difference between the observed stimulus s and its expectation μ_{prior} scaled by the weight on the sensory observation ω_2'.

$$\mu_{posterior} = (1 - \omega_2') \cdot \mu_{prior} + (\omega_2' \cdot s)\#\text{Eq. 8}$$

$$\mu_{posterior} - \mu_{prior} = \omega_2' \cdot \left(s - \mu_{prior}\right) = \omega_2' \cdot PE\#\text{Eq. 9}$$

In generalized predictive-coding models, the weighting of prediction errors at a given level is therefore the key variable controlling belief updates at that level. Within the active inference framework, this weight is adjusted by estimates from higher levels about the variability of the underlying generative process, with the ultimate goal of minimizing surprising outcomes (i.e., by optimizing predictions and acting to minimize surprise) to maintain long-term homeostasis (Friston, 2010). Misestimating the underlying process to be less variable than warranted (e.g., underestimating its volatility), will modify the weight of prediction errors, and belief updating, in lower levels. Under this framework, delusions are proposed to ultimately result from excessive weighting of high-level prior beliefs (as if a high-level $\omega_1{}'$ is overweighted; (Adams, 2018; Adams et al., 2014; Adams et al., 2013)). However, this is framed as a secondary, state-dependent compensation for a core alteration consisting of overweighting of sensory evidence at the lower levels (as if a low-level ω_2' is overweighted). Initially this alteration causes large fluctuations in beliefs, possibly boosting bottom-up salience of irrelevant sensory stimuli in line with theories of salience misattribution (Corlett et al., 2009; Fletcher and Frith, 2009; Heinz et al., 2019; Kapur, 2003; Sterzer et al., 2018). But the system's tendency towards minimizing surprise leads to a compensatory overweighting of high-level prior beliefs, which eventually stabilizes beliefs.

[1] We first obtain Eq. 8 from Eq. 5 via the substitution of a rearranged Eq. 7, namely $\omega_1' = 1 - \omega_2'$. We may then use Eq. 8 to examine the Bayesian update as the difference:

$$\mu_{posterior} - \mu_{prior} = (1 - \omega_2') \cdot \mu_{prior} + (\omega_2' \cdot s) - \mu_{prior}$$

Distributing and canceling the extraneous μ_{prior} terms gives:

$$\mu_{posterior} - \mu_{prior} = \mu_{prior} - (\omega_2' \cdot \mu_{prior}) + (\omega_2' \cdot s) - \mu_{prior}$$
$$\mu_{posterior} - \mu_{prior} = (\omega_2' \cdot s) - (\omega_2' \cdot \mu_{prior})$$

A simple reorganization of the above ω_2' terms then yields the desired result in Eq. 9.

In the belief propagation model (Denève and Jardri, 2016; Jardri and Denève, 2013; Leptourgos et al., 2017), in contrast, logit beliefs are iteratively updated based on logit likelihoods reflecting the strength of the evidence at a given level, with increasing levels representing beliefs about broader concepts (e.g., green → leaves → trees → forest). Critically, the top-down and bottom-up connections between levels are governed by independent self-inhibitory processes, presumed to depend on distinct subpopulations of inhibitory (GABAergic) interneurons. An adequate level of inhibition prevents reverberation of messages (i.e., the same message being sent multiple times) reflecting either bottom-up sensory evidence or top-down prior beliefs. In turn, disruptions in the inhibitory processes, hypothesized to derive from alterations in excitation-to-inhibition balance in schizophrenia, lead to alterations in inference characterized by overcounting messages. This scenario is termed ‘circular inference’. Bottom-up disinhibition leads to reverberation or overcounting of sensory evidence, which effectively implements a type of overweighting of sensory evidence; top-down disinhibition leads to reverberation or overcounting of prior beliefs, which effectively implements a type of overweighting of prior beliefs. In the short run, circular inference was shown to explain excessive belief certainty in the face of weak sensory evidence. In the long run, circular inference captured the development of strong and certain probabilistic associations between higher-level and lower-level constructs when these were actually unrelated and only weak evidence supported their association. The circular-inference model produces delusion-like conditional beliefs—false, overly certain, and rigid—only in ambiguous situations, which was proposed to explain the persecutory nature of delusions given the high inherent uncertainty of social inferences (relative to lower-level perceptual inference). Although Jardri and Denève (2013) suggested that bottom-up or top-down disinhibition could be consistent with different behaviors observed in schizophrenia, invoking in part the beads-task literature (see below), they proposed that psychotic symptoms such as delusions primarily originate from bottom-up disinhibition leading to overcounting of sensory evidence.

Empirical findings and gaps in the literature on inferential alterations in delusions

The inferential models of delusions described above inspired a substantial body of work aimed at empirically testing model predictions to isolate the cognitive and computational mechanisms underlying delusions in schizophrenia-spectrum disorders. Reframed in computational-psychiatry terms, the ultimate goal of this effort is to identify the failure mode(s) in inferential processes that give rise to delusions. This goal requires the ability to isolate interindividual variability in behaviors which can be selectively attributed to altered inferential processes and subprocesses, rather than to broader cognitive deficits such as those typically seen in schizophrenia (e.g., global neurocognitive deficits in working memory, verbal memory, and processing speed generally unrelated to positive symptoms like delusions) or other general factors associated with the illness (e.g., chronicity, institutionalization or hospitalization, socioeconomic conditions, medication, co-morbid psychiatric and medical conditions). So, can we do this?

The most prolific experimental paradigm in empirical studies of inference in schizophrenia is the "beads task" (also known as the "urn and beads task"), itself an instantiation of the so-called "bookbag and poker-chip" experiments (Benjamin, 2019). Based on Hemsley and Garety's theoretical framing for delusions, Huq et al. (Huq et al., 1988) conducted the first experiment using the beads task in schizophrenia. In their task, participants were shown two jars filled with a mixture of colored beads, with the majority color defining the identity of the jar (jar A: 85% beads of color a, 15% beads of color b; jar B: 85% beads of color b, 15% beads of color a). Next, the jars were hidden, and participants were informed that one of the jars would be chosen at random with equal probability. Participants were presented with one bead at a time from the chosen jar (randomly drawn from the jar with replacement) and after each bead was presented, participants could guess the identity of the chosen jar (jar A or jar B) or request another bead. With Eqs. 1-2 as a reference, it should now be straightforward to see how this task was designed to capture a process of causal inference on hidden states (the hidden jars): here, the observed color of the bead at a given draw provides an information sample s (where s can take on colors a or b) used to update beliefs about the identity of the chosen jar [$P(A|s)$ or $P(B|s)$], which according to Bayes' theorem should depend on the prior belief before observing this bead [$P(A)$ or $P(B)$] and the likelihood or strength of the evidence supporting each jar [in this case, $P(a|A) = 0.85$ and $P(b|A) = 0.15$ for jar A, and *vice versa* for jar B]. The main behavioral measures in this beads task were *draws-to-decision*, the total number beads requested before making a final guess, and reported probability estimates of the chosen jar being A or B elicited after each bead draw (a subjective estimate of the posterior belief of the chosen jar). No method to incentivize reporting of true beliefs or preferences was used. Task behavior was obtained from 15 participants diagnosed with schizophrenia and active, severe delusions, 10 psychiatric controls without a diagnosis of schizophrenia and without delusions, and 15 healthy controls. The main results were that patients with schizophrenia requested fewer beads before making a guess relative to both control groups, i.e., they exhibited reduced draws-to-decision, and tended to report higher probability estimates for the chosen jar after seeing only one bead. The reduction in draws-to-decision in schizophrenia was later dubbed the "jumping to conclusions" bias (Dudley et al., 1997a, b) and has been broadly replicated in subsequent research, as discussed below. Setting the stage for later work, Huq et al. evaluated these behavioral results against the Bayesian-inference benchmark described above and put forward the influential interpretation that patients with delusions tended to overweight the evidence associated with the bead samples. Concretely, the authors argued that patients with delusions were less susceptible to conservatism bias, which can be defined as the underweighting of the likelihood (i.e., as if the likelihood weight ω_2 in Eq. 4 was relatively greater in the schizophrenia patient group than in the control groups). This interpretation was supported by higher reported probability estimates after the first bead in patients with delusions, suggesting at least a relative overweighting of the likelihood. The authors also took the decrease in draws-to-decision to support this interpretation, assuming that more certain posterior beliefs (i.e., estimated probabilities closer to 1) would increase the probability of patients venturing a guess.

While compelling, this work stopped short of pinpointing a specific link between delusions and inferential alterations. Despite their laudable efforts to isolate delusional processes, the active delusions group in Huq et al. conflated delusions with active psychotic symptoms and with a diagnosis of schizophrenia, precluding the attribution of any group differences to delusions specifically. Furthermore, they did not discuss or rule out alternative explanations apart from inferential alterations, such as disproportionate effects in their active patient group of general cognitive deficits (e.g., broader, non-specific neurocognitive deficits that could interfere with performance on this task, as they do with a variety of other tasks) or other motivational determinants to stop sampling.

After the seminal work by Huq et al., the beads task became a widespread paradigm in studies on inference and delusions (Dudley et al., 2016; McLean et al., 2017; Ross et al., 2015), which heavily focused on draws-to-decision as a convenient measure of presumed relevance to inferential processes. Many of these subsequent studies have used the classic version of the task, with little or no modifications from Huq et al.'s task, although a common variant includes a memory aid indicating previous bead draws within a trial to control for potential working-memory confounds (Dudley et al., 1997b). Notably, these experiments typically included very few trials of the beads task—only 1 or 2 trials per likelihood condition in many cases—and often reused the same sequences from previous studies. Three recent meta-analyses have summarized this large body of work. In general, studies consistently find that patients with schizophrenia tend to exhibit the jumping-to-conclusions bias, characterized by decreased draws-to-decision compared to healthy or psychiatric controls. But critically, these meta-analyses do not provide clear evidence for a specific link to delusions. One of these meta-analyses (Dudley et al., 2016) found no evidence of differences in jumping-to-conclusions bias when comparing patients with schizophrenia who had active delusions to those who did not have active delusions after controlling for study quality and other factors. Another meta-analysis (McLean et al., 2017) did find group differences when comparing groups *with* active delusions to groups *without* active delusions, including schizophrenia and other psychiatric diagnoses. However, the sample descriptions suggest these groups may correspond more generally to '*actively psychotic*' and '*stable*' patients, respectively. Consequently, differences between these groups could be due to factors unrelated to delusions, such as interference of positive symptoms and disorganization with task performance, general illness severity, and several other cognitive, motivational, and treatment-related factors. To circumvent this issue, several studies have focused on correlating measures of task performance such as draws-to-decision with specific measures of delusion severity. A common measure of delusional and delusion-like ideation in this literature has been the Peters Delusion Inventory (PDI; (Peters et al., 2004)). The third meta-analysis (Ross et al., 2015) focused on studies examining correlations with interindividual variability in PDI scores. While this meta-analysis found a correlation between the jumping-to-conclusion bias and higher PDI scores, this effect was only present when analyzing clinical and non-clinical populations together or in non-clinical populations alone, but was absent when limiting the analysis to patients who were clinically delusional. Altogether, despite the consistent evidence for a jumping-to-conclusions bias in schizophrenia, clear support for a specific relationship between reduced draws-to-decision and clinical delusions in psychotic patients is lacking from this literature.

In addition to the classic, draws-to-decision version of the beads task, "graded estimates" or probability-estimation versions of the beads task show participants a predetermined number of bead draws and prompt them on a draw-by-draw basis to submit continuous probability estimates indicating their certainty about the hidden jars on a Likert or visual analogue scale (Moritz and Woodward, 2005; So et al., 2016; Speechley et al., 2010; Young and Bentall, 1997). Thus, these tasks aim to directly elicit the subjective posterior beliefs about the hidden jars given an observed sequence of beads [e.g., the subjective version of $P(A|aaba)$] instead of eliciting sampling decisions based on these beliefs. Studies using this probability-estimation method generally find that patients with schizophrenia and delusions tend to report higher levels of certainty earlier than healthy controls, which in principle accords with delusional beliefs being held with high certainty. At odds with the definition of delusions, however, these studies also show that patients change their estimates *more* in response to beads that represent "disconfirmatory" evidence or evidence against the most likely chosen jar up to that draw [e.g., the last bead b in the sequence $aaaab$, which counters the previous evidence for the chosen jar being A, decreasing the certainty of the posterior belief for jar A such that $P(A|aaaa) > P(A|aaaab)$]. Based on the argument laid out above, these results are consistent with the notion of a jumping-to-conclusions bias in patients. However, as with the draws-to-decision tasks, the definition of patient groups in these studies precludes attributing behavioral differences specifically to delusions (as opposed to schizophrenia or active psychosis). Further complicating this picture, the effects in the probability-estimation paradigms are less robust and less replicable (Fine et al., 2007) than those on the standard draws-to-decision measure (Ross et al., 2015). Moreover, despite notable exceptions (Adams, 2018; Schmack et al., 2013; Stuke et al., 2017; Stuke et al., 2019), common analytical approaches to probability-estimation beads tasks hinder their interpretation in terms of subjective beliefs. Continuous changes in reported probabilities as a function of draws are often discretized into measures such as draws-to-maximum-certainty, effectively treating the data in the same fashion as draws-to-decision. Beyond these considerations, even if the phenotypes from probability-estimates beads tasks had been empirically linked to delusions, a general account of delusions in terms of a presumed increase in weighting of evidence or likelihood (i.e., increased ω_2) would still face the critical challenge of explaining the rigidity and resistance to disconfirmatory evidence that defines delusional beliefs in general (with perhaps the exception of specific phenomena like 'delusional perception'; but see Adams, 2018).

Decreased draws-to-decision, and perhaps other behaviors elicited by beads-task paradigms, are associated with a diagnosis of schizophrenia but not specifically with delusions. If not a delusion-related process, what do these behaviors reflect? As with performance impairments on any cognitive task in a clinical population such as schizophrenia, an obvious culprit is the global neurocognitive deficit inherent to the illness. Against the backdrop of broad motivational (Green et al., 2012; Nakagami et al., 2008; Takeda et al., 2017) and neurocognitive deficits associated with schizophrenia (Fioravanti et al., 2005; Habtewold et al., 2020; Luck et al., 2019), impaired performance could be explained by an inability to comprehend or retain task instructions, insufficient task engagement, performance anxiety, or feeling rushed, among other factors. Indeed, several prior studies supporting this notion (Balzan et al., 2012a; Dudley et al., 1997b; Freeman et al., 2014; van der Leer and McKay, 2014) directly challenge the ability of the classic beads task to isolate inferential processes (Baker et al., 2019; Fine et al., 2007; McLean et al., 2020a; McLean et al., 2020b; Ross et al., 2015). But perhaps the most conclusive finding in this regard came from a recent beads-task study in the largest schizophrenia sample to date (Tripoli et al., 2020), which included 817 patients with first-episode psychosis and 1,294 controls from the general population. Here, the jumping-to-conclusions bias in patients with schizophrenia was fully explained by lower IQ (that is, diagnosis effects were no longer significant after accounting for IQ in a mediation analysis), indicating that the jumping-to-conclusions bias resulted from a global cognitive deficit rather than from a more circumscribed delusion-related process. Further supporting this notion, this study reported a correlation between delusion severity and *increased*—not decreased—draws-to-decision, although this effect was less robust.

Decreased draws-to-decision, and perhaps other behaviors elicited by beads-task paradigms, are associated with a diagnosis of schizophrenia but not specifically with delusions. If not a delusion-related process, what do these behaviors reflect? As with performance impairments on any cognitive task in a clinical population such as schizophrenia, an obvious culprit is the global neurocognitive deficit inherent to the illness. Against the backdrop of broad motivational (Green et al., 2012; Nakagami et al., 2008; Takeda et al., 2017) and neurocognitive deficits associated with schizophrenia (Fioravanti et al., 2005; Habtewold et al., 2020; Luck et al., 2019), impaired performance could be explained by an inability to comprehend or retain task instructions, insufficient task engagement, performance anxiety, or feeling rushed, among other factors. Although overlooked in earlier studies, more recent work indeed supports a role for these non-inferential factors in the jumping-to-conclusions bias observed in schizophrenia (Balzan et al., 2012a; Dudley et al., 1997b; Freeman et al., 2014; Tripoli et al., 2020; van der Leer and McKay, 2014; White and Mansell, 2009), directly challenging the ability of the classic beads task to isolate inferential processes (see Box 1 for a more detailed discussion). But perhaps the most conclusive finding in this regard came from a recent beads-task study in the largest schizophrenia sample to date (Tripoli et al., 2020), which included 817 patients with first-episode psychosis and 1,294 controls from the general population. Here, the jumping-to-conclusions bias in patients with schizophrenia was fully explained by lower IQ (that is, diagnosis effects were no longer significant after accounting for IQ in a mediation analysis), indicating that the jumping-to-conclusions bias resulted from a global cognitive deficit rather than from a more circumscribed delusion-related process. Further supporting this notion, this study reported a correlation between delusion severity and *increased*—not decreased—draws-to-decision, although this effect was less robust.

Taken together, these results strongly challenge the common assumption that the jumping-to-conclusions bias, and its hypothesized computational underpinnings (e.g., overweighting of likelihoods in inferences on hidden states), play a general and significant role in the genesis or maintenance of delusions in schizophrenia. More generally, the demonstrated susceptibility of the standard draws-to-decision measure to general cognitive impairment questions its suitability as a tool for selective interrogation of inferential processes relevant to delusions. How can we better probe these processes?

Distinguishing inferential and non-inferential processes

The preceding discussion implies the need to devise improved paradigms for isolating inferential processes and alterations therein. To expand further on our definition of inference, and dispel common misconceptions in the literature, we first distinguish inferential processes from other non-inferential processes involved in decision making.

In describing the different conventional beads-task paradigms, we focused on two metrics: the reported probabilities indicating certainty about the hidden jars (the main measure from the probability-estimation tasks) and the decisions to continue or stop drawing additional beads (the main measure from the draws-to-decision tasks). These behaviors are typically thought to map onto two distinct processes and are often studied with different paradigms: the first reflects subjective posterior beliefs about hidden states [e.g., $P(A|ab)$] such as those obtained through *belief-elicitation* tasks; the second reflects sampling decisions such as those studied via *information-sampling* paradigms. These two processes are fundamentally distinct. The first reflects a belief while the second reflects an action based on that belief. To further illustrate their precise differences, and to shed light on the process of making decisions on the basis of beliefs, we turn to an optimal model for sampling decisions that has been applied to solve the beads task and similar problems (Averbeck, 2015; Kaelbling et al., 1998): the partially observable Markov decision process (POMDP).

Again, the draws-to-decision version of the beads task is an information-sampling paradigm that measures decisions to sample or to stop sampling beads. Bayesian inference alone does not provide a solution for making this type of decision. The POMDP algorithm (Fig. 1) incorporates Bayesian inference and additionally maximizes rewards in sampling decisions by finding the turn (e.g., draw or sample number) at which the costs of information sampling (the costs of drawing an additional bead and the expected future gains derived from it) outweigh the costs of incorrectly guessing hidden states (guessing the identity of the chosen jar), at which point a rational agent should stop sampling. In the context of the beads task, the POMDP provides the optimal draws-to-decision for any given bead sequence and cost structure. Critically, the solution depends on the explicit costs of sampling and on choice accuracy—that is, the penalty associated with a bead draw and with an incorrect jar guess, as well as the reward associated with a correct guess (in monetary or other units). But more important for our illustration are the mechanics through which the POMDP reaches a sampling decision.

The POMDP can be portrayed as the combination of three modules that are hierarchically nested: Bayesian inference (Fig. 1b), value comparison (Fig. 1c), and choice (Fig. 1d). Bayesian inference is used to compute probabilistic beliefs about the hidden states (Fig. 1b) based on observed samples (Fig. 1a). Based on these beliefs, which reflect the intuited probabilities of different outcomes, and on the rewards and costs of those outcomes, an expected value for each alternative option (drawing and guessing in future turns versus guessing at the current turn) is calculated and compared (Fig. 1c). Finally, the option with the highest expected value is chosen (Fig. 1d). This approximately maps onto the consecutive steps which participants completing the beads task may follow, at least if they were given explicit costs for a bead draw and for an incorrect guess and an explicit reward for a correct guess. Intuitively, early in a trial and after observing only a few beads, participants will be uncertain about the identity of chosen jar [e.g., $P(A|ab) {\sim} P(B|ab) {\sim} 0.5$] because they have only gathered a small amount of evidence. If they were to make a guess at that point, the probability of an error would be high (~0.5). Assuming the cost of an incorrect guess is high enough and they are motivated to avoid it, participants would lean towards drawing another bead, assuming also its cost is low enough. In other words, at that point, the expected value of drawing is higher than that of guessing. But after drawing enough beads, once participants are very certain about the identity of chosen jar [e.g., $P(A|abaaaa) \gg P(B|abaaaa)$], the expected probability of an incorrect guess would be low and the expected value of guessing (and obtaining the reward associated with a correct guess) would exceed that of drawing, at which point the optimal choice would be to stop sampling and guess. The number of draws before the guess in this scenario would thus correspond to the optimal draws-to-decision behavior for that sequence and cost structure.

Critically, the POMDP illustrates that decisions to sample are based on beliefs about hidden states, but are still distinct from them. In the example above, the posterior belief about jar A after observing the bead sequence $abaaaa$ is the probability $P(A|abaaaa)$. In turn, the expected value of guessing A depends on the probability of an incorrect response, which is a function of the posterior belief, and on its cost. More generally, and beyond the POMDP (Glimcher and Rustichini, 2004), the expected value of choosing an option reflects the costs associated with the different possible outcomes (e.g., A being indeed the chosen jar or not) resulting from that choice, weighted by their probabilities. In the example case, this is given by the following equation (where positive costs would reflect rewards and negative costs penalties):

$$EV_{guess\ A} = \mathrm{P}(A|abaaaa) \cdot Cost_{correct} + \mathrm{P}(B|abaaaa) \cdot Cost_{incorrect} + \text{draw number} \cdot Cost_{draw} \quad \#\text{Eq. } 10$$

The POMDP calculates the expected value of all possible options: guessing A, guessing B, and drawing. The expected value of drawing is more complex as it involves the calculation of a tree of possible outcomes contingent of future choices as well as their costs (see Kaelbling et al., 1998 for the full algorithm, and Averbeck, 2015 and Baker et al., 2019 for its applications to the beads task). Even more importantly for our illustration, the decision to continue or stop sampling and guess the more likely jar is simply made by taking the option with the highest expected value, i.e. $max(EV_{guess\ A}, EV_{guess\ B}, EV_{draw})$[2]. Therefore, although sampling decisions and expected values depend on posterior beliefs, other factors like the costs associated with different outcomes also influence these variables. In the context of the beads task, this strongly suggests that draws-to-decision depends not only on inferences about hidden states but also on the costs attributed to different courses of action. These costs may be implicit or explicit, related to financial costs, cognitive effort, social rewards, or others related factors. This can be shown by parameterizing the POMDP, which allows for the simulation of changes in draws-to-decision by modifying costs and other variables. Increased (subjective) costs of drawing, for instance, produces decreased draws-to-decision (Baker et al., 2019).

Sampling decisions in information-sampling paradigms such as the draws-to-decision beads task are thus best conceptualized as a value-based decision. Interindividual differences in draws-to-decision would appear likely to depend on subjective valuation processes distinct from inference and cannot

[2] Here, in line with the standard POMDP model, we use a deterministic choice rule whereby the action (guessing or drawing) with the highest expected value is selected. However, a softmax choice rule is commonly implemented in parameterized models to select an action probabilistically as a function of expected value (Baker et al, 2019; Averbeck et al, 2015; Moutoussis et al, 2011). As the difference in expected value between actions increases, so does the likelihood that the action with higher expected value will be selected. Choice stochasticity is modeled by incorporating an additional 'temperature' parameter that scales these likelihoods.

provide a direct readout of inferential processes unless the non-inferential valuation processes are carefully controlled. This notion is supported by preliminary data from our group (Baker et al., 2019) and other direct demonstrations that beads-task behaviors depend on task incentives (Grether, 1992; van der Leer and McKay, 2014b), as well as on the subjective evaluation of those incentives (Ermakova et al., 2019). The corollary is that decreased draws-to-decision in schizophrenia may reflect a number of non-inferential, valuation processes (Box 1). Specifically, patients may tend to draw fewer beads simply because they attribute different subjective costs to drawing or incorrect guesses compared to controls, especially given that the classic beads task does not stipulate explicit costs. Patients may be less motivated to make accurate guesses or more sensitive to the cognitive costs of additional samples. Alternatively, decreased draws-to-decision could reflect a calculation involving the subjective value of the time spent performing the task at the expense of other activities. The possibility of terminating the classic beads task by deciding to stop drawing earlier further suggests that a participant focused on maximizing reward rate may decide to do just that, in which case the "jumping-to-conclusions" behavior would actually reflect an optimal strategy.

In sum, alterations in draws-to-decision could reflect a number of changes in value-based decisions apart from inference, and insufficient control over these non-inferential factors in classic versions of the beads task precludes their distinction from inferential processes (see Box 1 for a more detailed discussion of these factors and suggested approaches to minimize them). We now turn to more novel approaches to measuring inference that permit better control over these non-inferential factors.

Box 1. Potential non-inferential factors accounting for the jumping-to-conclusions bias in schizophrenia

In different sections of this paper, we discuss non-inferential factors that likely contribute to the common finding of decreased draws-to-decision in schizophrenia. These factors stand in contrast with the genuine and concrete alterations in causal inference that we hypothesize to underlie delusions—specifically, overweighting of prior beliefs in higher-level inference on hidden states. Here, we summarize these non-inferential factors and suggest concrete approaches to minimize or account for their contributions to sampling decisions such as those determining draws-to-decision behavior.

- Broader cognitive deficits that may generally interfere with task construal and performance. Broad neurocognitive deficits in schizophrenia (Fioravanti et al., 2005; Habtewold et al., 2020; Luck et al., 2019) include deficits in motivation (Green et al., 2012; Nakagami et al., 2008; Takeda et al., 2017), working memory (Forbes et al., 2009; Griffiths and Balzan, 2020), longer-term memory (Guo et al., 2019), and goal-directed planning (Siddiqui et al., 2019). Impaired performance on an information-sampling task may thus simply result from inability to comprehend or retain task rules and instructions (Balzan et al, 2012a; Balzan et al., 2012b; Ross et al., 2015), insufficient task engagement (e.g., due to motivational deficits or misunderstanding), anxiety (Lincoln et al., 2010a) or feeling rushed (White and Mansell, 2009) (e.g., due to awareness of cognitive deficits), among other factors. Cognitive deficits, including low IQ (Tripoli et al., 2020), working memory (Broome et al., 2007; Freeman et al., 2014; Garety et al., 2013), and generally poor performance on neuropsychological testing (Andreou et al., 2015; Falcone et al., 2015; González et al., 2018; Lincoln et al., 2010b), have been shown to explain some or all the variance in draws-to-decision (or discrete presence of the jumping-to-conclusions bias) associated with a diagnosis of schizophrenia. A trivial explanation for reduced draws-to-decision in schizophrenia could be that the default strategy of a participant experiencing miscomprehension, forgetting, and/or anxiety is to terminate the task as early as possible (e.g., to alleviate the discomfort associated with anxiety and confusion). It is also possible that these factors further compound the value-based decision-making factors discussed below. To minimize the contribution of broader cognitive deficits, decisions may be self-paced and experiments may include a comprehensive set of instructions, and comprehension and manipulation checks. Visual memory aids (Dudley et al., 1997b) and reminders of task instructions throughout the task may also be advantageous. Additionally, beads tasks should generally include sufficient trial repetitions to reliably ascertain task behaviors accounting for response variability (Balzan et al., 2017; McLean et al., 2018, 2020b; Moritz et al., 2017).

- Other general factors associated with schizophrenia that may generally interfere with task construal and performance. In addition to the broad cognitive deficits mentioned above, other disease-general factors that that tend to differ between patients with schizophrenia and controls may impact task performance. These include socioeconomic status (Hakulinen et al., 2020; Hudson, 2005), which may partly reflect impairments in cognitive functioning (Goldberg et al., 2011), co-morbid conditions, chronicity, institutionalization, and effects of psychiatric treatments. Some of these social factors may contribute to decreased familiarity to related tasks and the type of computer devices used to administer tasks. In addition, antipsychotic and other psychiatric medication may affect inference directly (Andreou et al., 2014; So et al., 2010) or indirectly (e.g., due to somnolence and inattention). These factors may result in decreased draws-to-decision for the reasons discussed in the point above and may be minimized using similar strategies. In addition, these issues may be addressed by conducting studies with larger samples of groups that are more closely matched on all relevant dimensions, including subsets of subjects with comparable socio-economic status and enough higher-functioning and unmedicated patients, patients in earlier stages of their psychotic illness, and appropriate psychiatric and healthy control groups (Fine et al., 2007). Testing and reporting the effects of these variables in specificity analyses is also desirable.

- **Specific alterations in value-based decision-making affecting sampling decisions**. Broad motivational deficits and more circumscribed alterations in value-based decision-making are common in schizophrenia (Gold et al., 2008; Strauss et al., 2014). In a non-incentivized sampling task, patients could exhibit decreased draws-to-decision because they assign less subjective value to possible incorrect guesses (e.g., due to differences in demand characteristics and the motivation to please the experimenter, possibly in relation to alterations in social reward processes; (Catalano et al., 2018; Fett et al., 2019; Lee et al., 2018)) or higher subjective value to collecting additional information samples (e.g., due to the additional time investment and the associated decrease in reward rate or perhaps due to increased perceived cognitive effort associated with integrating additional evidence, which could be related to alterations in cognitive-effort discounting; (Chang et al., 2020; Hartmann-Riemer et al., 2018; Kreis et al., 2020)). Choice stochasticity[2] could also contribute to diagnostic differences (Moutoussis et al., 2011). Financially incentivized tasks can minimize some of these factors (e.g., the contribution of social factors and their differential impact on clinical groups) and provide more experimental control over value-based decisions, which together with modeling can help parse contributions of valuation and choice (Baker et al., 2019). Disincentivizing certain strategies such as rushing through the task, for instance by imposing a minimum task duration, may also minimize the contribution of some of these factors and help homogenize task-solving strategies.

[FIGURE 1 HERE]

Enhanced approaches to probe inference and novel findings

With the abovementioned limitations in mind and building on prior modeling work (Furl and Averbeck, 2011; Moutoussis et al., 2011), we recently developed a variant of the beads task designed to isolate inferential alterations underlying delusions (Baker et al., 2019). This task is an information-sampling task where participants choose at each iteration within a trial whether to draw a bead or guess the identity of the chosen jar, which can thus measure draws-to-decision behavior. It also has a built-in belief-elicitation component consisting of prompts for probability estimates before each choice, recorded on a continuous sliding scale, to allow for a more direct readout of inferential processes. The establishment of an explicit cost structure (with an initial endowment of $30 and explicit costs for sampling, -$0.30, and incorrect guesses, -$15), along with a minimum task duration, further makes the task *incentive compatible* and renders the resulting data tractable to the POMDP framework. Consistent with the behavioral economics literature at large (Camerer, 1997; Camerer and Mobbs, 2017; Camerer et al., 2016; Ortmann, 2009; van der Leer and McKay, 2014) and specific clinically relevant applications (van der Leer and McKay, 2014), our experience suggests that an incentivized task is critical to engage participants and ensure their responses reflect their true preferences, particularly in clinical populations. Further, the task administration protocol includes comprehensive instructions which emphasize the objective of maximizing rewards on the task, practice trials that serve to ensure task comprehension, and a visual aid to control for possible working-memory deficits.

We obtained data with this controlled task in 24 patients with schizophrenia with varying levels of delusional severity (11 of them unmedicated with antipsychotics) and 21 healthy controls (Baker et al, 2019). First, a number of checks demonstrated the effectiveness of the various manipulations: sensitivity to task manipulations at the individual level and responses on a post-task questionnaire indicated participants adequately understood the task, which with the lack of systematic biases in initial (pre-bead) probability estimates, suggested that the data comported with model assumptions. A critical finding in this study was the strong correlation within patients between *increased* draws-to-decision and higher delusion severity scores, measured by PDI score, a finding at odds with the conventional wisdom of the beads task literature (but consistent with other data, including Tripoli et al, 2020). Importantly, this increase in draws-to-decision was specific to delusions, compared to a number of other clinical variables—even other positive symptoms—and cognitive and sociodemographic factors, and held in unmedicated patients alone. The insensitivity to general factors, including numeracy and working-memory performance, implied that global cognitive deficits were not a main driver of the observed variability in task behavior. Indeed, patients with delusions tended to exhibit better accuracy than non-delusional patients. Beyond the delusion-specific effect, we found that patients as a group showed the expected decrease in draws-to-decision compared to controls, but only when controlling for PDI scores, and this diagnosis effect disappeared after controlling for socioeconomic status. Altogether, these results describe (1) a more selective process linking increased information sampling to increased delusion severity and (2) a more general process linking decreased information sampling (a jumping-to-conclusions-type bias) to the lower socioeconomic status and cognitive deficits associated with schizophrenia, in line with later work (Moritz et al., 2020; Tripoli et al., 2020); Box 1). This result raised the question of whether inferential processes were driving the delusion-related increase in information-sampling behavior.

We turned to the draw-by-draw probability estimates provided by the participants for an answer. A weighted Bayesian model equivalent to that in Eq. 4 provided a reasonable fit to the probability estimates and captured qualitative differences in changes in the estimates over draws, which appeared to update more slowly in more delusional patients. More importantly, we used the fitted model parameters for the prior weight ω_1 and likelihood weights ω_2 (one for each likelihood condition in the task) for each participant to evaluate interindividual deviations as a function of delusion severity. In line with previous work, healthy individuals and patients with low delusion severity tended to underweight prior beliefs ($\omega_1 < 1$). Our central finding, however, was that higher fitted values of the prior weight ω_1 correlated with both higher delusion severity and with increased draws-to-decision behavior in patients, suggesting that both delusions and their effect on information sampling depended on a specific inferential failure mode consisting of a relative prior overweighting (or lessened prior underweighting[3]) compared to non-delusional patients. This interpretation was further corroborated by model-agnostic analyses and simulations of selective changes in the weight of prior beliefs in the context of the POMDP. This finding was specific to inferential processes as opposed to non-inferential processes. In a parameterized POMDP model, we showed that valuation and choice parameters based on subjective posterior beliefs were uncorrelated with delusions and draws-to-decision behavior, as were valuation parameters denoting subjective aversion to loss, risk, and ambiguity on other decision-making tasks.

Using a POMDP-inspired task design with a number of additional controls over standard designs, together with computational modeling of inference and information sampling, allowed us to uncover a candidate failure mode for delusions: a relative overweighting of prior beliefs in inference. This process appears to be clinically specific to delusions and computationally specific to inference. While these results certainly call for replication and extension, they may provide the foundation for a parsimonious, empirically supported model of delusions. Best practices in computational modeling include demonstrating the ability of selectively manipulated models to generate the observed behaviors via *in silico* simulations (Wilson and Collins, 2019), as we did in this work (Baker et al., 2019). In this vein, we will now use model simulations to illustrate how the proposed failure mode—increased prior weight ω_1—produces a dynamic primacy bias in probabilistic belief-updating that captures the defining characteristics of delusional beliefs.

[3] We have elected to refer to this computational phenotype as *relative prior overweighting* with respect to the non-delusional patients, who in absolute terms showed the commonly observed underweighting of prior beliefs. Delusional patients in Baker et al. exhibited prior weights ω_1 closer to the Bayesian benchmark of 1 and therefore this could also be framed as less absolute prior underweighting than non-delusional individuals. However, we find framing this computational phenotype in relative terms to be more intuitive.

Overweighting of prior beliefs as a candidate failure mode for delusions

Our previous empirical findings (Baker et al., 2019) suggest that an inferential alteration consisting of relative overweighting of prior beliefs could be responsible for delusions. It is worth considering whether the opposite is true: whether altered behaviors in delusional patients *result* from their delusions and general suspiciousness rather than reflecting an underlying alteration causing delusions. We considered and ultimately rejected the former possibility due to a number of observations that rendered it implausible (Baker et al., 2019). Instead, we ask here whether prior overweighting could theoretically cause the core phenomenological features of delusions. We mentioned in the introduction that delusional phenomena are highly variable across individuals; the content of delusional beliefs can involve any imaginable topic and varies widely with cultural and experiential context. Even falsity, part of the classical definitions of delusions, is now typically considered unnecessary to deem beliefs as delusional (e.g., as per the DSM-5 definition). The core features refer to their specific *form* as highly certain and rigid beliefs, which are generally considered necessary features of delusions. Could prior overweighting generate excessively rigid and certain beliefs akin to delusions?

We first consider the belief-updating dynamics induced by variations in prior weighting in the context of long-term sequential belief updating. This context is most relevant because in the real-world people usually sample ambiguous pieces of information over relatively long periods of time (Nastase et al., 2020), and because delusions are typically held over months or years with relative insensitivity to momentary situational factors (putting aside for expository purposes the roles of stress and negative emotion on delusion exacerbation (Ben-Zeev et al., 2012; Brenner and Ben-Zeev, 2014; Granholm et al., 2020).

Fig. 2a shows simulated data using the weighted Bayesian model (Eq. 4) in which two agents, identical except that one has a relatively lower prior weight ($\omega_1 = 0.950$) and the other a relatively higher prior weight ($\omega_1 = 0.995$), sequentially update their beliefs about hidden states upon receiving samples of information consistent with one of two complementary hypotheses with respect to the hidden states ($\omega_2 = 1$ for both agents). Note that the specific prior weights for these agents are selected here to visually highlight effects of interest the generality of which is proven later. This simulation is illustrated as the long-run posterior probability estimates produced by these two agents on a beads task where the evidence is weak (likelihoods $P(a|A) = P(b|B) = 0.55$). From the simulation in this ambiguous context, it becomes clear that the prior weight ω_1 affects the *dynamics* of sequential belief updating by controlling a *primacy-recency bias*. Higher ω_1 leads to a relative primacy bias characterized by the increased relative influence of older evidence (and decreased responsiveness to newer evidence) on current beliefs, or more "sticky" (less "leaky") beliefs; lower ω_1 leads to a recency bias characterized by a reduced influence of older evidence (and increased responsiveness to newer evidence) on current beliefs, or more "leaky" beliefs. This is in direct contrast to the likelihood weight ω_2, which scales the strength of all evidence equally, and consequently does not produce qualitative, dynamic changes in the belief trajectory (see below). While ω_2 is similar to the drift rate in evidence-accumulation models (Gold and Shadlen, 2007; Smith and Ratcliff, 2004), ω_1 makes the weighted Bayesian model a type of discrete, leaky accumulator (Bogacz et al., 2006; Busemeyer and Townsend, 1993; Usher and McClelland, 2001).

At least at face value, this primacy-recency bias associated with the prior weight ω_1 appears to capture the two core features of delusions. Higher ω_1, similar to that we observed in delusional patients, produces higher certainty and greater rigidity in beliefs, both specifically stemming from a change in ω_1. Higher belief *certainty* is manifest from posterior beliefs reaching asymptotic levels closer to 1 (Fig. 2a)—where 1 denotes complete certainty about the underlying hidden state and 0.5 reflecting total ambiguity. Higher rigidity (or equivalently more "stickiness") in beliefs is clear when examining the belief dynamics in response to randomly drawn samples. Assuming the chosen jar is A (or the black jar in Fig. 2a), if minority samples (b) happen to predominate early on, followed by more majority samples (a) later on, belief updates are more sluggish in the agent with higher ω_1; compared to the low-ω_1 agent, the high-ω_1 agent takes more samples to rectify its belief trajectory to start favoring of the correct hidden state A (Fig. 2a). That is, beliefs in the high-ω_1 agent are *more resistant* to evidence contrary to a favored hypothesis, or more *rigid*. Consistent with the observation from Jardri and Denève (Denève and Jardri, 2016; Jardri and Denève, 2013), these dynamic effects are more apparent in ambiguous contexts, which could explain why more complex and ambiguous social contexts may be fertile ground for the development of delusions. In contrast to the dynamic effects of the prior weight ω_1, changes in the likelihood weight ω_2 can only induce higher belief certainty but not belief rigidity (Fig. 2b).

The mathematics and generality of these effects can be derived from Eq. 4. To illustrate this, we start by re-writing Eq. 4 such that the *logit* posterior belief after seeing sample s, b_s, is the result of a weighted sum of the *logit* prior belief before observing this sample, b_{s-1}, with the *logit* likelihood (or log-likelihood ratio) of sample s, LLR_s. (In the beads task, the LLR_s is defined by the bead color in the current draw and the majority-to-minority ratio of bead colors in the hidden jar[4].)

$$b_s = \omega_1 \cdot b_{s-1} \; + \; \omega_2 \cdot LLR_s \qquad \text{Eq. 11}$$

[4] As in Eq. 2, the LLR_s of sample s is defined as $log\left(\frac{P(s|A)}{P(s|B)}\right)$ and it reflects the momentary evidence associated with this individual sample. For example, if the current sample s is a green bead (a) and the majority-to-minority ratio in the hidden jar is 60:40, the LLR_s for the green jar (A) based on this observed green bead (a), is given by $log\left(\frac{P(a|A)}{P(a|B)}\right)$ = $log\left(\frac{0.6}{0.4}\right) = 0.405$.

By expanding the prior term b_{s-1} to make explicit how the posterior belief would be influenced by evidence from previously observed samples through an iterative process, the effect of ω_1 starts becoming apparent. We illustrate this using a sequence of three samples, the evidence from which is given (in reverse chronological order) by LLR_s, LLR_{s-1}, and LLR_{s-2}.

$$b_s = \omega_1 \cdot (\omega_1 \cdot b_{s-2} \; + \; \omega_2 \cdot LLR_{s-1}) + \; \omega_2 \cdot LLR_s \#\text{Eq. } 12$$

$$b_s = \omega_1 \cdot (\omega_1 \cdot (\omega_1 \cdot b_{s-3} \; + \; \omega_2 \cdot LLR_{s-2}) + \; \omega_2 \cdot LLR_{s-1}) + \; \omega_2 \cdot LLR_s \#\text{Eq. } 13$$

Assuming that the initial prior belief before observing any samples is unbiased ($b_{s-3} = 0$), we can rearrange this formula to clearly see the effects of ω_1 and ω_2 on sequential belief updating.

$$b_s = \omega_1^{s-1} \cdot (\omega_2 \cdot LLR_{s-2}) \; + \; \omega_1^{s-2} \cdot (\omega_2 \cdot LLR_{s-1}) \; + \; (\omega_2 \cdot LLR_s) \qquad \text{Eq. } 14$$

$$b_s = (\textstyle\sum_{n=1}^{s-1} \omega_1^{s-n} \cdot (\omega_2 \cdot LLR_n)) \; + \; (\omega_2 \cdot LLR_s) \qquad \text{Eq. 15}$$

This shows that ω_1 controls the influence of older evidence on beliefs over time. For $0 < \omega_1 < 1$, each sample of older evidence is discounted more than the next due to the increasing powers on the ω_1 parameter. In contrast, ω_2, scales all samples of evidence equally.

Therefore, mathematically, the prior weight ω_1 controls the rate of exponential decay in the contribution of a sample of evidence on a given belief, a form of primacy-recency bias that determines rigidity and responsiveness to new evidence (Baker et al., 2019; Benjamin et al., 2019; Benjamin, 2019; Enke and Graeber, 2019; Grether, 1980). Furthermore, the prior weight ω_1 directly limits maximum belief certainty over the long term. For an infinite series of samples, the posterior belief is bounded as a function of ω_1 and the likelihood ratio (Benjamin et al, 2019), as:

$$max(b_s) = \lim_{s \to \infty} b_s \; = \frac{LLR}{1 - \; \omega_1} \; \#\text{Eq. } 16$$

Per Eq. 16, agents with higher ω_1 have a higher ceiling on belief certainty, consistent with the relatively *high certainty* associated with delusional beliefs. Per Eq. 15 they have a relative primacy bias whereby beliefs are more influenced by older evidence and less responsive to new evidence, consistent with the belief *rigidity* characteristic of delusions. Both core features of delusions stem from higher ω_1.

Eqs. 15-16 thus prove the generality of the effects exemplified in Fig. 2a, where higher values of the prior weight ω_1 simultaneously induce belief trajectories that are more rigid and reach higher certainty. In contrast, higher values of ω_2 only increase the certainty of beliefs without affecting their rigidity (Fig. 2b). Therefore, the dynamic changes in belief updating that capture belief rigidity (i.e., the relative primacy bias) uniquely depend on the prior weight ω_1.

For further clarification, Figs. 2c-e illustrate these belief-updating effects in the short term, over the course of a few samples. A lower-ω_1 agent (Fig. 2c; $\omega_1 = 0.70$; $\omega_2 = 1$), resembling healthy controls, exhibits a clear "leak" in prior beliefs, showing less certain posterior beliefs after observing a sequence, $aaaab$. For $0 < \omega_1 < 1$, because the weighted prior, $\omega_1 \cdot b_{s-1}$, is a fraction of the unweighted prior b_{s-1}, the leak is greater for more certain beliefs and becomes more obvious with more observed samples. This also explains its increased response to the "disconfirmatory" sample b at the end of the sequence, relative to the higher-ω_1 agent. Conversely, an agent resembling delusional patients with high ω_1 (Fig. 2d; $\omega_1 = 0.98; \omega_2 = 1$) exhibits less "leak", ends with higher certainty for A, and responds relatively less to the "disconfirmatory" sample. For contrast, Fig. 2e illustrates the isolated effects of changes in ω_2.

Above, we said that delusional patients in Baker et al. (2019) showed slower belief updating compared to non-delusional individuals. But, in Fig. 2d the delusion-like, higher-ω_1 agent mostly showed increased belief updates relative to the lower-ω_1 agent. How can we reconcile this? An important insight from the dynamics of the weighted Bayesian model is that, unlike the optimal Bayesian model, its belief trajectories depend on the *ordering* in which sequential samples of information are presented; this model's beliefs are *path-dependent*. The magnitude of the difference in belief updates for different values of ω_1 will thus depend on the specific sequence of samples (Figs. 3a-c). Under the POMDP, this has important consequences for draws-to-decision behavior on the beads task. Differences in the prior weight ω_1 induce order-dependent changes in beliefs (Figs. 3a-c) that, in turn, drive differences in the expected value of guessing versus drawing and consequently in draws-to-decision behavior (Figs. 3d-e). Thus, differences in draws-to-decision between delusional and non-delusional individuals—assuming these can be modeled via higher versus lower ω_1 values—will also depend on the sequence of samples, at least to some degree. We illustrate this point by showing that, depending solely on the sequence (the only difference between Figs. 3d and 3e), a higher-ω_1 agent ($\omega_1 = 0.98$) can in principle show either *decreased or increased* draws-to-decision relative to a lower-ω_1 agent ($\omega_1 = 0.89$). For this reason, the specific pattern of delusion-related effects in previous work may, among other things, depend on the specific bead sequences used in a given version of the task. This includes the pattern of delusion-related effects in Baker et al. (2019), where we observed slower belief updating and increased draws-to-decision in delusional patients. Model simulations using the specific bead sequences in that task showed that a selective increase in ω_1 drives increases in draws-to-decision over those particular sequences—this is because, for these sequences, increased ω_1 causes on average slower belief updating and consequently less certain beliefs about the identity of the chosen jar at a given point within a trial, which results in smaller expected values for guessing relative to drawing and an increased tendency to draw. But the predicted behavior would vary for a different set of sequences. This raises yet another foundational issue with using draws-to-decision as a proxy for inference. By introducing sequential dependencies in belief updating, the substantial variability in prior weighting observed across individuals calls into question the utility of an aggregated summary measure such as draws-to-decision to capture the dynamic inferential alterations hypothesized to underlie delusions.

[FIGURES 2 AND 3 HERE]

Normative explanations for changes in prior weighting

We began this paper by considering a *normative* Bayesian model of inference that optimizes estimation accuracy (Eqs. 1-3). One can think of this model as an idealized agent whose behavior is optimal, absent all constraints. Drawing on our own work, we then explored how a parameterized or weighted Bayesian model (Eq. 4) describes deviations from the optimal benchmark and between individuals that are relevant to delusions. We also showed how a particular deviation or failure mode in this *descriptive* model, a relative overweighting of the prior, may be theoretically sufficient to explain the core features of delusions. An unsatisfying aspect of this descriptive approach is that it does not provide a mechanistic explanation for why prior weighting may deviate from the normative optimum, or specify the constraints under which this deviation may actually not be suboptimal. *Prescriptive* models of inference, however, allow parameters (like the prior weight ω_1) to vary as a function of environmental circumstances and/or theorized internal limitations in information processing, permitting adaptations to these constraints. Consequently, in prescriptive models, the mathematically optimal value of a parameter may differ depending on these factors (as opposed to the fixed parameter values in the normative model). Prescriptive models can therefore point to maladaptations to presumed external or internal factors that might drive variability in parameter values. Here, we briefly introduce classes of prescriptive models where variable prior weighting is optimal, to gain theoretical insights into possible mechanistic causes of prior overweighting in delusional patients.

In one such model, the optimal weighting of prior beliefs is governed by environmental volatility, or the frequency of unannounced changes in hidden states (Glaze et al., 2015). The intuition is the following. In a situation where hidden states change abruptly (e.g., the identity of the chosen jar in the beads task suddenly changes mid trial), evidence presented before that change becomes uninformative.

Rationally, if one were able to identify or surmise the changepoint, then they should discount all beliefs formed on the basis of samples presented before the changepoint and start forming new beliefs “from scratch”. More generally, if changes in hidden states are frequent, then it is adaptive to diminish the contribution of (or increase the “leak” of) prior beliefs in a manner approximately equivalent to decreasing ω_1 (although in this model the weight on the prior depends non-linearly on both the likelihood and the hazard rate, H – the probability of a change in the hidden state per unit of time). In short, prior underweighting is optimal when the perceived environmental volatility is high. The corollary is that individuals who underestimate volatility may overweight prior beliefs compared to optimal agents. Therefore, the finding of relative prior overweighting in delusional patients could reflect underestimation of environmental volatility, which could in turn depend on alterations in neuromodulator and neural systems thought to contribute to this process, including the norepinephrine (Silvetti et al., 2013; Vincent et al., 2019) or dopamine (Cools, 2019; Diederen and Fletcher, 2020) systems. We have proposed a related mechanism for hallucinations whereby hallucinating patients with excess nigrostriatal dopamine may overweight lower-level perceptual priors through an inability to encode prior uncertainty (Cassidy et al., 2018), with other data supporting overweighting of lower-level perceptual priors in hallucinators that co-exist with—but do not necessarily depend on—alterations in volatility estimation in psychotic patients (Powers et al., 2017). Other related ideas are indeed commonplace in computational psychiatry, not only in schizophrenia but for several other disorders (Huang et al., 2017 2017; Lawson et al., 2017; Paliwal et al., 2019; Palmer et al., 2017), possibly due to the extensive use of algorithms implementing volatility-dependent hierarchical inference in this literature (Adams, 2018; Adams et al., 2014; Heinz et al., 2019; Mathys, 2011; Stephan and Mathys, 2014; Sterzer et al., 2018). However, whether a volatility account could explain the delusion-related prior overweighting we observed in Baker et al. (2019) is unclear. Arguing against this, our task explicitly instructed participants that hidden states were stable during a trial (i.e., there was no volatility; H=0), so interindividual variability on this task appears more likely to depend on factors other than volatility estimation (although one counterargument is that a neuromodulatory or other neural alteration giving rise to volatility misestimation may be present even in stable environments and still impact behavior in this context). So are there other possible accounts, unrelated to volatility?

Another relevant model posits that inference depends on noisy neural samples that represent prior beliefs with some level of imprecision, and that optimal prior weighting is governed in part by the internal costs of improving precision in the representation of prior beliefs. This model can be placed within a larger class of models popular in the economics literature, the so-called “bounded rationality” models (Simon, 1990). Instead of solely focusing on environmental constraints, these models also consider optimal adaptations to internal limitations, or constraints, in information processing. In other words, these models prescribe how optimal agents like humans and other animals should behave given their limited cognitive resources. In the case of the noisy sampling model of inference recently proposed by Azeredo da Silveira and Woodford (Azeredo da Silveira and Woodford, 2019), resource-limited agents are assumed to access a representation of prior evidence through noisy sampling, providing an imprecise reproduction of prior beliefs (Note that the term ‘sample’ is not to be confused with that we used in the context of information-sampling tasks, where a sample corresponded to an observed piece of objective evidence in the task, like a bead draw; here we use this term to refer to neural samples or instances of a cognitive retrieval process that represents prior information without requiring full access to it). The precision of this prior estimate can increase, reducing noise in the samples, but that comes at the cost of allocating more cognitive resources. This creates a tradeoff between the costs of cognitive precision and the cost of inaccurate predictions. An optimal agent can find the balance between these two costs by diminishing its reliance on prior evidence, which would be reflected in our descriptive model by decreasing the prior weight ω_1. This is consistent with data showing that humans tend to underweight prior beliefs, as mentioned above, which leads to posterior beliefs that are more responsive to new evidence and which always retain some level of uncertainty (like the lower-ω_1 agents in Fig. 2a and 2c). The notion of prior sampling is also consistent with other work supporting the plausibility of sampling-based models of approximate Bayesian inference (Bornstein et al., 2018; Haefner et al., 2016 2010; Heng et al., 2020; Hoyer and Hyvärinen, 2003; Shadlen and Shohamy, 2016). Applied to delusions, this framing may suggest that prior overweighting could result either from alterations in the prior-sampling process itself (e.g., increased redundancy and decreased noise in prior samples) or from alterations in strategies used to resolve the tradeoff (e.g., if delusional patients underestimate the cost of cognitive precision).

Beyond these two models, which can broadly explain prior overweighting as a consequence of maladaptations to environmental volatility or limited cognitive resources, a third possibility goes back to the standard algorithm of normative Bayesian inference. As mentioned above (Eqs. 5-8), a tradeoff between the prior weight ω_1 and the likelihood weight ω_2 is commonly assumed in Bayesian inference on continuous variables and consistent with empirical data demonstrating reliability-weighting in inference (Aller and Noppeney, 2019; Chambon et al., 2017; Chambon et al., 2011a; Chambon et al., 2011b; Fetsch et al., 2012; French and DeAngelis, 2020; Orbán and Wolpert, 2011). Under such a tradeoff, the overweighting of prior beliefs could result from decreased reliability in the representation of new evidence (Teufel et al., 2015). More work is thus needed to arbitrate between this and the other possible explanations discussed in this section.

Evidence for hierarchical-inference models of delusions

As mentioned earlier, weighting of prior beliefs and sensory evidence can also be accomplished through hierarchical message passing. What is the evidence that delusions result from alterations in these hierarchical processes?

The hierarchical-inference models discussed earlier theorize that delusions result directly or indirectly from increased weighting of sensory evidence. Generalized predictive-coding models suggest that overweighting of sensory evidence at low levels of the hierarchy, which initially causes amplified belief updating, secondarily result in an overcompensation characterized by overweighting of prior beliefs at higher levels (Adams et al., 2013). The latter stage is in principle consistent with the proposed failure mode we discussed at length. In contrast, the proposed version of circular inference discussed above posits that delusions primarily arise from disinhibition of bottom-up messages conveying sensory evidence (Jardri and Denève, 2013). While the belief-propagation model is itself hierarchical, the proposed alteration affects bottom-up connections similarly across the levels of the hierarchy. That is, the proposed alteration is not level-specific, although the hierarchical architecture of the model still enables level-dependent changes in belief updating. In any case, the proposed failure mode in circular inference would effectively manifest as overweighting of sensory evidence.

While empirical work supports hierarchical-inference models in general (Iglesias et al., 2013) and initial work is generally consistent with hierarchical alterations in schizophrenia (Diaconescu et al., 2014; Diaconescu et al., 2017; Haarsma et al., 2020a; Heinz et al., 2019; Henco et al., 2020; Sterzer et al., 2019), specific links to clinical delusions have been more elusive in this emerging literature (Cole et al., 2020; Diaconescu et al., 2019). Recent empirical studies inspired by generalized predictive-coding principles, however, hint at delusion-relevant hierarchical alterations. These studies investigated paranoid and persecutory ideation in the general population using tasks that manipulate volatility in underlying hidden states. Consistent with the notion of overweighting of prior beliefs at higher levels, these studies showed that more paranoid ideation was associated with overweighting of prior beliefs about volatility in non-social contexts (Reed et al., 2020) and overweighting of beliefs about the advice fidelity in social contexts (Diaconescu et al., 2020; Wellstein et al., 2020). More work is needed to probe this failure mode hypothesized to drive delusions, which given its hierarchical, state-dependent nature may require longitudinal investigations.

Some evidence supports circular inference in schizophrenia. In a probability-estimation version of a beads-like task with explicit cueing of prior information, patients with schizophrenia exhibited behaviors consistent with undercounting of prior beliefs and overcounting of sensory evidence compared to healthy controls (Jardri et al., 2017). Furthermore, the severity of delusional beliefs correlated with a fitted parameter reflecting bottom-up disinhibition. In principle, this result fits well with the predictions of the circular-inference model. However, its specificity to delusions versus other symptom dimensions like disorganization was less clear. One concern is that working-memory or general cognitive deficits likely interfered with the acquisition of prior knowledge, introducing variability in the formation of prior beliefs based on briefly presented visual cues (interindividual variability in working-memory performance indeed correlated with a prior weight parameter). Thus, it is not entirely clear that alterations in the relative weighting of prior beliefs and sensory evidence reported using this paradigm can be confidently attributed to alterations in the integration of this information—i.e., the inference process itself—or that a more general cognitive deficit interfering with its acquisition could be definitively ruled out. Notwithstanding, further testing of the failure modes proposed within the circular-inference framework, and contrasting these against those proposed under the generalized predictive-coding framework, would be a fruitful future direction.

One appealing aspect of the proposed failure mode for delusions is that it may complement a mechanistic explanation of hallucinations that has received growing empirical support: namely, that hallucinations result from overweighting of perceptual prior beliefs (Corlett et al., 2019). As implied by the definition of the psychotic syndrome, hallucinations and delusions typically co-occur and evolve in parallel. A parsimonious explanation of psychosis would thus invoke a common driver for these symptoms. However, these individual symptoms sometimes occur in isolation, suggesting the existence of symptom-specific pathways. This may be reconciled within the hierarchical-inference framework discussed above, which generally posits that inferential neural systems feature different but interdependent levels of processing. In this context, one possibility (Davies et al., 2018; Horga and Abi-Dargham, 2019) is that delusions and hallucinations result from similar algorithmic alterations occurring at different levels of the hierarchy supporting different computational goals. Both symptoms could be explained by a similar failure mode—i.e. overweighting of prior beliefs—with hallucinations arising from prior overweighting at lower hierarchical levels supporting inference on stimulus properties and delusions, in contrast, arising from prior overweighting at higher hierarchical levels supporting causal inference on hidden abstract states. This scenario would predict that hallucination severity should correlate preferentially with prior biases in perceptual tasks involving signal detection or magnitude estimation and delusion severity instead with prior biases in hidden-state inference tasks such as the beads task, consistent respectively with our prior behavioral work in hallucinations (Cassidy et al., 2018) and delusions (Baker et al., 2019). Critically, the interdependence between hierarchical levels inherent to this framework suggests that alterations at one level of the hierarchy may propagate to, or otherwise impact, other levels (Chaudhuri et al., 2015; Cicchini et al., 2020). Alternatively, partially shared elements within circuit motifs present at several levels may provide similar, although not necessarily identical, levels of susceptibility to common drivers (e.g., dopamine or glutamatergic dysfunction). Therefore, in principle this framework could readily accommodate the usual association of psychotic symptoms as well as their possible dissociation, for instance if differences in circuitry at specific levels (e.g., long-range connectivity or presence of certain cell populations) render them more susceptible or resilient than other levels. Examining neuroanatomical hierarchies of intrinsic neural timescales in fMRI data, we found initial support for this notion by showing that hallucinations and delusions correlate with distinct hierarchical alterations in the auditory and somatosensory systems (Wengler et al., 2020b).

Despite the valuable contribution of hierarchical-inference models to computational psychiatry, specific alterations in hierarchical inference linked selectively to delusions have not been conclusively established. Given this, and since the failure mode we have focused on—relative overweighting of high-level priors in causal inference on hidden states—can indeed be accommodated within the hierarchical-inference framework, we argue that this failure mode remains a top candidate the implementation of which is worth considering further.

Potential neurobiological implementations of prior weighting and delusions

To attain a holistic perspective on the merit of prior overweighting as a failure mode driving delusions, one must consider what is known about the neurobiological implementation of prior weighting in the brain and how it intersects with the pathophysiological substrates of delusions. Here we briefly discuss a selection of relevant neurobiological findings, starting with the pathophysiology of delusions.

The expression of psychosis and its response to antipsychotic treatment has long been linked to mesostriatal dopamine excess (Howes et al., 2012; Weinstein et al., 2017). Given the established role of phasic dopamine signals in associative learning (Glimcher, 2011; Schultz, 2016; Schultz et al., 1997), current theories posit that delusions result from disruptions in associative learning caused by aberrant dopamine signaling (Kapur, 2003). Such alterations, more typically framed in the context of reinforcement learning (Maia and Frank, 2011; Sterzer et al., 2018), are thought to drive unwarranted beliefs about the relevance or informativeness of neutral events and their bearing on causal inferences—sometimes referred to as salience misattribution (Fletcher and Frith, 2009; Heinz et al., 2019; Kapur, 2003; Sterzer et al., 2018)—and can thus be framed in the context of the type of inferential processes we have discussed so far (Fletcher and Frith, 2009). This parallels the growing appreciation of a broader role of phasic dopamine signals in updating of beliefs that go beyond reward expectations (Gershman and Uchida, 2019). Some empirical studies in delusional patients generally suggest alterations in inferential processes. For instance, in one such study delusional patients exhibited an attenuation of fMRI signals reflecting violation of expected outcomes acquired through associative learning in a region of right lateral prefrontal cortex (Corlett et al., 2007). Similar regions of anterior-lateral prefrontal cortex have been implicated in belief updating in health (Edelson et al., 2014; Fleming et al., 2018) and in the development of post-lesion delusions in a network-localization lesion study (Darby et al., 2017). This suggests that prefrontal circuits relevant to belief updating may be dysfunctional in delusional patients, but does not implicate dopamine. A recent study in healthy individuals provided more direct evidence for an involvement of dopamine in belief updating during an inference task (Nour et al., 2018). Here, molecular-imaging markers of striatal dopamine function correlated negatively with fMRI belief-updating signals in the striatum. In turn, decreased belief updating correlated with subclinical paranoid ideation, altogether providing feasibility for a model whereby excess striatal dopamine impairs inferential processes leading to delusional ideation. Despite many open questions, this literature broadly suggests that the pathophysiology of delusions involves mesostriatal dopamine excess and dysfunctions in prefrontal-striatal circuits supporting associative learning and inferential processes. Yet, the exact nature of the contributions from dopamine and different elements of this associative circuitry to delusions remain obscure. And so does their potential role in neurally instantiating prior weighting and its hypothesized alterations.

Some fMRI studies in health speak to plausible neural implementations of prior weighting. One study examined this by manipulating the consistency across sequential samples of evidence to induce more or less reliable prior knowledge (Vilares et al., 2012). By also manipulating and controlling the reliability of the likelihood within a trial, this work showed that fMRI activations in the striatum and in orbitofrontal parts of the prefrontal cortex specifically scaled with the reliability of prior knowledge. These activations correlated with behavioral weighting of prior beliefs in response to the statistics of the environment, suggesting a potential implementation of prior weighting in frontostriatal circuits. Other lines of work also suggest that prefrontal cortex and its interactions with parietal regions contribute to balancing the relative weight of prior beliefs and sensory evidence (Chambon et al., 2017; Flounders et al., 2019). Taken together, this suggests that fronto-parietal-striatal circuits may control the weight of prior beliefs in inference.

Electrophysiology and biophysical modeling have also shed light into the neuronal and circuit-level implementation of inferential processes similar to those we have discussed here. Many of these studies have used the “weather prediction task” (Knowlton et al., 1996 1996). Like the beads task, the weather prediction task probes behaviors relevant to inference on hidden states from a series of predictive samples (e.g., prediction of weather conditions, like a rainy day, A). But unlike the beads task, the likelihood associated with the samples of evidence is not explicitly instructed and needs to be learned through trial and error. Distinct samples provide different levels of evidence strength or likelihoods [e.g., $P(x|A) > P(y|A) > P(z|A)$] and participants need to infer the hidden state by iteratively updating their beliefs as they observe a sequence combining several distinct samples [e.g., $P(A|xyz)$]. Single-unit recordings from nonhuman primates performing a two-alternative-forced-choice version of this task revealed a neural substrate for sequential belief updating, which consisted of signals encoding the logit likelihood in a region of parietal association cortex (Kira et al., 2015; Yang and Shadlen, 2007).

A biophysical neural-network model was developed to recapitulate the neuronal and behavioral findings on this task and provide insights into a plausible circuit-level implementation (Soltani and Wang, 2010). Importantly, this model learned the expected value of each sample via simple Hebbian synaptic-plasticity rules like those involved in dopamine-dependent associative learning. As a result, synapses from neurons selective to specific samples that project onto expected-value neurons reflected the conditional probability of a state given that a specific sample appeared in the series [$\tilde{P}(A|x)$]. Using this *'naïve' posterior belief* as conservative proxy for the sample likelihood [$P(x|A)$], this model was able to infer hidden states. This biophysical model not only suggests plausible circuit mechanisms for approximate Bayesian inference but also for variability in prior weighting. Even though the model's architecture was determined by biophysically realistic principles, its behavior exhibited deviations from normative Bayesian inference similar to deviations in humans. Like humans, the model tended to underweight prior beliefs after a single sample and overweight priors in other circumstances where human participants tend to do so (Gluck and Bower, 1988; Soltani et al., 2016). This modeling thus suggests a potential dopamine-dependent synaptic mechanism for non-normative prior weighting in some forms of inference. Further modeling work is warranted to examine this intriguing mechanism, particularly in the context of the beads task and other online inference paradigms that do not require trial-and-error learning.

Altogether, this work suggests potential neurobiological substrates for changes in prior weighting that could implement the hypothesized inferential alterations behind delusions. Although much work is still needed in this area, one possibility is that dysregulated dopamine signals may disrupt inferential processes implemented in part in the striatum. Converging evidence also points to an involvement of higher-order prefrontal-parietal cortical regions that participate in inferential processes in health. Other brain regions and neuromodulatory systems involved in inference (e.g., norepinephrine) may be important candidates requiring further investigation. So far, however, an underlying substrate for prior overweighting in delusions remains unknown.

Conclusions and future directions

In this review, we have discussed inferential theories of delusions in psychosis and the empirical evidence favoring certain models and challenging others. Implicit in the notion of these inferential theories is that delusions result from narrow failure modes that should manifest as quantitative deviations from inferential biases common in health, not as broad deficits in neurocognition. Indeed, delusion severity tends to be uncorrelated with overall performance on standard neuropsychological tests (Baker et al., 2019; Keefe et al., 2006). And at least a subset of patients with schizophrenia do not exhibit obvious neuropsychological impairment, yet they still present with delusions and other symptoms of psychosis (Goldstein et al., 2005; Palmer et al., 1997). Likely in that group was John Nash, the Nobel laureate mathematician whose experiences marked the beginning of our review. By all accounts a brilliant logician, and a seminal contributor to the subject of game theory, Nash nonetheless suffered from severely disruptive and persistent delusions. In a famous exchange (Nasar, 1998), a colleague asked him, "How could you, a mathematician, a man devoted to reason and logical proof [...] believe that extraterrestrials are sending you messages?". To which Nash replied, "Because the ideas I had about supernatural beings came to me the same way my mathematical ideas did, so I took them seriously." As far as he was concerned, he arrived at his conclusions through logical reasoning; when he recovered, he even referred to his delusions in inference terms as "delusional hypotheses" (Nash, 1994). While anecdotal, the selective inferential alterations implied by his case suggest the need for similarly selective investigations to isolate the mechanisms of delusions in others.

Based on a critical review of the beads-task literature and theoretical considerations (Figs. 1 and 3), we have presented an argument against the utility of the classic beads task to isolate inferential processes. Our reading of the literature suggests there is insufficient evidence to conclude that the jumping-to-conclusions bias indicates an inferential alteration relevant to delusions. Instead, we take the literature to provide substantial support that this bias, and draws-to-decision behavior in the classic beads task more generally, mainly reflects general cognitive deficits or motivational factors rather than genuine alterations in inferential processes. The arguments we present caution against assuming that a specific relationship between the jumping-to-conclusions bias and clinical delusions has been established, or that such a presumed relationship supports an account of clinical delusions characterized by the overweighting of sensory evidence during inference. Further discussing other lines of work that may favor this interpretation (e.g., in subclinical populations or using other paradigms) is beyond the scope of this review; we simply contend here that invoking the beads-task literature in schizophrenia as direct support for this view is unwarranted.

We also describe enhanced approaches that show more promise in isolating delusion-specific inferential alterations. We focused on describing our novel approach combining a controlled paradigm and computational modeling, which has produced results pointing to a concrete failure mode in inference that is selectively associated with delusions: relative overweighting of prior beliefs. Through *in silico* simulations based on a weighted Bayesian model, we went on to show that this single failure mode can theoretically explain the two formal features that define delusional beliefs, namely their high certainty and rigidity (Figs. 2a-c). We also discussed possible extensions of this work based on prescriptive models that cast prior weighting as an adaptive response to external changes in the environment or internal constraints in information processing, suggesting that maladaptation to these conditions could explain the proposed failure mode. We then assessed the neurobiological intersections between the pathophysiology of delusions and the potential neural implementation of prior weighting during inferential processes. Despite our limited understanding, the available data support the biological plausibility of the proposed failure mode and hint at possible implementations at the system and circuit levels. Taking all this together, and drawing on early empirical support, we propose prior overweighting in causal inference as a parsimonious, and plausible, candidate failure mode for delusions. Future studies are needed to confirm and further investigate this mechanism, including its precise neural implementation. To this end, we offer several future directions which we believe will be fruitful avenues for deepening our neurocomputational understanding of delusions.

First, we believe there is room for further improvements in experimental paradigms, which we take as perhaps the most critical aspect of future work. Incentive compatibility is thought to contribute to the high replicability of economics paradigms, encouraging the reporting of true preferences and beliefs (Camerer, 1997; Camerer and Mobbs, 2017; Camerer et al., 2016). This feature may be critical for future belief-elicitation paradigms trying to isolate inference in delusions, in line with previous reports (van der Leer and McKay, 2014). Second, we believe that the independent replication of key behavioral and modeling results, comparisons across paradigms and models, and the confirmation of specific associations with delusions will be necessary to establish a solid foundation for further work (including backtranslation, causal investigations, and forward translation towards treatment development). Although our simulations indicate that the proposed failure mode for delusions could parsimoniously explain the gradual development and maintenance of delusional beliefs (Fig. 2a), an important milestone will be to show whether this prior overweighting is indeed associated with attenuated delusions in psychosis high-risk populations, and whether the evolution of this computational phenotype predicts clinical trajectories. If alterations in higher-level inferences on hidden causal states are indeed confirmed to be specific to delusions, and computationally distinct (albeit algorithmically similar) from lower-level inferential alterations linked to hallucinations (Horga and Abi-Dargham, 2019; Wengler et al., 2020a), that would lend further support for hierarchical frameworks with potential to provide an integrative understanding of psychosis as a whole. Third, connecting the proposed algorithmic mechanisms to underlying biological implementations will lend further support for their feasibility and provide targets for interventions. Given that inputs from different hierarchical levels are thought to segregate into specific cortical layers within a brain region (Lawrence et al., 2019; Stephan et al., 2019), new layer-specific, high-resolution fMRI techniques (de Hollander et al., 2021; Haarsma et al., 2020b) may be a promising avenue in this regard (for further discussion, see Haarsma et al., 2020b).

Specific alterations in social inferences and social cognition have also been proposed to underlie paranoid ideation and delusions (Bell et al., 2020; Diaconescu et al., 2019; Diaconescu et al., 2020;

Wellstein et al., 2020), as well as schizophrenia more generally (Henco et al., 2020; Patel et al., 2020). The link to delusions seems at odds with our findings in Baker et al. (2019), including the strong correlation between paranoid delusions and prior overweighting in a non-social, emotionally neutral context, and with other recent findings in paranoid ideation using a similarly neutral reversal-learning task (Reed et al., 2020). As noted by Diaconescu et al.(2020; 2019), the open question here is whether delusions result from basic inferential alterations that manifest in generally ambiguous contexts (like social situations), or whether they result specifically from alterations in social inference. Direct comparisons of social and non-social inference in delusional patients would help settle this debate. Finally, once abnormalities in inferences governing the form of delusional beliefs are identified, a comprehensive model of delusions can and should aspire to address the thematic content of delusions. Despite the issues we have raised about the content of delusions, focusing on the more consistent and tractable aspects of their content may help elucidate the overrepresentation of delusional themes with negative emotional valence (Appelbaum et al., 1999; Sharot and Garrett, 2016; Woodward et al., 2014). Moving beyond the specter of the jumping-to-conclusions bias and pursuing the goals set out above may yet transform our understanding of delusions, and bring us ever closer to a comprehensive, computational model of this enigmatic symptom.

Acknowledgments: The authors would like to express our respect and gratitude to the memory of John Nash, whose story we invoke in the pursuit of a greater scientific understanding of delusions. This work was supported by National Institute of Mental Health Grant Nos. R01MH117323 and R01MH114965 (to G.H.). B.K.A. was supported by a T32 Postdoctoral Fellowship (T32-MH018870) at Columbia University.

Figure captions

Figure 1. Distinct, nested processes linking inference and sampling decisions in the POMDP framework. For (a) a sequence of observed samples (grayed-out samples reflect future samples that the agent never sees), this instantiation of the POMDP model shows (b) the logit posterior beliefs of the ideal Bayesian observer ($\omega_1 = \omega_2 = 1$) after each sample and (c) the difference in expected value between the best guess (the guess associated with the jar that has highest expected value) and drawing another sample. (d) A stopping decision is made when the expected value of the best guess is higher than the expected value of drawing another sample, i.e., the first point at which the difference in expected values is above 0. This point represents the optimal draws-to-decision (DTD). Note that it takes the optimal agent 6 samples (draws) to reach the stopping point based on valuation, even though the exact same level of

belief certainty was achieved after only 4 samples (draws). This illustrates that DTD is depends on value-related factors beyond inference. The simulation uses cost parameters (starting endowment of \$30; \$0 for a correct response; -\$15 for an incorrect response; -\$0.30 for a draw) consistent with the experimental parameters from Baker et al. (2019).

Figure 2. Dynamic effects of prior weighting on inference and relevance to the form of delusions. (a) Long-term trajectory of beliefs with respect to a black jar (in probability space) for two agents (higher ω_1 = 0.995; lower ω_1= 0.950; $\omega_2 = 1$ for both agents) over 450 randomly selected samples (with replacement) in the beads task. Here, and in general, please note that parameter values were selected to illustrate the belief-updating effects highlighted in the main text. The correct (black) jar has a ratio of 55 black beads to 45 white beads, reflecting an ambiguous situation of weak sensory evidence (likelihood of 0.55). This simulation illustrates an ω_1-driven rigidity effect, whereby the beliefs of the higher-ω_1 agent take more disconfirmatory samples to return to an uncertain level, and a concomitant certainty effect, whereby its beliefs tend to be more certain, relative to the lower-ω_1 agent. (b) Long-term trajectory of beliefs with respect to a black jar (in probability space) for two agents (higher-ω_2 = 1; lower-ω_2= 0.40; $\omega_1 = .95$ for both agents) over the same 450 randomly selected samples in (a) in the beads task. For reference, the higher-ω_2 agent in (b) is identical to the lower-ω_1 agent in (a). Changes in ω_2 induce a certainty effect, i.e., the higher-ω_2 agent tends to reach more certain beliefs than the lower-ω_2 agent, but has no effect on belief rigidity. (c, d, e) Simulations illustrating local belief-updating dynamics over 5 samples for a (c) lower-ω_1 agent (ω_1= 0.70; $\omega_2 = 1$; similar to healthy individuals in Baker et al.), a (d) higher-ω_1 agent (ω_1= 0.98; $\omega_2 = 1$; consistent with delusional patients in Baker et al.), and a (e) lower-ω_2 agent (ω_1= 0.70; $\omega_2 = 0.40$). The dotted diagonal lines depict the “leak” of logit prior beliefs and their endpoints indicate the value of the weighted prior for the next belief update. The solid horizontal line is a reference to indicate the value of the unweighted prior. Thus, the distance between the solid line and the dotted line reflects the magnitude of the prior leak for each update. The dashed vertical lines reflect the contribution of the logit likelihood (LLR) to the belief update. It is apparent in (a) that for lower-ω_1 agents, prior beliefs “leak” more, gradually decreasing the magnitude of belief updates over samples leading to relatively less certain and less rigid beliefs; and (b) shows that these effects are attenuated for higher-ω_1 agents, leading to relatively more certain and more rigid beliefs. Comparing (a) and (c) highlights that differences in ω_2 only scale belief certainty and do not affect belief rigidity.

Figure 3. Evidence-order effects on belief updating and draws-to-decision under the weighted Bayesian model. (a, b) Simulation of logit posterior beliefs favoring the black jar for a higher-ω_1 agent (ω_1 = 0.98) and a lower-ω_1 agent (ω_1 = 0.70) in two sequences. In (a) evidence favoring the black jar (the correct jar) occurs earlier in the sequence, and the higher-ω_1 agent generally exhibits more certain beliefs than the lower ω_1 agent that the majority black jar is the correct jar. In (b) evidence favoring the black jar occurs later in the sequence, and the higher-ω_1 agent instead exhibits less certain beliefs than the lower-

ω_1 agent. Note that parameters were selected to visually exaggerate the effects of interest, although their generality is addressed in the main text. (c) Simulations for various sequence orders including the same samples of evidence show order-dependent differences in beliefs (in probability space) on a sample-by-sample basis between a higher-ω_1 (ω_1 = 0.98; similar to delusional patients in Baker et al.) and a lower-ω_1 agent (ω_1= 0.89; ω_2 = 1 for all simulations). Positive values (shades of red) in the heatmap indicate that the higher-ω_1 agent exhibits more certain beliefs than the lower-ω_1 agent that the black jar was the correct jar, and negative values (shades of blue) indicate that the lower-ω_1 agent was more certain. (d, e) Simulations of the POMDP valuation process comparing two agents (the same agents from 3c) across different sequences to illustrate how evidence order affects sampling (draws-to-decision) behavior. The remaining POMPD parameters are equivalent to those in Fig. 1 except for the cost of drawing a bead (here $0.10 instead of $0.30 for illustrative purposes). Note that DTD differences between the two agents are opposite between the two sequences. The asterisk in d indicates the point at which the lower-ω_1 agent is forced to make a guess because the maximum number of samples is 8 (as in the Baker et al. task).

References

Adams, R.A., 2018. Chapter 7 - Bayesian Inference, Predictive Coding, and Computational Models of Psychosis, in: Anticevic, A., Murray, J.D. (Eds.), Computational Psychiatry. Academic Press, pp. 175-195.

Adams, R.A., Brown, H.R., Friston, K.J., 2014. Bayesian inference, predictive coding and delusions. AVANT V(3), 51-88.

Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D., Friston, K.J., 2013. The Computational Anatomy of Psychosis. Front. Psychiatry 4.

Aller, M., Noppeney, U., 2019. To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. PLOS Biology 17(4), e3000210.

Ambuehl, S., Li, S., 2018. Belief updating and the demand for information. Games and Economic Behavior 109, 21-39.

American Psychological Association, 1980. Diagnostic and Statistical Manual of Mental Disorders, 3rd ed.

American Psychological Association, 2013. Diagnostic and statistical manual of mental disorders, 5th ed. American Psychiatric Association.

Andreou, C., Moritz, S., Veith, K., Veckenstedt, R., Naber, D., 2014. Dopaminergic Modulation of Probabilistic Reasoning and Overconfidence in Errors: A Double-Blind Study. Schizophrenia Bulletin 40(3), 558-565.

Andreou, C., Schneider, B.C., Balzan, R., Luedecke, D., Roesch-Ely, D., Moritz, S., 2015. Neurocognitive deficits are relevant for the jumping-to-conclusions bias, but not for delusions: A longitudinal study. Schizophr Res Cogn 2(1), 8-11.

Appelbaum, P.S., Robbins, P.C., Roth, L.H., 1999. Dimensional Approach to Delusions: Comparison Across Types and Diagnoses. American Journal of Psychiatry 156(12), 1938-1943.

Aschebrock, Y., Gavey, N., McCreanor, T., Tippett, L., 2003. Is the Content of Delusions and Hallucinations Important? Australas Psychiatry 11(3), 306-311.

Averbeck, B.B., 2015. Theory of Choice in Bandit, Information Sampling and Foraging Tasks. PLOS Computational Biology 11(3), e1004164.

Azeredo da Silveira, R., Woodford, M., 2019. Noisy Memory and Over-Reaction to News. AEA Papers and Proceedings 109, 557-561.

Baker, S.C., Konova, A.B., Daw, N.D., Horga, G., 2019. A distinct inferential mechanism for delusions in schizophrenia. Brain 142(6), 1797-1812.

Balzan, R., Delfabbro, P., Galletly, C., 2012a. Delusion-proneness or miscomprehension? A re-examination of the jumping-to-conclusions bias. Australian Journal of Psychology 64(2), 100-107.

Balzan, R., Delfabbro, P., Galletly, C., Woodward, T., 2012b. Over-adjustment or miscomprehension? A re-examination of the jumping to conclusions bias. Aust N Z J Psychiatry 46(6), 532-540.

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.

Bar-Hillel, M., 1980. The base-rate fallacy in probability judgments. Acta Psychologica 44(3), 211-233.

Bell, V., Raihani, N., Wilkinson, S., 2020. Derationalizing Delusions. Clinical Psychological Science 0(0), 2167702620951553.

Ben-Zeev, D., Morris, S., Swendsen, J., Granholm, E., 2012. Predicting the Occurrence, Conviction, Distress, and Disruption of Different Delusional Experiences in the Daily Life of People with Schizophrenia. Schizophrenia Bulletin 38(4), 826-837.

Benjamin, D., Bodoh-Creed, A., Rabin, M., 2019. Base-Rate Neglect: Foundations and Implications. 62.

Benjamin, D.J., 2019. Errors in probabilistic reasoning and judgment biases, Handbook of Behavioral Economics: Applications and Foundations 1. Elsevier, pp. 69-186.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J.D., 2006. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review 113(4), 700-765.

Bornstein, A.M., Aly, M., Feng, S.F., Turk-Browne, N.B., Norman, K.A., Cohen, J.D., 2018. Perceptual decisions result from the continuous accumulation of memory and sensory evidence. Neuroscience.

Brenner, C.J., Ben-Zeev, D., 2014. Affective forecasting in schizophrenia: Comparing predictions to real-time Ecological Momentary Assessment (EMA) ratings. Psychiatric Rehabilitation Journal 37(4), 316-320.

Broome, M.R., Johns, L.C., Valli, I., Woolley, J.B., Tabraham, P., Brett, C., Valmaggia, L., Peters, E., Garety, P.A., McGuire, P.K., 2007. Delusion formation and reasoning biases in those at clinicalhigh risk for psychosis. The British Journal of Psychiatry 191(S51), s38-s42.

Busemeyer, J.R., Townsend, J.T., 1993. Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. 28.

Camerer, C., 1997. Rules for Experimenting in Psychology and Economics, and Why They Differ, in: Albers, W., Güth, W., Hammerstein, P., Moldovanu, B., van Damme, E. (Eds.), Understanding Strategic Interaction: Essays in Honor of Reinhard Selten. Springer, Berlin, Heidelberg, pp. 313-327.

Camerer, C., Mobbs, D., 2017. Differences in Behavior and Brain Activity during Hypothetical and Real Choices. Trends in Cognitive Sciences 21(1), 46-56.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351(6280), 1433-1436.

Cassidy, C.M., Balsam, P.D., Weinstein, J.J., Rosengard, R.J., Slifstein, M., Daw, N.D., Abi-Dargham, A., Horga, G., 2018. A Perceptual Inference Mechanism for Hallucinations Linked to Striatal Dopamine. Current Biology 28(4), 503-514.e504.

Catalano, L.T., Heerey, E.A., Gold, J.M., 2018. The valuation of social rewards in schizophrenia. Journal of Abnormal Psychology 127(6), 602-611.

Chambon, V., Domenech, P., Jacquet, P.O., Barbalat, G., Bouton, S., Pacherie, E., Koechlin, E., Farrer, C., 2017. Neural coding of prior expectations in hierarchical intention inference. Scientific Reports 7(1), 1278.

Chambon, V., Domenech, P., Pacherie, E., Koechlin, E., Baraduc, P., Farrer, C., 2011a. What Are They Up To? The Role of Sensory Evidence and Prior Knowledge in Action Understanding. PLOS ONE 6(2), e17133.

Chambon, V., Pacherie, E., Barbalat, G., Jacquet, P., Franck, N., Farrer, C., 2011b. Mentalizing under influence: abnormal dependence on prior expectations in patients with schizophrenia. Brain 134(12), 3728-3741.

Chang, W.C., Westbrook, A., Strauss, G.P., Chu, A.O.K., Chong, C.S.Y., Siu, C.M.W., Chan, S.K.W., Lee, E.H.M., Hui, C.L.M., Suen, Y.M., Lo, T.L., Chen, E.Y.H., 2020. Abnormal cognitive effort allocation and its association with amotivation in first-episode psychosis. Psychol Med 50(15), 2599-2609.

Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H., Wang, X.-J., 2015. A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. Neuron 88(2), 419-431.

Cicchini, G.M., Benedetto, A., Burr, D.C., 2020. Perceptual history propagates down to early levels of sensory analysis. Current Biology.

Cole, D.M., Diaconescu, A.O., Pfeiffer, U.J., Brodersen, K.H., Mathys, C.D., Julkowski, D., Ruhrmann, S., Schilbach, L., Tittgemeyer, M., Vogeley, K., Stephan, K.E., 2020. Atypical processing of uncertainty in individuals at risk for psychosis. NeuroImage: Clinical 26, 102239.

Coltheart, M., Menzies, P., Sutton, J., 2010. Abductive inference and delusional belief. Cognitive Neuropsychiatry 15(1), 261-287.

Cools, R., 2019. Chemistry of the Adaptive Mind: Lessons from Dopamine. Neuron 104(1), 113-131.

Corlett, P., Krystal, J., Taylor, J., Fletcher, P., 2009. Why do delusions persist? Frontiers in Human Neuroscience 3, 12.

Corlett, P.R., Horga, G., Fletcher, P.C., Alderson-Day, B., Schmack, K., Powers, A.R., 2019. Hallucinations and Strong Priors. Trends in Cognitive Sciences 23(2), 114-127.

Corlett, P.R., Murray, G.K., Honey, G.D., Aitken, M.R.F., Shanks, D.R., Robbins, T.W., Bullmore, E.T., Dickinson, A., Fletcher, P.C., 2007. Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. Brain 130(0 9).

Corlett, P.R., Taylor, J.R., Wang, X.J., Fletcher, P.C., Krystal, J.H., 2010. Toward a neurobiology of delusions. Progress in Neurobiology 92(3), 345-369.

Darby, R.R., Laganiere, S., Pascual-Leone, A., Prasad, S., Fox, M.D., 2017. Finding the imposter: brain connectivity of lesions causing delusional misidentifications. Brain 140(2), 497-507.

Davies, D.J., Teufel, C., Fletcher, P.C., 2018. Anomalous Perceptions and Beliefs Are Associated With Shifts Toward Different Types of Prior Knowledge in Perceptual Inference. Schizophrenia Bulletin 44(6), 1245-1253.

Daw, N.D., 2014. Chapter 16 - Advanced Reinforcement Learning, in: Glimcher, P.W., Fehr, E. (Eds.), Neuroeconomics (Second Edition). Academic Press, San Diego, pp. 299-320.

de Hollander, G., van der Zwaag, W., Qian, C., Zhang, P., Knapen, T., 2021. Ultra-high field fMRI reveals origins of feedforward and feedback activity within laminae of human ocular dominance columns. NeuroImage 228, 117683.

Denève, S., Jardri, R., 2016. Circular inference: mistaken belief, misplaced trust. Current Opinion in Behavioral Sciences 11, 40-48.

Diaconescu, A.O., Hauke, D.J., Borgwardt, S., 2019. Models of persecutory delusions: a mechanistic insight into the early stages of psychosis. Molecular Psychiatry 24(9), 1258-1267.

Diaconescu, A.O., Mathys, C., Weber, L.A.E., Daunizeau, J., Kasper, L., Lomakina, E.I., Fehr, E., Stephan, K.E., 2014. Inferring on the Intentions of Others by Hierarchical Bayesian Learning. PLOS Computational Biology 10(9), e1003810.

Diaconescu, A.O., Mathys, C., Weber, L.A.E., Kasper, L., Mauer, J., Stephan, K.E., 2017. Hierarchical prediction errors in midbrain and septum during social learning. Soc Cogn Affect Neurosci 12(4), 618-634.

Diaconescu, A.O., Wellstein, K.V., Kasper, L., Mathys, C., Stephan, K.E., 2020. Hierarchical Bayesian models of social inference for probing persecutory delusional ideation. Journal of Abnormal Psychology 129(6), 556-569.

Diederen, K.M.J., Fletcher, P.C., 2020. Dopamine, Prediction Error and Beyond. The Neuroscientist, 1073858420907591.

Dudley, R., Taylor, P., Wickham, S., Hutton, P., 2016. Psychosis, Delusions and the “Jumping to Conclusions” Reasoning Bias: A Systematic Review and Meta-analysis. Schizophrenia Bulletin 42(3), 652-665.

Dudley, R.E.J., John, C.H., Young, A.W., Over, D.E., 1997a. The effect of self-referent material on the reasoning of people with delusions. British Journal of Clinical Psychology 36(4), 575-584.

Dudley, R.E.J., John, C.H., Young, A.W., Over, D.E., 1997b. Normal and abnormal reasoning in people with delusions. British Journal of Clinical Psychology 36(2), 243-258.

Edelson, M.G., Dudai, Y., Dolan, R.J., Sharot, T., 2014. Brain Substrates of Recovery from Misleading Influence. Journal of Neuroscience 34(23), 7744-7753.

Enke, B., Graeber, T., 2019. Cognitive Uncertainty. Social Science Research Network, Rochester, NY.

Ermakova, A.O., Gileadi, N., Knolle, F., Justicia, A., Anderson, R., Fletcher, P.C., Moutoussis, M., Murray, G.K., 2019. Cost Evaluation During Decision-Making in Patients at Early Stages of Psychosis. Computational Psychiatry 3, 18-39.

Falcone, M.A., Murray, R.M., Wiffen, B.D., O'Connor, J.A., Russo, M., Kolliakou, A., Stilo, S., Taylor, H., Gardner-Sood, P., Paparelli, A., Jichi, F., Di Forti, M., David, A.S., Freeman, D., Jolley, S., 2015. Jumping to conclusions, neuropsychological functioning, and delusional beliefs in first episode psychosis. Schizophr Bull 41(2), 411-418.

Fetsch, C.R., Pouget, A., DeAngelis, G.C., Angelaki, D.E., 2012. Neural correlates of reliability-based cue weighting during multisensory integration. Nature Neuroscience 15(1), 146-154.

Fett, A.-K.J., Mouchlianitis, E., Gromann, P.M., Vanes, L., Shergill, S.S., Krabbendam, L., 2019. The neural mechanisms of social reward in early psychosis. Soc Cogn Affect Neurosci 14(8), 861-870.

Fine, C., Gardner, M., Craigie, J., Gold, I., 2007. Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. Cognitive Neuropsychiatry 12(1), 46-77.

Fioravanti, M., Carlone, O., Vitale, B., Cinti, M.E., Clare, L., 2005. A Meta-Analysis of Cognitive Deficits in Adults with a Diagnosis of Schizophrenia. Neuropsychol Rev 15(2), 73-95.

Fischhoff, B., Beyth-Marom, R., 1983. Hypothesis evaluation from a Bayesian perspective. Psychological Review 90(3), 239-260.

Fleming, S.M., van der Putten, E.J., Daw, N.D., 2018. Neural mediators of changes of mind about perceptual decisions. Nature Neuroscience 21(4), 617-624.

Fletcher, P.C., Frith, C.D., 2009. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nature Reviews Neuroscience 10(1), 48.

Flounders, M.W., González-García, C., Hardstone, R., He, B.J., 2019. Neural dynamics of visual ambiguity resolution by perceptual prior. eLife 8, e41861.

Forbes, N.F., Carrick, L.A., McIntosh, A.M., Lawrie, S.M., 2009. Working memory in schizophrenia: a meta-analysis. Psychological Medicine 39(6), 889-905.

Freeman, D., Startup, H., Dunn, G., Černis, E., Wingham, G., Pugh, K., Cordwell, J., Mander, H., Kingdon, D., 2014. Understanding jumping to conclusions in patients with persecutory delusions: working memory and intolerance of uncertainty. Psychological Medicine, 3017-3024.

French, R.L., DeAngelis, G.C., 2020. Multisensory neural processing: from cue integration to causal inference. Current Opinion in Physiology 16, 8-13.

Friston, K., 2008. Hierarchical Models in the Brain. PLOS Computational Biology 4(11), e1000211.

Friston, K., 2010. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 11(2), 127-138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., 2016. Active Inference: A Process Theory. Neural Computation 29(1), 1-49.

Furl, N., Averbeck, B.B., 2011. Parietal Cortex and Insula Relate to Evidence Seeking Relevant to Reward-Related Decisions. Journal of Neuroscience 31(48), 17572-17582.

Gaines, A.D., 1995. Culture-Specific Delusions: Sense and Nonsense in Cultural Context. Psychiatric Clinics of North America 18(2), 281-301.

Garety, P., Joyce, E., Jolley, S., Emsley, R., Waller, H., Kuipers, E., Bebbington, P., Fowler, D., Dunn, G., Freeman, D., 2013. Neuropsychological functioning and jumping to conclusions in delusions. Schizophr Res 150(2-3), 570-574.

Gershman, S.J., Uchida, N., 2019. Believing in dopamine. Nature Reviews Neuroscience 20(11), 703-714.

Glaze, C.M., Kable, J.W., Gold, J.I., 2015. Normative evidence accumulation in unpredictable environments. eLife 4, e08825.

Glimcher, P.W., 2011. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences 108(Supplement 3), 15647-15654.

Glimcher, P.W., Rustichini, A., 2004. Neuroeconomics: The Consilience of Brain and Decision. Science 306(5695), 447-452.

Gluck, M.A., Bower, G.H., 1988. From conditioning to category learning: an adaptive network model. J Exp Psychol Gen 117(3), 227-247.

Gold, J., Gold, I., 2012. The “Truman Show” delusion: Psychosis in the global village. Cognitive Neuropsychiatry 17(6), 455-472.

Gold, J.I., Shadlen, M.N., 2007. The Neural Basis of Decision Making. Annual Review of Neuroscience 30(1), 535-574.

Gold, J.M., Waltz, J.A., Prentice, K.J., Morris, S.E., Heerey, E.A., 2008. Reward processing in schizophrenia: A deficit in the representation of value. Schizophrenia Bulletin 34(5), 835-847.

Goldberg, S., Fruchter, E., Davidson, M., Reichenberg, A., Yoffe, R., Weiser, M., 2011. The relationship between risk of hospitalization for schizophrenia, SES, and cognitive functioning. Schizophr Bull 37(4), 664-670.

Goldstein, G., Shemansky, W.J., Allen, D.N., 2005. Cognitive function in schizoaffective disorder and clinical subtypes of schizophrenia. Archives of Clinical Neuropsychology 20(2), 153-159.

González, L.E., López-Carrilero, R., Barrigón, M.L., Grasa, E., Barajas, A., Pousa, E., González-Higueras, F., Ruiz-Delgado, I., Cid, J., Lorente-Rovira, E., Pélaez, T., Ochoa, S., 2018. Neuropsychological functioning and jumping to conclusions in recent onset psychosis patients. Schizophrenia Research 195, 366-371.

Gonzalez, R., Wu, G., 1999. On the Shape of the Probability Weighting Function. Cognitive Psychology 38(1), 129-166.

Granholm, E., Holden, J.L., Mikhael, T., Link, P.C., Swendsen, J., Depp, C., Moore, R.C., Harvey, P.D., 2020. What Do People With Schizophrenia Do All Day? Ecological Momentary Assessment of Real-World Functioning in Schizophrenia. Schizophrenia Bulletin 46(2), 242-251.

Green, M.F., Hellemann, G., Horan, W.P., Lee, J., Wynn, J.K., 2012. From Perception to Functional Outcome in Schizophrenia: Modeling the Role of Ability and Motivation. Archives of General Psychiatry 69(12), 1216-1224.

Grether, D.M., 1980. Bayes Rule as a Descriptive Model: The Representativeness Heuristic. The Quarterly Journal of Economics 95(3), 537-557.

Grether, D.M., 1992. Testing bayes rule and the representativeness heuristic: Some experimental evidence. Journal of Economic Behavior & Organization 17(1), 31-57.

Griffiths, O., Balzan, R., 2020. EXPRESS: Schizotypy is associated with difficulty maintaining multiple hypotheses. Q J Exp Psychol (Hove), 1747021820982256.

Guo, J.Y., Ragland, J.D., Carter, C.S., 2019. Memory and cognition in schizophrenia. Molecular Psychiatry 24(5), 633-642.

Haarsma, J., Knolle, F., Griffin, J.D., Taverne, H., Mada, M., Goodyer, I.M., The, N.C., Fletcher, P.C., Murray, G.K., 2020a. Influence of prior beliefs on perception in early psychosis: Effects of illness stage and hierarchical level of belief. Journal of Abnormal Psychology 129(6), 581-598.

Haarsma, J., Kok, P., Browning, M., 2020b. The promise of layer-specific neuroimaging for testing predictive coding theories of psychosis. Schizophrenia Research.

Habtewold, T.D., Rodijk, L.H., Liemburg, E.J., Sidorenkov, G., Boezen, H.M., Bruggeman, R., Alizadeh, B.Z., 2020. A systematic review and narrative synthesis of data-driven studies in schizophrenia symptoms and cognitive deficits. Translational Psychiatry 10(1), 1-24.

Haefner, Ralf M., Berkes, P., Fiser, J., 2016. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron 90(3), 649-660.

Hakulinen, C., Webb, R.T., Pedersen, C.B., Agerbo, E., Mok, P.L.H., 2020. Association Between Parental Income During Childhood and Risk of Schizophrenia Later in Life. JAMA Psychiatry 77(1), 17-24.

Hartmann-Riemer, M., Kirschner, M., Kaiser, S., 2018. Effort-based decision-making paradigms as objective measures of apathy in schizophrenia? Current Opinion in Behavioral Sciences 22, 70-75.

Heinz, A., Murray, G.K., Schlagenhauf, F., Sterzer, P., Grace, A.A., Waltz, J.A., 2019. Towards a Unifying Cognitive, Neurophysiological, and Computational Neuroscience Account of Schizophrenia. Schizophrenia Bulletin 45(5), 1092-1100.

Heinze, K., Lin, A., Nelson, B., Reniers, R.L.E.P., Upthegrove, R., Clarke, L., Roche, A., Lowrie, A., Wood, S.J., 2018. The impact of psychotic experiences in the early stages of mental health problems in young people. BMC Psychiatry 18(1), 214.

Hemsley, D.R., Garety, P.A., 1986. The Formation of Maintenance of Delusions: a Bayesian Analysis. Br J Psychiatry 149(1), 51-56.

Henco, L., Diaconescu, A.O., Lahnakoski, J.M., Brandi, M.-L., Hörmann, S., Hennings, J., Hasan, A., Papazova, I., Strube, W., Bolis, D., Schilbach, L., Mathys, C., 2020. Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. PLOS Computational Biology 16(9), e1008162.

Heng, J.A., Woodford, M., Polania, R., 2020. Efficient sampling and noisy decisions. eLife 9, e54962.

Horga, G., Abi-Dargham, A., 2019. An integrative framework for perceptual disturbances in psychosis. Nature Reviews Neuroscience 20(12), 763-778.

Howes, O.D., Kambeitz, J., Kim, E., Stahl, D., Slifstein, M., Abi-Dargham, A., Kapur, S., 2012. The nature of dopamine dysfunction in schizophrenia and what this means for treatment. Archives of General Psychiatry 69(8), 776-786.

Hoyer, P.O., Hyvärinen, A., 2003. Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior, in: Becker, S., Thrun, S., Obermayer, K. (Eds.), Advances in Neural Information Processing Systems 15. MIT Press, pp. 293-300.

Huang, H., Thompson, W., Paulus, M.P., 2017. Computational Dysfunctions in Anxiety: Failure to Differentiate Signal From Noise. Biological Psychiatry 82(6), 440-446.

Hudson, C.G., 2005. Socioeconomic Status and Mental Illness: Tests of the Social Causation and Selection Hypotheses. American Journal of Orthopsychiatry 75(1), 3-18.

Huq, S.F., Garety, P.A., Hemsley, D.R., 1988. Probabilistic judgements in deluded and non-deluded subjects. The Quarterly Journal of Experimental Psychology Section A 40(4), 801-812.

Iglesias, S., Mathys, C., Brodersen, Kay H., Kasper, L., Piccirelli, M., den Ouden, Hanneke E.M., Stephan, Klaas E., 2013. Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. Neuron 80(2), 519-530.

Jardri, R., Denève, S., 2013. Circular inferences in schizophrenia. Brain 136(11), 3227-3241.

Jardri, R., Duverne, S., Litvinova, A.S., Denève, S., 2017. Experimental evidence for circular inference in schizophrenia. Nature Communications 8, 14218.

Jaspers, K., 1913. General Psychopathology, translated by Hoenig J and Hamilton, MW. University Press, Manchester, England.

Kaelbling, L.P., Littman, M.L., Cassandra, A.R., 1998. Planning and acting in partially observable stochastic domains. Artificial Intelligence 101(1-2), 99-134.

Kahneman, D., Tversky, A., 1973. On the psychology of prediction. Psychological Review 80(4), 237-251.

Kapur, S., 2003. Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. The American Journal of Psychiatry 160(1), 13-23.

Keefe, R.S.E., Bilder, R.M., Harvey, P.D., Davis, S.M., Palmer, B.W., Gold, J.M., Meltzer, H.Y., Green, M.F., Miller, D.D., Canive, J.M., Adler, L.W., Manschreck, T.C., Swartz, M., Rosenheck, R., Perkins, D.O., Walker, T.M., Stroup, T.S., McEvoy, J.P., Lieberman, J.A., 2006. Baseline neurocognitive deficits in the CATIE schizophrenia trial. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology 31(9), 2033-2046.

Kira, S., Yang, T., Shadlen, M.N., 2015. A neural implementation of Wald's sequential probability ratio test. Neuron 85(4), 861-873.

Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27(12), 712-719.

Knowlton, B.J., Mangels, J.A., Squire, L.R., 1996. A Neostriatal Habit Learning System in Humans. Science 273(5280), 1399-1402.

Kreis, I., Moritz, S., Pfuhl, G., 2020. Objective Versus Subjective Effort in Schizophrenia. Frontiers in Psychology 11(1469).

Lawrence, S.J.D., Formisano, E., Muckli, L., de Lange, F.P., 2019. Laminar fMRI: Applications for cognitive neuroscience. NeuroImage 197, 785-791.

Lawson, R.P., Mathys, C., Rees, G., 2017. Adults with autism overestimate the volatility of the sensory environment. Nature Neuroscience 20(9), 1293-1299.

Lee, J., Jimenez, A.M., Reavis, E.A., Horan, W.P., Wynn, J.K., Green, M.F., 2018. Reduced Neural Sensitivity to Social vs Nonsocial Reward in Schizophrenia. Schizophrenia Bulletin 45(3), 620-628.

Leptourgos, P., Denève, S., Jardri, R., 2017. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? Current Opinion in Neurobiology 46, 154-161.

Lincoln, T.M., Lange, J., Burau, J., Exner, C., Moritz, S., 2010a. The effect of state anxiety on paranoid ideation and jumping to conclusions. An experimental investigation. Schizophr Bull 36(6), 1140-1148.

Lincoln, T.M., Ziegler, M., Mehl, S., Rief, W., 2010b. The jumping to conclusions bias in delusions: Specificity and changeability. Journal of Abnormal Psychology 119(1), 40-49.

Luck, S.J., Hahn, B., Leonard, C.J., Gold, J.M., 2019. The Hyperfocusing Hypothesis: A New Account of Cognitive Dysfunction in Schizophrenia. Schizophrenia Bulletin.

Maia, T.V., Frank, M.J., 2011. From reinforcement learning models to psychiatric and neurological disorders. Nature Neuroscience 14(2), 154-162.

Mathys, C., 2011. A Bayesian foundation for individual learning under uncertainty. Frontiers in Human Neuroscience 5.

McLean, B.F., Balzan, R.P., Mattiske, J.K., 2020a. Jumping to conclusions in the less-delusion-prone? Further evidence from a more reliable beads task. Consciousness and Cognition 83, 102956.

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(2), 344-354.

McLean, B.F., Mattiske, J.K., Balzan, R.P., 2018. Towards a reliable repeated-measures beads task for assessing the jumping to conclusions bias. Psychiatry Research 265, 200-207.

McLean, B.F., Mattiske, J.K., Balzan, R.P., 2020b. Jumping to conclusions in the less-delusion-prone? Preliminary evidence from a more reliable beads task. Journal of Behavior Therapy and Experimental Psychiatry 68, 101562.

Moritz, S., Göritz, A.S., Balzan, R.P., Gawęda, Ł., Kulagin, S.C., Andreou, C., 2017. A new paradigm to measure probabilistic reasoning and a possible answer to the question why psychosis-prone individuals jump to conclusions. Journal of Abnormal Psychology 126(4), 406-415.

Moritz, S., Scheunemann, J., Lüdtke, T., Westermann, S., Pfuhl, G., Balzan, R.P., Andreou, C., 2020. Prolonged rather than hasty decision-making in schizophrenia using the box task. Must we rethink the jumping to conclusions account of paranoia? Schizophrenia Research.

Moritz, S., Woodward, T.S., 2005. Jumping to conclusions in delusional and non-delusional schizophrenic patients. British Journal of Clinical Psychology 44(2), 193-207.

Moutoussis, M., Bentall, R.P., El-Deredy, W., Dayan, P., 2011. Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. Cognitive Neuropsychiatry 16(5), 422-447.

Nakagami, E., Xie, B., Hoe, M., Brekke, J.S., 2008. Intrinsic motivation, neurocognition and psychosocial functioning in schizophrenia: Testing mediator and moderator effects. Schizophrenia Research 105(1), 95-104.

Nasar, S., 1998. A Beautiful Mind: The Life of Mathematical Genius and Nobel Laureate John Nash. Simon and Schuster.

Nash, J., 1994. The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1994. NobelPrize.org.

Nastase, S.A., Goldstein, A., Hasson, U., 2020. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. NeuroImage 222, 117254.

Nour, M.M., Dahoun, T., Schwartenbeck, P., Adams, R.A., FitzGerald, T.H.B., Coello, C., Wall, M.B., Dolan, R.J., Howes, O.D., 2018. Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. Proceedings of the National Academy of Sciences 115(43), E10167-E10176.

Orbán, G., Wolpert, D.M., 2011. Representations of uncertainty in sensorimotor control. Current Opinion in Neurobiology 21(4), 629-635.

Ortmann, A., 2009. The way in which an experiment is conducted is unbelievably Important: on the experimentation practices of economists and psychologists. CESifo Working Paper.

Paliwal, S., Mosley, P.E., Breakspear, M., Coyne, T., Silburn, P., Aponte, E., Mathys, C., Stephan, K.E., 2019. Subjective estimates of uncertainty during gambling and impulsivity after subthalamic deep brain stimulation for Parkinson’s disease. Scientific Reports 9(1), 14795.

Palmer, B.W., Heaton, R.K., Paulsen, J.S., Kuck, J., Braff, D., Harris, M.J., Zisook, S., Jeste, D.V., 1997. Is it possible to be schizophrenic yet neuropsychologically normal? Neuropsychology 11(3), 437-446.

Palmer, C.J., Lawson, R.P., Hohwy, J., 2017. Bayesian approaches to autism: Towards volatility, action, and behavior. Psychological Bulletin 143(5), 521-542.

Patel, G.H., Arkin, S.C., Ruiz-Betancourt, D.R., DeBaun, H.M., Strauss, N.E., Bartel, L.P., Grinband, J., Martinez, A., Berman, R.A., Leopold, D.A., Javitt, D.C., 2020. What you see is what you get: visual scanning failures of naturalistic social scenes in schizophrenia. Psychological Medicine, 1-10.

Peters, E., Joseph, S., Day, S., Garety, P., 2004. Measuring Delusional Ideation: The 21-Item Peters et al. Delusions Inventory (PDI). Schizophrenia Bulletin 30(4), 1005-1022.

Powers, A.R., Mathys, C., Corlett, P.R., 2017. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science 357(6351), 596-600.

Redish, A.D., Jensen, S., Johnson, A., 2008. A unified framework for addiction: vulnerabilities in the decision process. Behav Brain Sci 31(4), 415-437; discussion 437-487.

Reed, E.J., Uddenberg, S., Suthaharan, P., Mathys, C.D., Taylor, J.R., Groman, S.M., Corlett, P.R., 2020. Paranoia as a deficit in non-social belief updating. eLife 9, e56345.

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(5), 1183-1191.

Schmack, K., Castro, A.G.-C.d., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., Heinz, A., Petrovic, P., Sterzer, P., 2013. Delusions and the Role of Beliefs in Perceptual Inference. Journal of Neuroscience 33(34), 13701-13712.

Schultz, W., 2016. Dopamine reward prediction error coding. Dialogues Clin Neurosci 18(1), 23-32.

Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275(5306), 1593-1599.

Shadlen, M.N., Shohamy, D., 2016. Decision Making and Sequential Sampling from Memory. Neuron 90(5), 927-939.

Sharot, T., Garrett, N., 2016. Forming Beliefs: Why Valence Matters. Trends in Cognitive Sciences 20(1), 25-33.

Siddiqui, I., Saperia, S., Fervaha, G., Da Silva, S., Jeffay, E., Zakzanis, K.K., Agid, O., Remington, G., Foussias, G., 2019. Goal-directed planning and action impairments in schizophrenia evaluated in a virtual environment. Schizophrenia Research 206, 400-406.

Silvetti, M., Seurinck, R., van Bochove, M., Verguts, T., 2013. The influence of the noradrenergic system on optimal control of neural plasticity. Frontiers in Behavioral Neuroscience 7.

Simon, H.A., 1990. Bounded Rationality, in: Eatwell, J., Milgate, M., Newman, P. (Eds.), Utility and Probability. Palgrave Macmillan UK, London, pp. 15-18.

Smith, P.L., Ratcliff, R., 2004. Psychology and neurobiology of simple decisions. Trends in Neurosciences 27(3), 161-168.

Smith, R., Badcock, P., Friston, K.J., 2020. Recent advances in the application of predictive coding and active inference models within clinical neuroscience. Psychiatry and Clinical Neurosciences n/a(n/a).

So, S.H.-w., Siu, N.Y.-f., 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.

So, S.H., Garety, P.A., Peters, E.R., Kapur, S., 2010. Do Antipsychotics Improve Reasoning Biases? A Review. Psychosomatic Medicine 72(7), 681.

Soltani, A., Khorsand, P., Guo, C., Farashahi, S., Liu, J., 2016. Neural substrates of cognitive biases during probabilistic inference. Nature Communications 7(1), 11393.

Soltani, A., Wang, X.-J., 2010. Synaptic computation underlying probabilistic inference. Nature Neuroscience 13(1), 112-119.

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. J Psychiatry Neurosci 35(1), 7-17.

Spitzer, M., 1990. On defining delusions. Comprehensive Psychiatry 31(5), 377-397.

Stephan, K.E., Mathys, C., 2014. Computational approaches to psychiatry. Current Opinion in Neurobiology 25, 85-92.

Stephan, K.E., Petzschner, F.H., Kasper, L., Bayer, J., Wellstein, K.V., Stefanics, G., Pruessmann, K.P., Heinzle, J., 2019. Laminar fMRI and computational theories of brain function. NeuroImage 197, 699-706.

Sterzer, P., Adams, R.A., Fletcher, P., Frith, C., Lawrie, S.M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., Corlett, P.R., 2018. The Predictive Coding Account of Psychosis. Biological Psychiatry 84(9), 634-643.

Sterzer, P., Voss, M., Schlagenhauf, F., Heinz, A., 2019. Decision-making in schizophrenia: A predictive-coding perspective. NeuroImage 190, 133-143.

Stompe, T., Ortwein-Swoboda, G., Ritter, K., Schanda, H., 2003. Old Wine in New Bottles? Psychopathology 36(1), 6-12.

Strauss, G.P., Waltz, J.A., Gold, J.M., 2014. A review of reward processing and motivational impairment in schizophrenia. Schizophrenia Bulletin 40(Suppl 2), S107-S116.

Stuke, H., Stuke, H., Weilnhammer, V.A., Schmack, K., 2017. Psychotic Experiences and Overhasty Inferences Are Related to Maladaptive Learning. PLoS Comput Biol 13(1), e1005328.

Stuke, H., Weilnhammer, V.A., Sterzer, P., Schmack, K., 2019. Delusion Proneness is Linked to a Reduced Usage of Prior Beliefs in Perceptual Decisions. Schizophrenia Bulletin 45(1), 80-86.

Takeda, K., Matsumoto, M., Ogata, Y., Maida, K., Murakami, H., Murayama, K., Shimoji, K., Hanakawa, T., Matsumoto, K., Nakagome, K., 2017. Impaired prefrontal activity to regulate the intrinsic motivation-action link in schizophrenia. NeuroImage: Clinical 16, 32-42.

Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P.R., Goodyer, I.M., Fletcher, P.C., 2015. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. Proceedings of the National Academy of Sciences 112(43), 13401-13406.

Tripoli, G., Quattrone, D., Ferraro, L., Gayer-Anderson, C., Rodriguez, V., Cascia, C.L., Barbera, D.L., Sartorio, C., Seminerio, F., Tarricone, I., Berardi, D., Szöke, A., Arango, C., Tortelli, A., Llorca, P.-M., Haan, L.d., Velthorst, E., Bobes, J., Bernardo, M., Sanjuán, J., Santos, J.L., Arrojo, M., Del-Ben, C.M., Menezes, P.R., Selten, J.-P., Group, E.-G.W., Jones, P.B., Jongsma, H.E., Kirkbride, J.B., Lasalvia, A., Tosato, S., Richards, A., O'Donovan, M., Rutten, B.P.F., Os, J.v., Morgan, C., Sham, P.C., Murray, R.M., Murray, G.K., Forti, M.D., 2020. Jumping to conclusions, general intelligence, and psychosis liability: findings from the multi-centre EU-GEI case-control study. Psychological Medicine, 1-11.

Upthegrove, R., 2018. Delusional Beliefs in the Clinical Context, in: Bortolotti, L. (Ed.), Delusions in Context. Springer International Publishing, Cham, pp. 1-34.

Usher, M., McClelland, J.L., 2001. The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review 108(3), 550-592.

van der Leer, L., McKay, R., 2014. "Jumping to conclusions" in delusion-prone participants: an experimental economics approach. Cognitive Neuropsychiatry 19(3), 257-267.

Vilares, I., Howard, J.D., Fernandes, H.L., Gottfried, J.A., Kording, K.P., 2012. Differential Representations of Prior and Likelihood Uncertainty in the Human Brain. Current Biology 22(18), 1641-1648.

Vincent, P., Parr, T., Benrimoh, D., Friston, K.J., 2019. With an eye on uncertainty: Modelling pupillary responses to environmental volatility. PLOS Computational Biology 15(7), e1007126.

Walters, C.J., Redish, A.D., 2018. Chapter 8 - A Case Study in Computational Psychiatry: Addiction as Failure Modes of the Decision-Making System, in: Anticevic, A., Murray, J.D. (Eds.), Computational Psychiatry. Academic Press, pp. 199-217.

Weinstein, J.J., Chohan, M.O., Slifstein, M., Kegeles, L.S., Moore, H., Abi-Dargham, A., 2017. Pathway-Specific Dopamine Abnormalities in Schizophrenia. Biological Psychiatry 81(1), 31-42.

Wellstein, K.V., Diaconescu, A.O., Bischof, M., Rüesch, A., Paolini, G., Aponte, E.A., Ullrich, J., Stephan, K.E., 2020. Inflexible social inference in individuals with subclinical persecutory delusional tendencies. Schizophrenia Research 215, 344-351.

Wengler, K., Goldberg, A.T., Chahine, G., Horga, G., 2020a. Distinct hierarchical alterations of intrinsic neural timescales account for different manifestations of psychosis. eLife 9, 2020.2002.2007.939520.

Wengler, K., He, X., Abi-Dargham, A., Horga, G., 2020b. Reproducibility assessment of neuromelanin-sensitive magnetic resonance imaging protocols for region-of-interest and voxelwise analyses. NeuroImage 208, 116457.

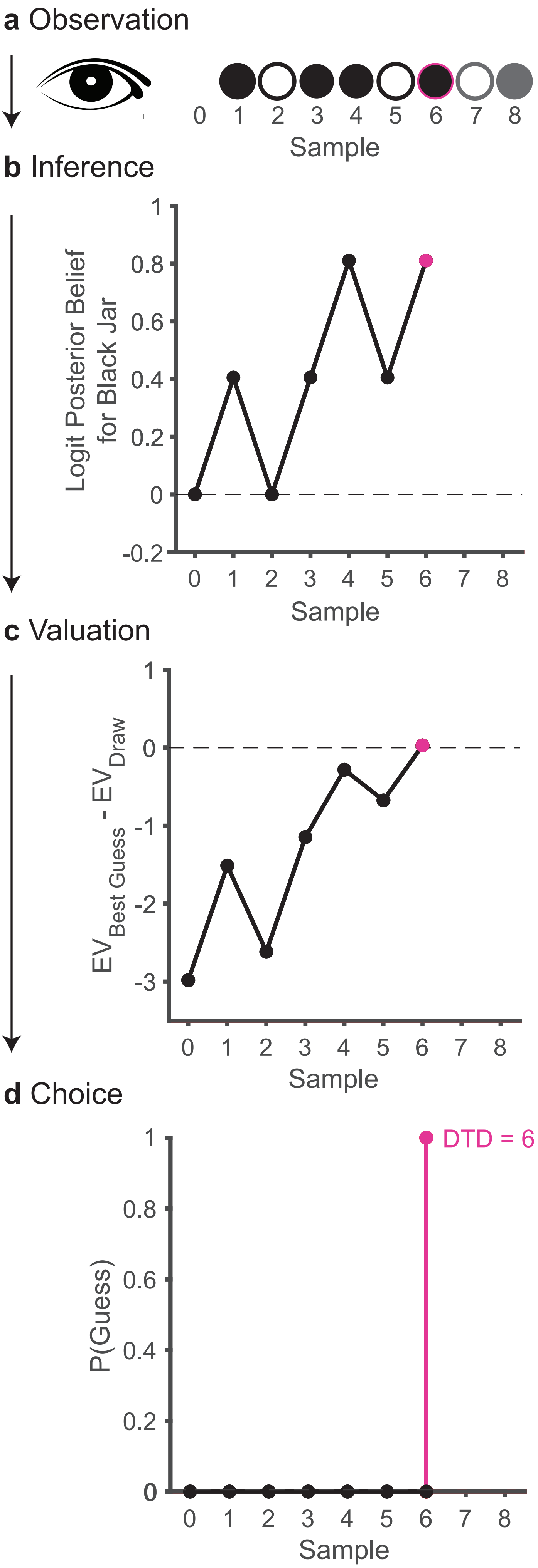
White, L.O., Mansell, W., 2009. Failing to ponder? Delusion-prone individuals rush to conclusions. Clin Psychol Psychother 16(2), 111-124.

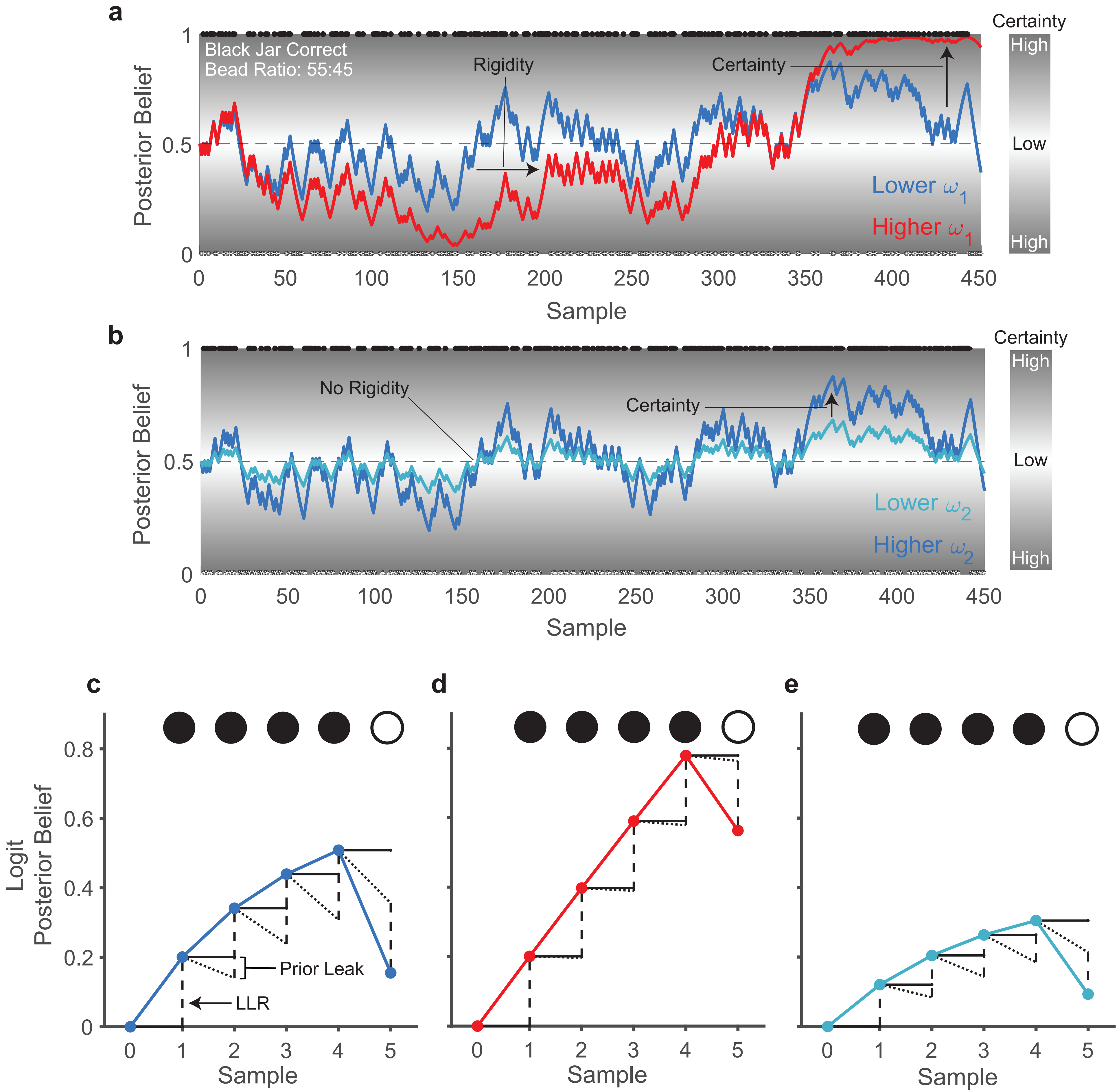
Wilson, R.C., Collins, A.G.E., 2019. Ten simple rules for the computational modeling of behavioral data. eLife 8, e49547.

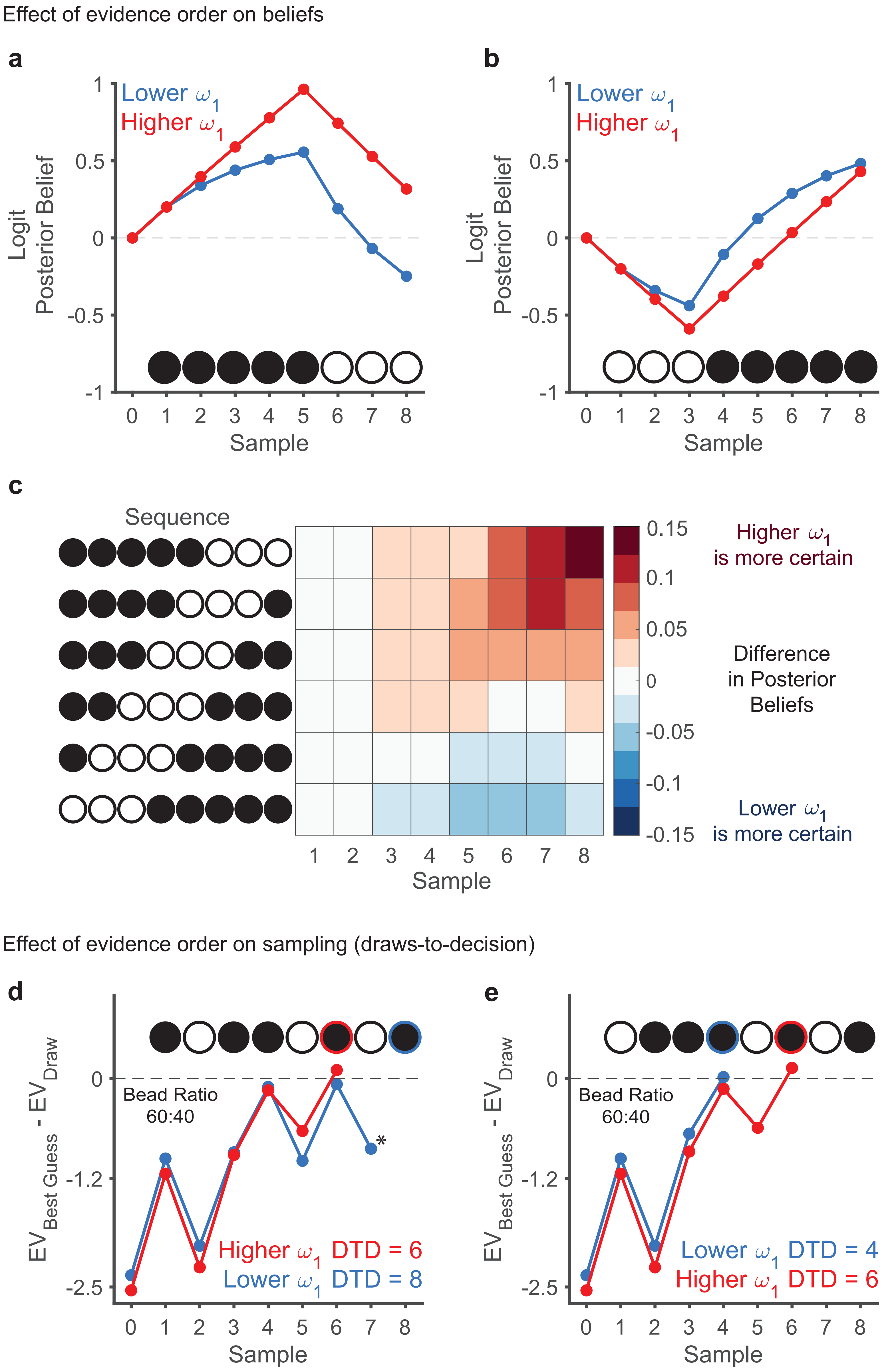
Woodward, T.S., Jung, K., Hwang, H., Yin, J., Taylor, L., Menon, M., Peters, E., Kuipers, E., Waters, F., Lecomte, T., Sommer, I.E., Daalman, K., van Lutterveld, R., Hubl, D., Kindler, J., Homan, P., Badcock, J.C., Chhabra, S., Cella, M., Keedy, S., Allen, P., Mechelli, A., Preti, A., Siddi, S., Erickson, D., 2014. Symptom Dimensions of the Psychotic Symptom Rating Scales in Psychosis: A Multisite Study. Schizophrenia Bulletin 40(Suppl_4), S265-S274.

Yang, T., Shadlen, M.N., 2007. Probabilistic reasoning by neurons. Nature 447(7148), 1075-1080.

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(2), 455-465.


a Observation
0
1
2
3
4
5
6
7
8
Sample
b Inference
Logit Posterior Belief for Black Jar
1
0.8
0.6
0.4
0.2
0
-0.2
0
1
2
3
4
5
6
7
8
Sample
c Valuation
$EV_{Best\ Guess} - EV_{Draw}$
1
0
-1
-2
-3
0
1
2
3
4
5
6
7
8
Sample
d Choice
P(Guess)
DTD = 6
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8
Sample


a
Certainty
High
Low
High
Black Jar Correct
Bead Ratio: 55:45
Rigidity
Certainty
Lower ω_1
Higher ω_1
Posterior Belief
1
0.5
0
0
50
100
150
200
250
300
350
400
450
Sample
b
Certainty
High
Low
High
No Rigidity
Certainty
Lower ω_2
Higher ω_2
Posterior Belief
1
0.5
0
0
50
100
150
200
250
300
350
400
450
Sample
c
d
e
Logit
Posterior Belief
0.8
0.6
0.4
0.2
0
Prior Leak
LLR
0
1
2
3
4
5
Sample
0
1
2
3
4
5
Sample
0
1
2
3
4
5
Sample


Effect of evidence order on beliefs
a
b
Lower ω_1
Higher ω_1
Logit Posterior Belief
1
0.5
0
-0.5
-1
0 1 2 3 4 5 6 7 8
Sample
c
Sequence
1 2 3 4 5 6 7 8
Sample
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
Higher ω_1 is more certain
Difference in Posterior Beliefs
Lower ω_1 is more certain
Effect of evidence order on sampling (draws-to-decision)
d
e
$EV_{Best Guess} - EV_{Draw}$
0
-1.2
-2.5
Bead Ratio 60:40
Higher ω_1 DTD = 6
Lower ω_1 DTD = 8
*
Lower ω_1 DTD = 4
Higher ω_1 DTD = 6
0 1 2 3 4 5 6 7 8
Sample