Project Description

Matthew Nassar[1], Noham Wolpe[2], Oriel FeldmanHall[1], Frederike Petzschner[1]
[1]Carney Institute for Brain Science, Brown University, Providence RI, USA
[2]Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

Specific Aims

People form beliefs about themselves and the world that contribute to nearly every aspect of behavior. Such beliefs can facilitate adaptive behaviors when updated appropriately, but failures to update them can lead to inaccurate beliefs that have devastating consequences, such as delusions in psychosis, cognitive distortions in depression, and conspiracy theories in otherwise healthy individuals (e.g., QAnon). Despite active research across research disciplines, there is a limited understanding of the cognitive and neural mechanisms of belief updating. This is largely because of three major limitations: 1) a focus on individual tasks and lack of understanding of their link to real-life behavior, 2) over-simplified models that miss the contextual nature of real-world inference problems, and 3) a failure to link computational descriptions of behavior to their neural circuitry. Overcoming these limitations is critical for understanding the mechanisms of belief updating and developing interventions for mental health conditions with abnormal belief updating.

Here, we draw on our experience in computational neuroscience and clinical psychiatry to overcome these challenges. First, instead of focusing on one neuroscience domain, we develop a belief updating battery and computational framework capable of identifying domain-general belief updating phenotypes that apply broadly across a range of tasks, and examine how they relate to real-world belief updating and mental health traits. Second, we build on a Bayesian framework that embraces the complexity of real-world belief updating, by considering contextual inference (Heald et al., 2021). In this model, belief updating critically depends not only on prior and new information for incremental learning, but also on beliefs about the context the agent is currently in. Third, we use a novel neural network instantiation of this model (Razmi & Nassar, 2022) to make testable predictions regarding how biological differences could lead to individual differences in belief updating. This will provide a new biological insight into normal and abnormal belief updating.

Aim 1: Identify task-general parameters that predict real-world belief updating in healthy adults. We will use a battery of tasks that probe belief updating across neuroscience domains, including sensorimotor, social, and perceptual neuroscience tasks. We will administer the battery to a large online cohort (n=1000) and collect self-report measures to probe belief updating in everyday life. We will extend a contextual Bayesian inference model of belief updating to test if specific biases in contextual inference explain deviations in belief updating behavior across tasks and self-report measures. We will test the ability of this model to predict participant behavior 'out-of-task' better than a combination of previously explored models.

Aim 2: Identify the biological mechanisms of belief updating in healthy adults. In the neural network instantiation of our contextual inference model, incremental learning is mediated by cortico-striatal synaptic plasticity, but more drastic changes in behavior are afforded by switches in prefrontal attractor states that represent the relevant behavioral context. Critically, such transitions depend on thalamic projections to cortex whose impact in turn depends on the E/I balance in cortex. To test key predictions of this model, we will administer a subset of the tasks from Aim 1 while collecting fMRI, MRS, EEG, and eye-tracking data to track biological markers of incremental learning and contextual changes, namely thalamocortical connectivity, glutamate/GABA differences, EEG-based P300 and pupil dilation changes.

Aim 3: Identify the biological mechanisms of abnormal belief updating in schizophrenia. We will examine whether schizophrenia patients update context representations abnormally. We will test whether patients show a greater tendency for the 'all-or-nothing' belief updating, particularly in high variability conditions. We will further test whether this behavioral change is mediated by neurophysiological markers for context transition (EEG-based P300 and pupil dilation) and whether these measures scale with the severity of delusion symptoms across patients. By simulating a model with reduced parvalbumin positive interneurons, as seen in schizophrenia, we will test whether our biologically informed model can reproduce behavioral and neurophysiological changes in patients.

Intellectual Merit:
Our project has the potential to explain a key aspect of human behavior, that can go awry in many mental health conditions. Our proposed project is unique in that it sets a new standard for 'construct validity' – that is, testing whether and how things measured in the lab apply to real-life behavior.

Broader Impacts:
Our project can define a new biologically informed model for understanding delusional beliefs, which can be used to inform new treatments. We will also make our rich dataset widely available, so that other groups will be able to test their models or further develop our proposed model and test new emerging hypotheses.

1 Overview & Significance

People's behavior is largely determined by their beliefs about themselves and the world around them. For example, a student may stay up the night before an exam to revise if they believe they are unprepared for the exam. A 'belief' in this sense refers to an inferential process, whereby one combines available information to establish what they consider to be a true or accurate representation of reality. To reflect reality as accurately as possible, a belief needs to be constantly updated based on new information. For example, the student might change their mind and consider they are prepared for the exam if they successfully complete a mock exam, and can then choose to stop studying and go to sleep. Belief updating is an integral part of everyday adaptive behavior and is often done seamlessly.

This seamless belief updating is impressive because the 'right' way to update beliefs is not always clear. Updating beliefs in the face of new information is not always beneficial. For example, if the mock exam is old and does not reliably reflect the current level of exams in the course, the student may choose to continue with their 'all-nighter', despite apparent evidence they are ready for the exam. While beneficial in some circumstances, ignoring new evidence and adhering to prior beliefs instead can be problematic in other circumstances, as in the maintenance of conspiratorial beliefs or fixed beliefs seen in delusions.

Normative accounts have suggested that the brain combines information in a precision-weighted manner. That is, precise (reliable) data have a stronger influence on one's belief, whereas uncertain or noisy data are weighted less when updating a belief (Ernst & Banks, 2002). Such accounts rely on Bayesian inference, whereby stored beliefs are represented as 'priors' over possible states of the world, which are updated according to the 'likelihood' with which these states would give rise to observations (Fletcher & Frith, 2009). The past two decades have seen a substantial increase in the number of studies using Bayesian inference to explain a wide range of brain functions, such as learning and motor control (Behrens et al., 2007; Körding & Wolpert, 2004; Ma et al., 2006). PIs Wolpe and Nassar have used this approach to examine how people update beliefs about their performance (Hezemans et al., 2020, 2022; Wolpe et al., 2014, 2015) and latent features of the environment (Nassar et al., 2010, 2012; Nassar, Bruckner, et al., 2019), identifying both normative and counter-normative aspects of belief updating in healthy and clinical populations (Nassar et al., 2021; Nassar & Troiani, 2021).

The significant interest in these normative and counter-normative aspects of belief updating stems from their possible implications to real-world behavior, and mental health disorders. Delusion is perhaps an extreme example of reduced belief updating in mental health disorders, including schizophrenia. PI Wolpe is a clinical psychiatrist who has worked with patients with delusional disorders and witnessed the unmet clinical need to improve our understanding and treatment in this clinical population. The Diagnostic and Statistical Manual of Mental Disorders 5 (DSM 5) defines delusion as "fixed beliefs that are not amenable to change in light of conflicting evidence" (American Psychiatric Association, 2013). That is, a delusion, in its broadest definition, is a belief that cannot be or is not updated. While we note that there have been debates as to whether one or two processes are required to explain the formation and maintenance of a delusional belief (Coltheart, 2010; Corlett, 2019; Miyazono & McKay, 2019), it appears that an inability to update beliefs in light of contrasting evidence is central to most accounts. In the basic Bayesian framework, delusional belief could be represented by an overly precise prior, which is indeed consistent with some laboratory tasks (Baker et al., 2019; Powers et al., 2017). However, in other paradigms, those prone to delusions under-utilize prior beliefs (Evans et al., 2015; Stuke et al., 2019). Recent work from PI Nassar has shown both of these seemingly antithetical behavioral features in a single predictive inference task, where on any given trial, schizophrenia patients are both 1) more likely to completely update their beliefs

in accordance with new information, and 2) more likely to ignore new information altogether, leading to 'all-or-nothing' updating behavior (Nassar et al., 2021).

While Bayesian inference has intuitive and normative appeal, its predictions depend critically on assumptions about the structure of the world. These assumptions, we argue, are unrealistically simplistic in how they have been implemented in some of the previous work, as they ignore the context in which beliefs are updated. Our proposal aims to incorporate this additional level of complexity to explain the mechanisms underlying the counterintuitive behavioral updating seen in schizophrenia patients.

The work of PI Nassar over the past decade has examined the mechanisms of belief updating in complex changing environments (McGuire et al., 2014; Nassar, Bruckner, et al., 2019; Nassar et al., 2012). Importantly, the lab has recently developed a computational model that can explain belief updating in predictive inference tasks along with its EEG, fMRI, and pupillometric correlates (Razmi & Nassar, 2022; Yu et al., 2021). A key component of the model is latent state (or context) representations, which encode the unobservable context through which new observations are interpreted. Flexible belief updating, which involves rapid belief updates in some situations but robustness to discordant information in others, emerge through changes to the active latent state (Razmi & Nassar, 2022). Normal belief updating thus requires not only the integration of priors with new evidence, but also the correct identification of latent state for which this belief is updated (Heald et al., 2021). This poses a central credit assignment problem: are my expectations wrong? Or am I just in a different context than I thought I was in? We argue that this problem has important implications for belief updating in healthy and clinical populations.

In terms of their underlying brain correlates, such latent state representations have been observed in the prefrontal cortex (Nassar, McGuire, et al., 2019; Schuck et al., 2016; Vaidya et al., 2021). In neural populations, switches from one active state to another are envisioned as resets to a cortical attractor state – that is, the set of recurrently connected neurons in a neural network that are active. These resets are thought to be facilitated by thalamocortical signaling (Rikhye et al., 2018; Schmitt et al., 2017).

This model could provide a useful framework for thinking about how biological changes in schizophrenia might impair belief updating. A primary biological feature of schizophrenia is prefrontal impairments, due to, at least in part, reductions in parvalbumin positive inhibitory interneurons (Lodge et al., 2009). Such interneurons are thought to mediate the stabilizing effects of conflict-sensitive thalamocortical projections that are active when new sensory information is only weakly related to the underlying context giving rise to them (Mukerjee 2022). Loss of such neurons could lead to overly permissive attractor switching in the cortex. Neurally, this would mean that even a small discrepancy between actual and observed outcomes (or 'prediction errors') could force prefrontal networks to shift into a new attractor state. Jumping between attractor states is expected to prevent the circuit from integrating new information. This is because associations learned from each new outcome would be assigned to a different state and stored in projection weights of a different subset of prefrontal neurons. The exact behavioral consequence would depend on which attractor state is active when beliefs are next probed: if the new attractor state is active, then beliefs would be completely updated in accordance with the newest observation ('all' belief updating), whereas if the previous attractor state is active, then behavior would look like a lack of learning ('nothing' belief updating). Considering both possibilities provides a potential computational explanation for the 'all-or-nothing' learning behavior that we recently observed in schizophrenia patients, and the spuriously created attractor state bears some resemblance to delusions in schizophrenia. Following this logic, such a belief updating pathology might also prevent updating of fixed delusions, as information discrepant with the delusion would promote a shift in the cortical attractor state, thereby assigning the delusion disconfirming information to an alternate cause/attractor. Our proposed project will investigate these mechanisms in healthy and clinical populations, after first establishing their construct validity.

2 Innovation

A major gap in the belief updating literature is in the through line connecting laboratory tasks to broader constructs and real-world behavior. A key innovation of our work is to bridge this gap, by combining a large-scale multi-domain behavioral study with a computational framework for modeling behavior across tasks, thereby setting a new standard for construct validity. A second innovation of our proposal is that it embraces the complexity and diversity of real-world belief updating through a flexible cognitive modeling framework

in which beliefs are updated through contextual inference. We have recently developed this framework (Yu et al., 2021); shown how it can be implemented (Razmi & Nassar, 2022); and demonstrated its broad application to social interactions (FeldmanHall & Nassar, 2021). Third, our work innovates by mapping our cognitive framework onto canonical cortico-thalamic circuitry for a biologically constrained model of flexible belief updating. This allows us to draw on recent advances in our understanding of thalamocortical signaling (Mukherjee et al., 2021) in order to make predictions about how specific biological perturbations, including those thought to drive symptoms in schizophrenia, would affect belief updating across tasks and domains.

3 Approach and preliminary data

Our overall approach is to first identify domain-general features of belief updating that map onto real-life behavior, using novel computational models of context-dependent belief updating. We will then test the ability of these features to predict behavioral and neuropsychological data in a healthy population and in patients with schizophrenia. We will examine the cognitive and neural bases for normal and abnormal belief updating through three distinct aims that each tightly integrates experiments with computational modeling.

Aim 1: Identify task invariant parameters that predict real-world belief updating in healthy adults. In the first Aim, we examine which aspects of belief updating are domain-general and predict real-world behavior. We will extend a model of contextual Bayesian belief updating to capture behavior across multiple distinct tasks from different domains of research in neuroscience and use it to infer self-reported behavioral measures in a large-scale behavioral study.

Over the past two decades, belief updating has become a central topic of research in psychology, psychiatry, and neuroscience (Bromberg-Martin & Sharot, 2020; FeldmanHall & Nassar, 2021; Kube & Rozenkrantz, 2021). Numerous tasks have emerged to measure individual differences in belief updating (Behrens et al., 2007; Kobayashi et al., 2021; Körding & Wolpert, 2004; Lamba et al., 2020; Nassar et al., 2010; Sharot, 2011; Wolpe et al., 2013), and in many cases individual differences in these tasks relate to differences in underlying brain function (Krugel et al., 2009; McGuire et al., 2014; Nassar et al., 2012) or psychiatric conditions (Browning et al., 2015; Garrett & Sharot, 2014; Lamba et al., 2020; Nassar et al., 2021; Nassar & Troiani, 2021). The implicit assumption in this line of research is as follows: 1) lab-based experimental tasks serve as models for the belief updating behavior realized in everyday life; and 2) by understanding the neural underpinnings of task behavior, and by developing interventions to alter them, we might nudge people away from pathological belief updating behaviors (e.g., conspiratorial, or delusional beliefs) through targeted interventions. However, the success of such an approach requires, at minimum, an understanding of how belief updating in one task or domain will relate to that in another, and whether domain-general behavior in laboratory tasks relates to real-world belief updating.

In Aim 1, we will collect the empirical behavioral data and develop the computational modeling framework necessary for such an understanding of how belief updating constructs manifest across tasks and in the real-world. We will collect behavioral data from a large online multinational cohort of participants on a battery of belief updating tasks (Table 1) that includes two tasks created explicitly to measure belief updating, namely predictive inference (Nassar et al., 2021) and belief updating tasks (Sharot, 2011); as well as tasks that measure the degree of belief updating in specific neuroscience domains [Social: iterative trust game (Lamba et al., 2020), Motor: Adaptation (Tsay et al., 2021; Wolpe et al., 2020), Perceptual: magnitude estimation (Petzschner et al., 2015; Petzschner & Glasauer, 2011), Self-assessment: performance belief task (Wolpe et al., 2014)]. Completion of the entire battery of tasks will require 4 hours spread across three sessions with order randomized across a large sample of 1000 participants. Given the range of task domains and details, *ad hoc* strategy developed for a specific task will be unlikely to apply generally across the whole task set. Nonetheless, each task will have a similar structure, as it will require the updating of beliefs in accordance with new information, and knowledge about the context, which can change during the task (Table 1). On each task, the participant will provide information about his or her belief, either through a choice (e.g., endowing money in the iterative trust game) or through direct report (estimating an updated risk in the belief updating task). The participant then receives additional information, often in conflict with current belief, and then provides information about his or her updated belief through either choice or direct report. Given this shared structure across tasks, it is possible that belief updating pattern in all tasks could be described by a single computational process, however, whether this is the case, and if so, what that process is, remain open empirical questions.

Table 1. Brief description of belief updating tasks in our task battery.

Task name	Brief description	Principal measures
Predictive inference task (Nassar et al., 2021)	On each trial, participants are asked to move a 'bucket' to a position on the screen to where they predict coins will be dropped (by a 'helicopter'), in order to maximize the coins they collect (reward). They then get feedback as to the actual position where the coins are dropped and reward earned, requiring them to update the predicted location on the next trial.	**Belief updating** is measured as the extent to which participants adjust the bucket location on each trial. This is calculated separately for an 'oddball' condition, in which unexpected coin position are a one-off outlier, and in a 'changepoint' condition, in which the underlying distribution generating the coin position is changed.
Belief update task (Sharot, 2011)	The task includes two sequential blocks: pre- and post-feedback blocks. On each trial, participants see a short description of an adverse life event (e.g., 'card fraud') and are asked to estimate how likely this event can occur to them in the future. In the pre-feedback block, after estimation they see the actual probability of that event occurring to someone with a similar demographic background to them.	**Belief update** is calculated as the difference between probability estimates in pre- and post-feedback blocks. This is calculated separately for 'good news' and 'bad news' trials, in which the initial estimate is higher or lower than the feedback, respectively.
Iterative trust game (Lamba et al., 2020)	On each trial, participants are asked how much 'money' they wish to invest with a given partner. Investment is multiplied by 4 and then the partner returns some fraction of the money. One set of partners changes their mean return proportion occasionally, whereas another set is variable but overall stable in their returns.	**Belief updating** is measured as the extent to which participants adjust their investments in the face of a positive or negative trial outcome. The specificity of belief updating is assessed by how trial outcomes with one partner affect future investments with another.
Motor adaptation (Tsay et al., 2021; Wolpe et al., 2020)	Participants use their mouse/trackpad to move a cursor on the screen so as to hit a target, pseudorandomly displayed 45° or 135° with respect to the cursor starting position on each trial. The visual feedback for the cursor position is displayed with a rotation angle with respect to the real movement direction (0°, 30° or 60° – experimental 'context').	Trial-by-trial **belief updating** is calculated as the difference between hand movement in trial *n* and that on trial *n-1*. angle in one trial and the previous trial. This is calculated for each 'context' in the experiment.
Perceptual magnitude task (Petzschner et al., 2015)	In every trial, participants are asked to estimate the duration of a time interval (production-reproduction or 2AFC). The durations of the sample intervals are drawn from different discrete partially overlapping uniform distributions that change over time.	**Belief update** is measured on a trial-by-trial basis by the degree to which time estimates are biased by the underlying context (sample distribution) in which they are presented.
Performance belief task (Wolpe et al., 2014)	In this variant, participants are asked to stop a ball when it is aligned with a target. Pseudorandomly, partially intermittent feedback is given as to the position of the moving ball. On each trial, ball speed is drawn from a Gaussian distribution. At some point in the experiment, the ball speed distribution changes. The task is to press a button when the ball is vertically aligned with the target. The ball and target then disappear, and participants move a cursor to estimate where the ball stopped.	**Performance belief** is calculated by the difference between estimated and true stopping position of the ball with respect to the target. **Belief updating** is calculated from a 'feedback' block where participants see the real and estimated position of the ball at the end of each trial. Both measures are calculated separately for different contexts (ball speed distributions).

To aid in answering this question, we will develop an integrated modeling framework through which behavioral data from one task could be used to predict how an individual might update their beliefs in another. The model set will rely on a Bayesian Hierarchical framework that builds on existing single-task models that provide descriptions of behavior within each task in our battery (Carpenter et al., 2017). The Bayesian hierarchical model will infer covariance parameter estimates from different tasks, thereby allowing it to predict the parameters that will best describe participant belief updating in a given task, even without observing participant data on that task. This will allow us to test the degree to which computational descriptions of belief updating generalize across tasks. As with computational models of individual tasks, fits from the hierarchical model set will be scrutinized through posterior predictive checking to identify specific generalization failures, which in turn inform adjustments to the individual task models so as to better capture cross-task differences in behavior (Nassar & Frank, 2016). We note that since the goal of this modeling is to maximize the predictive accuracy of leave-one-task-out behavioral predictions, the winning model may contain 'single-task models' that are quite different from those that provide the best fit to the individual tasks. In cases where multiple models could capture the basic pattern of single-task results, our approach would prefer the model containing parameters that best generalize across tasks. *Ad hoc* (task-specific) models have indeed been suggested to provide an overfit to the data (Navarro, 2018), thus limiting generalizability and application to real-life behavior.

Of particular interest in this regard is a class of cognitive models that approximate contextual Bayesian inference by assuming that new information is conditional on the latent states of the world (i.e., context) that gave rise to it (Heald et al., 2021; Yu et al., 2021). Such models rely on more realistic assumptions than typical Bayesian models of cognition in that they account for the possibility that an observation fails to match expectations due to an unobserved change of latent state. For example, having food poisoning after eating in my favorite restaurant may indicate that there was an untrained chef preparing my meal. In this example, how such an inference affects my beliefs about the future depends on my structural assumptions about the world: If I think that my favorite chef quit, then I might expect future meals to be of a low quality ('changepoint') whereas if I think that the untrained chef was just substituting in for a day ('oddball') then I might expect a high quality on subsequent meals and attribute my food poisoning to bad luck. The incorporation of structural assumptions allows contextual inference models to capture a broader range of behaviors than standard Bayesian models of cognition. That is because the same model can effectively update beliefs across different generative environments, so long as the model is given (or learns) the relevant transition structure (Fig. 1). In tasks where participants must make sequential predictions about a noisy variable, latent state models can describe key features of behavioral data (Razmi & Nassar, 2022), and unlike previous models (Nassar et al., 2010), they can do so for a wide variety of task structures (i.e., changepoints, oddballs, reversals). In addition, latent variables in contextual Bayesian inference models better match the primary neural correlates of performance in changepoint and oddball tasks (Razmi & Nassar, 2022), including fMRI representations that mimic representation of the latent state (Nassar, Bruckner, et al., 2019) and EEG & pupil signals that closely match context update

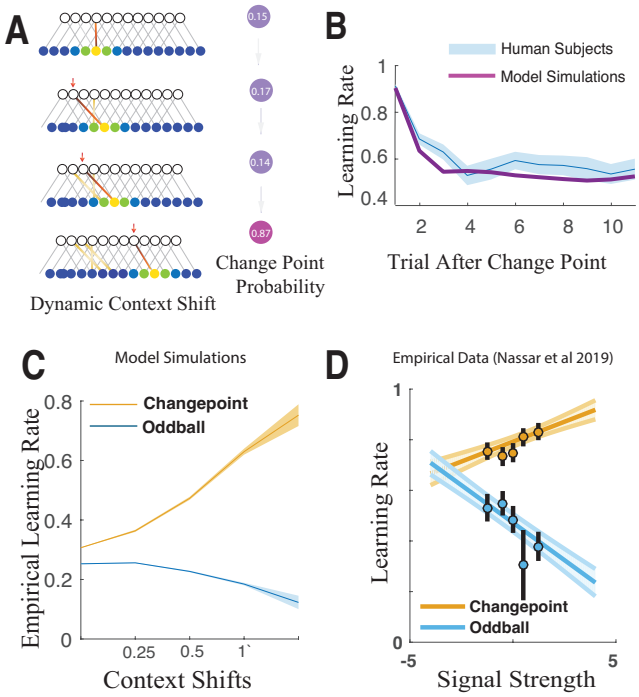


Figure 1. Contextual belief updating explains human behavior and EEG signals. A) Neural network from Razmi & Nassar 2021 utilizes dynamic input layer transitions to afford adaptive learning that matches human behavioral data **(B)**. **C)** Input layer transitions correspond to context shifts (abscissa) that differentially relate to belief updating (ordinate) across statistical contexts (color) in a similar manner to feedback locked EEG signals (P300) measured in humans **(D)**.

signals in the model (Nassar et al., 2012; Nassar, Bruckner, et al., 2019; O'Reilly et al., 2013; Razmi & Nassar, 2022). One possibility is that contextual Bayesian inference using latent states provides a general framework for belief updating that can be shared across domains and task structures.

Here, we will test this idea by constructing a hierarchical Bayesian model that includes models of each task in our battery that rely on learned associations to inferred latent states. Each individual task-model will have at least two components. One component will be a system for incremental learning, through which the beliefs for a given latent state or context can be adjusted slowly in proportion to the errors made in predicting new observations (prediction errors). The second component is a system for controlling which context a given observation will be attributed to, thereby allowing the model to rapidly update the active context in the case of a changepoint or a one-off outlier event in the case of an oddball. We will create variants of this model that map onto belief updating in each task in our battery, and combine these models into a single hierarchical model that pools data across participants and tasks by estimating cross-task parameter covariance. As described above, fitting will be evaluated according to out-of-task prediction – that is, how well can observing belief updating of a participant in all but one tasks predict their unique belief updating in the held-out task. The out-of-task predictive accuracy for this contextual Bayesian inference model will be compared to that of the original hierarchical model to test whether latent states provide a better general description of belief updating differences across tasks.

To understand how belief updating in laboratory tasks relates to behavior in real-world settings and to broader mental health constructs, we will also collect self-report data that measures 1) self-reported real-world belief updating style. To this end, we have developed and piloted a new questionnaire that probes real-life belief updating behavior (Fig. 2); 2) delusional ideation (Peters Delusion Inventory, Peters et al., 1999); 3) Intolerance of uncertainty (Buhr & Dugas, 2002); 4) Dimensional traits linked to autism (Autism Questionnaire 50, Baron-Cohen et al., 2001); 5) Depression (Patient Healthy Questionnaire-9, PHQ9, Kroenke et al., 2001); 6) Anxiety (General Anxiety Disorder-7, GAD7 Spitzer et al., 2006); and 7) apathy (Apathy Evaluation Scale, AES Marin et al., 1991). We will combine data from our multi-domain belief updating battery, our questionnaire measures, and our Bayesian Hierarchical model set to test the following specific hypotheses:

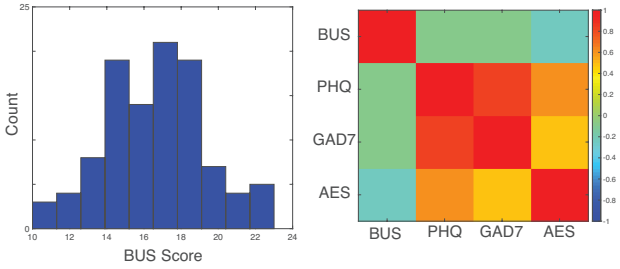


Figure 2. Belief updating survey. A) Score distribution from preliminary data (N=100) using belief updating survey characterizing self-reported real-world belief updating strategies. Higher scores indicate greater tendency to revise beliefs in response to new information. **B)** Correlations (coefficients in color) show that our belief updating scale (BUS) is independent of mood (PHQ9) and anxiety (GAD7) but does negatively correlate with apathy (AES).

H1.1: The latent states framework can capture belief updating across all tasks. Latent state models will be formulated for each of the tasks in Table 1 using our previously published framework (Razmi & Nassar, 2022). The output layer of the neural network will be adjusted for each task such that output units fully represent the action space and the state transition function will be modified to match the transition structure for each task. Each modified neural network will include: 1) a parameter that governs the sensitivity to detecting a state transition and 2) a parameter that controls a base rate of state transitioning (irrespective of the new observation). Gaussian and softmax likelihood functions will be used to map output layer onto actions for continuous and discrete choice tasks, respectively. Model fits from the newly developed latent states models will be compared to those from the best-fitting previously published models using Watanabe-Akaike information criterion (WAIC) (Vehtari et al., 2017), along with posterior predictive checks to identify specific model shortcomings. We anticipate that latent states models will capture core features of behavior across tasks, but not quite achieve the within task fit metrics of the originally published models, which may be over-fits to the specific details of a single task.

H1.2: Measures of belief updating in one task will predict how an individual will update beliefs in another. The fits of hierarchical models (original and latent states) allowing for cross-task parameter covariance will be compared to those of models in which cross-task parameter covariances are fit to zero in terms of 'leave-one-task-out' cross validated likelihood (Vehtari et al., 2017). Difference in fits between

the models will be compared not just for the entire dataset to select a winning model, but also for each individual task, to examine how well individual differences in a task can be explained based on belief updating behavior in all other tasks. We anticipate that models with free parameter-covariance terms will provide a better overall out-of-task fit by sharing information across a subset of tasks that draw on domain-general belief updating constructs.

H1.3: Hierarchical Bayesian model based on latent states better predicts out-of-task belief updating behavior due to conserved state transition constructs. The fits of the two hierarchical models (Original: using the models originally developed for each individual tasks, and Latent States: using individual task model that all rely on the latent state framework) with cross-task covariance parameters will be compared to one another using cross-validated 'leave-one-task-out' likelihood. Posterior-predictive checks will be used to identify cross-task characteristics of behavior that may or may not be captured by each model. We anticipate that the latent state model will better predict behavior in a left-out task, and that such predictions will hinge on high fit covariance values for the latent state transition sensitivity and base rate parameters (described in H1.1), but that these parameter fits manifest in slightly different patterns of behavior in each task. Domain-general constructs will be identified by computing the eigenvectors of the cross-task parameter covariance matrix. Constructs identified in this way will be tested in a follow-up fitting procedure, in which the covariance matrix is re-parameterized according to the eigenvectors obtained from the initial fit proceeding from the largest to smallest eigenvalues, and with each additional eigenvector determining whether it improved fit over the prior model. A key question is whether latent states provide better cross-task fits by capturing consistent cross-task individual differences in the base rate and sensitivity latent state parameters.

H1.4: Laboratory measures of belief updating predict self-reported measures of real-world belief updating behavior. Cross-task constructs will be related to self-report measures of real-world belief updating and mental health traits in two ways. First, subject-level estimates of the constructs identified in H1.3 will be extracted and correlated directly with aggregate measures from questionnaires through canonical correlations analysis. Second, the Bayesian hierarchical model will be extended to include aggregate questionnaire scores with additional covariance parameters added allowing behavioral task constructs to covary with self-report measures. We predict that a construct capturing differences in the sensitivity parameter across tasks will positively covary with 1) self-reported flexibility of belief updating, 2) attention to detail (from AQ), and 3) intolerance of uncertainty. We predict that a construct capturing the base rate of transitions across tasks will relate to delusional ideation (PDI), consistent with our motivating hypothesis that delusions could emerge through a failure to maintain latent states in prefrontal attractor networks leading to unprovoked latent state transitions.

Preliminary data supporting our ability to execute Aim: Our ability to successfully carry out this Aim is supported by our previous modeling work showing how one model can capture different adaptive learning behaviors in different temporal contexts (Razmi & Nassar, 2022), Moreover, previous our work showed that belief updating in individual tasks examined here correlate with mental health constructs related to autism (Nassar & Troiani, 2021) and anxiety (Lamba et al., 2020). Lastly, our previous work demonstrated our ability to collect and apply Bayesian hierarchical modeling to a large multi-session online study datasets (Jang et al., 2019).

Potential pitfalls/alternative outcomes: It is possible that we find limited evidence of belief updating constructs that generalize across task. While this would seem to be a negative outcome at first glance, it would probably be the most important result that could come out of this study, as it would be an indication of a 'construct crisis' that would have critical implications for the future of neuroscience. In particular, such a negative result would imply that a good deal of research efforts and funding are going to projects that are unlikely to generalize beyond the particular tasks that they employ. If we were to find such a result in this study, we would examine which individual tasks in our battery show the strongest relationships with real-world measures, and adjust our task inclusion for Aims 2 & 3 to focus on the tasks with the highest direct real-world relevance.

Aim 2: Identify the biological origins of belief updating in healthy adults. Despite a considerable number of studies examining the biological basis for belief updating, the exact mechanisms through which the brain updates beliefs according to new information remain unclear. While correlative studies using fMRI, EEG, and pupillometry have identified neural signatures of belief updating (Behrens et al., 2007; Fischer & Ullsperger, 2013; Jepma et al., 2016; McGuire et al., 2014; Nassar et al., 2012; Payzan-LeNestour et al., 2013; Vilares et al., 2012), more recent work has shown that such signatures are conditional on the exact structure of the specific belief updating task (d'Acremont & Bossaerts, 2016; Nassar, Bruckner, et al., 2019; O'Reilly et al., 2013). In particular, relationships between neural signaling and belief updating that have been observed in tasks with persisting changes have been shown to be absent, or even reverse, in tasks where occasional outliers require minimizing the influence of outlying data to improve the robustness of beliefs (Cheadle et al., 2014; d'Acremont & Bossaerts, 2016; Nassar, Bruckner, et al., 2019; O'Reilly et al., 2013). As described above, we have developed a model that updates latent states rapidly in the face of surprising feedback to explain belief updating across statistical environments (Razmi & Nassar, 2022). A key advantage of this model is that it explains the discrepant empirical observations in the literature by showing that: 1) latent state transitions can speed learning in environments where new latent states persist in time (changepoints); but 2) slow learning in environments with so-called oddballs that reflect transient changes in the latent state (Razmi & Nassar, 2022, see figure 1). The latent state transitions that enable the model to capture behavior match the observed characteristics of feedback locked P300 signals (Nassar, Bruckner, et al., 2019) and pupil dilations (Nassar et al., 2012; O'Reilly et al., 2013) across changepoint and oddball conditions, providing two potential neurophysiological markers for latent state transitions in the brain.

We have recently developed a biologically constrained version of the model that represents latent states in a prefrontal recurrent neural network (RNN). Learning occurs through associations between these latent states and medium spiny cells in the striatum that pass learned beliefs through a striato-thalamic-cortical loop to (motor) cortex. Conflict between supervised feedback and motor cortical predictions is used to drive thalamic neurons that project to the prefrontal RNN to promote attractor state transitions (Fig 3A).

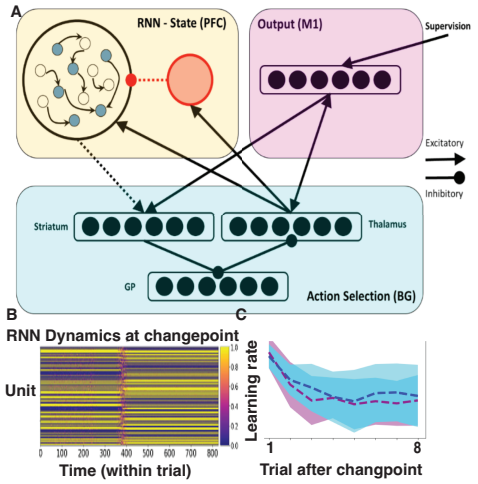


Figure 3. Contextual belief updating using cortico-striatal-thalamic loops. A) Recurrent neural networks (RNNs) in cortex represent attractor states (yellow) and project onto striatal synapses (blue, striatum) that are incrementally updated in accordance with supervised feedback to a cortical output layer (pink). Thalamic units receive conflict signals from the output layer to drive feedback inputs to the RNN. **B)** When conflict is high, it can cause the RNN to shift to a new attractor state (left), which allows the model to rapidly update beliefs at changepoints. Model simulations (purple) can successfully reproduce human behavior (blue).

Our preliminary model simulations suggest that this biologically constrained model can capture belief updating behavior in a predictive inference task (Fig 3B). In particular, the model rapidly adjusts beliefs after changepoints because thalamic units respond to conflict between predicted and observed outcomes by driving changes in the cortical attractor state (latent state reset). During periods of stability, when predictions are closely matched to supervised feedback, thalamic inputs stabilize the attractor state in the PFC such that belief updating is achieved through incremental learning at cortical-striatal synapses (incremental learning). Thalamocortical connections in the model are inspired by recent work showing two pathways that project from medial dorsal thalamus to prefrontal cortex in rodents that antagonistically control the stability of cortical networks (Mukherjee et al., 2021). A key finding is that one of these pathways contains neurons that respond to anticipated conflict and act through parvalbumin positive interneurons, which effectively control excitation/inhibition (E/I) balance, to stabilize active cortical representations (Mukherjee et al., 2021). From this biological perspective, a natural question is whether thalamocortical connectivity, or the E/I balance in prefrontal regions, might underlie individual differences in how individuals update beliefs according to new information (Nassar et al., 2010).

Here we will test these ideas using MR spectroscopy to non-invasively measure levels of glutamate and GABA, giving us an indirect, albeit crude handle on individual differences in E/I balance in regions thought to represent latent states, and fMRI to probe cortico-thalamic connectivity. Individual differences in E/I balance and corticothalamic connectivity will be linked to individual differences in belief updating behavior in the predictive inference task and iterative trust game that were part of the larger behavioral battery in Aim 1 (Fig. 4). In previous work using an iterative trust game, we used fMRI to identify prefrontal 'state' representations that translate experienced feedback into behavioral adjustments (Lamba et al., 2022). Here, we will examine whether such representations: 1) transition rapidly at reversals in the trustworthiness of a partner; 2) are accompanied by changes in corticothalamic connectivity; and 3) transition more frequently in people with higher cortical E/I balance. Our previous work with the predictive inference task has identified pupil dilations and feedback locked P300 as convergent markers for transitions in latent states. We will therefore collect EEG and pupil data during task performance (Nassar et al., 2012; Nassar, Bruckner, et al., 2019; Razmi & Nassar, 2022), allowing us to examine whether increased E/I balance and abnormal cortico-thalamic connectivity alter belief updating behavior by promoting latent state transitions when none should be necessary.

Eighty participants will be recruited for a two-session study giving 95% power to detect correlations of R=0.4 (similar to behavior-MRS relationships in our preliminary data) and >99% power to detect effects of R=0.55 (similar to our previous relationships between behavior and pupil- and EEG-proxies for latent state transitions). In the first session, we will acquire T1-weighted MPRAGE structural images from each participant, followed by spectroscopy using PRESS and megaPRESS sequences to measure glutamate and GABA concentrations, respectively, in orbitofrontal and prefrontal regions previously shown to represent latent states (Lamba et al., 2022; Nassar, McGuire, et al., 2019). After MRS, participants will undergo fMRI while performing an iterative trust game. A key feature of the task is that it will involve two types of partners that are encountered in short blocks. One type of partners is highly variable in their returns, whereas the other type of partners is more precise in their returns, but undergoes periodic reversals in their level of returns over the course of the experiment, thereby creating different latent states (Fig. 4). These two blocked conditions will allow us to assess thalamocortical connectivity during periods that demand stable latent states, which we propose will depend on conflict-sensitive thalamic neurons that act to stabilize cortical representations via parvalbumin positive interneurons, and during periods that demand latent state transitions, which we propose will depend on GIRK-mediated thalamocortical signal amplification pathways (Mukherjee et al., 2021). During the second session, participants will complete a predictive inference task that includes both changepoint and oddball contexts (Nassar, Bruckner, et al., 2019) while EEG and eye-tracking data are recorded. A key advantage of the predictive inference task is that it requires participants to report their beliefs directly, thus providing model-free measures of belief updating that can be used to test and validate models (Nassar & Gold, 2013). MRS, fMRI, EEG, pupillometry, and behavioral data will be combined to test the following hypotheses:

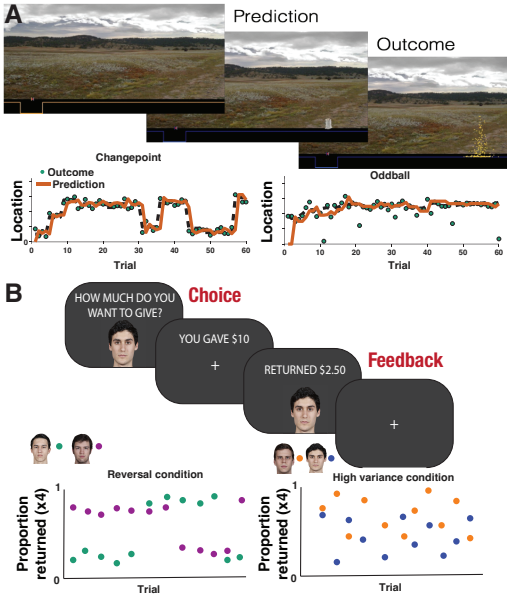


Figure 4. Belief updating tasks. A) Predictive inference task requires participants to predict the location of a hidden helicopter that drops bags of gold on each trial. In the changepoint condition, (left) the helicopter undergoes occasional abrupt changes, whereas in the oddball condition the helicopter moves slowly, but bags are occasionally unrelated to helicopter position. **B)** Iterative trust game requires participants to decide how much to invest in a partner who will receive 4X the investment and return some proportion. In reversal blocks, mean return will occasionally reverse (left) whereas in the high variance condition mean returns will be stable for each partner, but include considerable variability across trials.

H2.1: Cortical E/I balance predicts individual differences in thalamocortical connectivity. Glutamate and GABA concentrations will be measured from OFC and PFC ROIs, processed using LC model, and E/I balance will be computed according to the normalized abundance of the two neurotransmitters (Shibata et

al., 2017). Functional neuroimaging data will be preprocessed and projected into the MNI standard space, where medial dorsal thalamic (MD) activations will be extracted using a probabilistic map of thalamic nuclei (Iglesias et al., 2018). Connectivity between the two regions will be assessed through a psychophysiological interaction analysis (PPI). In the PPI, the cortical BOLD signal is regressed onto an explanatory matrix that contains: trial and feedback onsets, model-based modulatory terms that capture effects of partner identity and reward prediction error (Lamba et al., 2022), the MD timeseries, and the interaction between the MD timeseries and block type (those requiring stabilization versus switching of latent states). We hypothesize that cortical E/I balance will predict the degree to which MD and prefrontal signals are coupled during task performance, with high E/I balances leading to greater coupling between MD and cortical regions in the latent state switching context, and low E/I balances leading to greater coupling between MD and cortical regions during the stable latent state context. We refer to this difference in connectivity, which will be directly assessed in our PPI, as differential connectivity.

H2.2: Cortical E/I balance and thalamocortical connectivity predict individual differences in EEG- and pupil-based measures of latent state transitions. If E/I balance and differential connectivity control the stability of cortical attractor states, then they should also determine how big of an error is required to trigger a latent state transition. With this in mind, we will use a data-driven approach to extract single trial measures of P300 and pupil diameter that have previously been related to state transitions in the predictive inference task (Nassar et al., 2012; Nassar, Bruckner, et al., 2019; Razmi & Nassar, 2022). We will examine how these measures scale as a function of error magnitude in both changepoint and oddball contexts. We will test whether P300 and pupil dilations are more sensitive to small errors in participants who have higher E/I balance and differential thalamocortical connectivity. Observing this effect would confirm a key prediction from our model and support the notion that individual differences in thalamocortical signaling can control whether discrepant observations are seen as coming from the same or different latent states.

H2.3: Cortical E/I balance and thalamocortical connectivity predict individual differences in belief updating. If higher cortical E/I balance shapes cortico-striatal signaling to increase the number of latent state transitions, this should lead to increased belief updating when such transitions persist in time (changepoints) but less belief updating when transitions are transient (oddballs). We will test this by measuring the degree to which participants updated their predictions toward the most recent outcome after changepoints and oddballs. We will examine whether 'learning rates', which capture this updating, are higher in the changepoint condition and lower in the oddball condition for individuals with higher E/I balance and differential connectivity. We predict a similar behavioral dissociation in the iterative trust game, with high E/I balance and differential connectivity improving performance immediately after trustworthiness reversals, but hindering performance with stable but variable partners, where integration over a large number of interactions is required to accurately assess trustworthiness. We further hypothesize that high E/I balance will predict real-life behavior in our belief updating scale (see Aim 1), such that individuals with high E/I will score higher in questions describing new environment contexts.

H2.4: EEG- and pupil-based measures of latent state transitions mediate the effect of E/I balance and thalamocortical connectivity on individual differences in belief updating. Our model predicts that structural measures (E/I balance and differential thalamocortical connectivity) can impact behavior (belief updating on observing discordant information) by affecting the likelihood of cortical latent state transitions. Thus, we predict that our measures of latent state transitions (single trial P300 and pupil dilation) will also predict learning, and statistically mediate the relationships between brain measures and behavior.

H2.5: Biological model of belief updating will be extended to incorporate experimental observations Our experiments will not only test core assertions of the model, but also provide quantitative constraints on key variables, such as the frequency of latent state transitions and their impact on behavior. After obtaining experimental data, the model will be refined to better align it with new quantitative constraints and then used to make more nuanced predictions that will be backtested in the data.

Preliminary data supporting our ability to execute Aim: Our ability to execute this Aim is supported by preliminary data demonstrating the ability of our cortico-striato-thalamic loop to switch attractor states to achieve contextual inference (Figure 3). Moreover, our previous work demonstrates our ability to measure

pupil dilations and P300 signals that track latent state transitions (Nassar et al., 2012; Nassar, Bruckner, et al., 2019; Razmi & Nassar, 2022), and to reliably measure glutamate and GABA with MRS (Fig. 5).

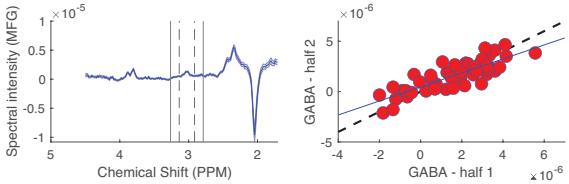


Figure 5. Reproducible GABA measurements with MRS. Left: Average spectral intensity (ordinate) from a megaPRESS sequence to isolate GABA (3 PPM peak on abscissa) in a prefrontal voxel. **Right**: Split half reliability of GABA quantification via peak integration was R=0.8 for N = 49 participants (data points).

Potential pitfalls/alternative outcomes: The principal potential pitfall in this aim relates to our use of MRS to measure bulk concentrations of GABA and glutamate. Indeed, the primary strength of this proposal, which is to directly relate biological measures to real-world constructs in humans, limits our use of more invasive direct biological measures in human participants. That said, our derived measure of GABA is supported by a considerable body of work linking it to inhibition (Stagg et al., 2011; Takado et al., 2022), our preliminary data showing its stability and reliability (Fig. 5), and the fact that parvalbumin positive interneurons are the most abundant GABAergic cell type in prefrontal cortex (Chung et al., 2016). Nonetheless, we will attempt to mitigate this risk by computing other EEG-based proxies for E/I balance (Gao et al., 2017) and by examining other EEG measured gamma oscillations that are thought to depend specifically on parvalbumin positive interneurons (Gonzalez-Burgos et al., 2015).

Aim 3: implement our computational framework in patients with schizophrenia. In Aim 3, we will examine how the cognitive and neural mechanisms explored in Aims 1 and Aim 2 play out in schizophrenia patients. The aim is to develop a more ecologically relevant computational framework through which biological and behavioral measures collected in the lab could be used to explain the emergence of pathologies in belief updating, as epitomized in delusions.

As described above for studies in the healthy population, Bayesian integration models have been used to explore the underlying mechanisms of abnormal belief updating in psychosis (Fletcher & Frith, 2009, Sterzer et al., 2018). At first look, this framework maps onto delusions seamlessly, particularly in the social domain where delusions, such as persecutory delusions, are most common in patients (Stuke et al., 2021). As delusions are fixed beliefs that are not updated in the face of new evidence, they can be explained by an overreliance on priors with minimal influence of new sensory evidence or likelihood. Indeed, several studies have found evidence for this hypothesis (Schmack et al., 2017; Stuke et al., 2021; Teufel et al., 2015). For example, psychotic individuals are better able to use prior information for detecting human figures in ambiguous two-toned images (Teufel et al., 2015). Similarly, the strength of the biasing effect of experimentally induced beliefs about stimulus motion direction positively correlates with the severity of psychotic symptoms in schizophrenia (Schmack et al., 2017).

There is, however, a large body of literature suggesting that delusions are associated with an under-reliance on prior beliefs, and an exaggerated influence of sensory evidence on belief updating. For example, weaker influence of prior beliefs was found in both patients with schizophrenia and healthy individuals who are more prone to delusions (Schmack et al., 2015; Stuke et al., 2019). Further evidence in favor of overreliance on sensory evidence comes from the 'jumping to conclusion' phenomenon in schizophrenia (Dudley et al., 2016). That is, using decision-making tasks, such as the 'beads task', researchers have shown that patients with delusions make 'hasty' decisions that are biased towards the minimally sampled sensory information (Dudley et al., 2016). Moreover, in an n-of-1 study (Gabler et al., 2011), we tested a patient with treatment-resistant schizophrenia, who had bilateral deep brain stimulation electrodes in his ventral tegmental area. We found that the influence of priors on performance beliefs (inversely related to prior variance) changed as a function of stimulation parameters. Importantly, the precision of priors closely matched the patient expression of positive symptoms (Fig. 6), such that higher variance and thus weaker the priors were related to more severe psychotic symptoms (Wolpe et al., in preparation).

Together, these results show contrasting evidence for over- or under-use of prior in belief updating in patients with delusions. Attempts to resolve this 'paradox' (Furl et al., 2022) suggest time-dependent effects (Haarsma et al., 2020), whereby proneness to delusion develops as a result of weak priors, but with a transition to strong priors as delusions form (Corlett & Fletcher, 2021; Sterzer et al., 2018). However, it is

not easy to reconcile why such a transition would occur. Another explanation has emphasized the hierarchical nature of priors, whereby experiments may capture priors of different 'levels', which can have different precision and therefore different influence on decision-making and perception (Adams et al., 2013; Corlett et al., 2010; Diaconescu et al., 2020; Sterzer et al., 2018). We find the latter explanation particularly appealing, and we consider our contextual Bayesian inference model to be one implementation of this idea. By including categorical contexts or latent states, our model can respond to disconfirmatory information, either by incrementally updating beliefs about the active latent state, or by inferring that the active latent state has changed. In principle, a predisposition toward latent state switch due to increased E/I balance might lead to a behavior consistent with 'jumping to conclusions' (under-use of prior) in conditions where environmental context has in fact changed ('changepoints'). But when environmental context switches are transient ('oddball'), such a predisposition will lead to slower belief updating (over-use of prior).

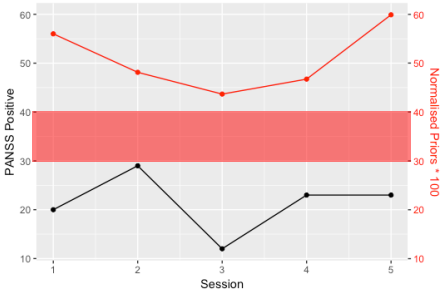


Figure 6. Weaker performance belief priors correspond to positive symptom severity. Preliminary data from n-of-1 patient study, showing changes in the variance of performance belief priors (normalized by performance variance) corresponding to positive symptom severity (measured using the Positive and Negative Syndrome Scale). Variance of performance belief priors in controls is shown in shaded red (mean ±1 standard error).

Our computational framework for abnormal belief updating in schizophrenia is not the only instantiation for the possible consequences of E/I imbalance on behavior (Jardri et al., 2017). But one interesting point of divergence between our framework and other Bayesian inference models applied to schizophrenia is that contextual inference need not always include integration of new and old information. That is because in many cases, new experiences in new trials will be erroneously associated with new latent states. Fluctuations in trial-by-trial behavior in belief updating tasks seem to provide some support for this idea. Across trials of an experiment, behavior appears to reflect, on average, an integration between sensory evidence and priors that can be captured by standard Bayesian models. However, when considered on a trial-by-trial basis, people often use an 'all-or-nothing' belief update strategy (Nassar et al., 2021), whereby only a single (previous) sample is used to make inferences. That is, on a given trial, people may not integrate prior experience with sensory evidence and instead rely completely on either the previous sample (the prior) or on current sample (the sensory evidence). In schizophrenia, stably medicated patients show an exaggerated tendency for such an all-or-nothing behavior in in our predictive inference task (Fig. 7A) (Nassar et al., 2021). Similarly, we found such a tendency of all-or-nothing belief updating in patients with first episode psychosis (Fig. 7B) (data kindly shared by Haarsma et al., 2021). These findings suggest that previous research may have overlooked an important and common, yet poorly characterized neurocomputational mechanism of belief updating which is captured by our computational framework. Namely, 'all-or-nothing' updates (i.e., updates that reflect only a single piece of new or old information) might be produced by overly frequent context switches in our contextual inference model.

Our biological implementation of contextual inference with cortico-striato-thalamic loops gives some insight into why this behavioral phenotype might emerge in schizophrenia. Attractor transitions in our model function to assign data to an appropriate context using thalamocortical projections. This is based on animal work showing that one of the two thalamic cell types projects to cortical parvalbumin positive interneurons to increase activation with expected conflict to stabilize cortical attractors (Mukherjee et al., 2021). However, this pathway is likely

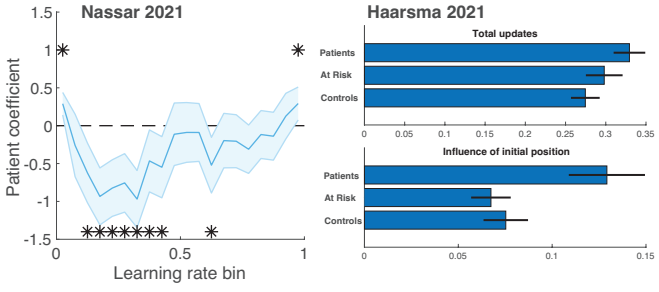


Figure 7. All-or-nothing belief updating in schizophrenia and psychosis. A) Predictive inference task data from Nassar 2021 (*Brain*) showing that schizophrenia patients tend to overuse both very large and very small belief updates (learning rates of 0 and 1 on abscissa) relative to controls. **B)** Data from Haarsma 2021 reanalyzed to show a similar pattern, with patients overusing total updates on one hand (all), but on the other hand tending to be more influenced by their randomly initialized position (nothing).

disrupted in schizophrenia, where prefrontal parvalbumin positive neurons are reduced (Chung et al., 2016; Gonzalez-Burgos et al., 2015). In our model, this disruption would lead to frequent attractor switching, even in seemingly stable environmental context, that could produce all-or-nothing updating, and prevent observations to be integrated into a single representation. Given that the thalamic inputs to this pathway depend on the expected level of conflict, we would predict a loss of parvalbumin positive neurons to be particularly deleterious in situations with high variability, or noise, where thalamo-cortical projections would typically play an important role in stabilizing cortical attractor representations.

Our computational modeling approach can therefore potentially resolve a tension between under- and over-reliance on priors in previous accounts of delusions and explains how both under-updating and over-updating phenomena might co-exist even within a single context. In Aim 3, we test key predictions of our model to investigate the neuro-computational mechanisms underlying abnormal belief updating.

A cohort of 80 patients with a DSM 5 diagnosis of schizophrenia and 80 sex- and age-matched controls will be recruited for a one-session study. This sample size will give 88% power to detect a medium effect size of group difference d=0.5 and 78% power to detect within-patient group correlation of R=0.3 (effect size assumed to be similar or more stringent than our previous patient studies). PI Wolpe is a clinical psychiatrist who will lead patient recruitment in Tel Aviv. Participants will complete the predictive inference and iterative trust game tasks described in Aim 2, with the exception that the predictive inference task will include two different noise levels of the changepoint condition, rather than including an oddball condition. As described above, this variant of the task can better test the specific predictions of our biological model, that is – as noise increases the lack of parvalbumin positive neurons in PFC will lead to greater latent state transitions (seeing higher conflict without stabilizing thalamocortical signals).

As in Aim 2, EEG and pupillometry data will be acquired while participants perform the predictive inference task. To facilitate wider patient participation and a larger sample size, we will omit the MR scan session. In-depth clinical assessment will involve the assessment of delusions (Peters Delusions Inventory, PDI (Peters et al., 1999)), negative symptoms (Brief Negative Symptom Scale (Kirkpatrick et al., 2011)), cognitive abilities (Brief Assessment of Cognition, BACS (Keefe et al., 2008)), depression (Calgary Depression Scale (Addington et al., 1993), self-reported motivation (Apathy Evaluation Scale, AES (Marin et al., 1991)), and real-life belief updating behavior (Belief Updating Scale, see Aim 1 preliminary data). Behavioral data will be fit with hierarchical Bayesian models developed in Aim 1 and synthetic data will be created using different perturbations of the cortico-striato-thalamic loop model from Aim 2.

H3.1: Patients will show exaggerated all-or-nothing belief updating that is attributable to a high base rate of latent state transitions. Similar to our preliminary data (see above), we expect patients to show an exaggerated behavioral pattern of trial-by-trial all or nothing belief updating in both the predictive inference and iterative trust game tasks. Specifically, compared to controls, patients will show fewer trials where priors and sensory evidence are integrated, and more trials where they rely completely on either the previous trial ('all') or their previous belief or investment ('nothing'). We predict that this all-or-nothing behavioral pattern will be even more pronounced in the high vs. low noise condition, providing us an additional internal control condition to test our specific hypothesis. When fit with the latent state variant of the Bayesian hierarchical model described in Aim 1, we expect that patients will be fit with higher base rates for latent state transitions. In posterior predictive checks, this will capture all-or-nothing belief updating by promoting both persisting latent state transitions (contributing to 'all' updating) and transient latent state transitions (contributing to 'nothing' updating).

H3.2: Base rate of latent state transitions will be enhanced in the high noise condition and correlated with the severity of delusions as measured by PDI. Given that schizophrenia patients have reduced parvalbumin positive interneurons in PFC, and that these neurons play a critical role in mediating stabilizing effects of thalamocortical signals when conflict is expected to be high (Mukherjee et al., 2021), our biological model would predict that patients would have particular problems maintaining attractor states (and thus latent states) in the high noise condition. We will test this idea by 1) examining all-or-nothing updating behaviors across the two conditions and quantifying their difference; and 2) adding parameters to the hierarchical model allowing base rate and sensitivity of latent state transitions to depend on underlying noise condition (i.e., expected conflict). We expect that patients will have exaggerated all-or-nothing

updating and base rate parameter fits in the high noise condition, and that conditional differences in these measures will relate to severity of delusions as measured by PDI.

H3.3: Patients will show increased frequency of neural markers of latent state transitions. We will use feedback locked P300 potentials measured by EEG and pupil diameter changes measured by an eyetracker (as described in Aim 2) to infer when participants undergo latent state transitions. As in Aim 2, we will take a data-driven approach to quantifying trial-to-trial P300 and pupil responses (Krishnamurthy et al., 2017; Nassar, Bruckner, et al., 2019). We will test the degree to which these trial-to-trial measures scale with overall prediction error magnitude separately in the two different noise conditions. We predict that patients will have a weaker overall scaling of both measures with prediction errors, as would occur if many of the small error trials drove inappropriate neural state transitions. Most notably, we expect to see a reduced difference between the high and low noise conditions in the patients, consistent with reduced thalamocortical stabilization signaling when high conflict is expected.

H3.4: Specific computational lesions will reproduce behavior and neurophysiological data of patients. The cortico-striato-thalamic loop model will be modified to reflect the effects of a reduction in prefrontal parvalbumin positive neurons, characteristic of schizophrenia (PVi-depleted model) (Gonzalez-Burgos et al., 2015). We will also examine the effects of parvalbumin positive neurons on thalamo-cortical signaling as well as on local cortical processing. The goal of this exercise will be to test whether the depletion of parvalbumin positive neurons in our model can reproduce belief updating and neurophysiological changes that mimic those observed in patients, and if so, whether it does so through effects on cortico-cortical or thalamocortical signaling in the model. To this end, we will run simulations from PVi-depleted models along with those from our original model. We hypothesize that the PVi-depleted model will show destabilized prefrontal attractor networks that will facilitate rapid attractor switches. Behaviorally, this will lead to rapid latent state transitions and an exaggerated all-or-nothing belief updating, particularly in the high noise condition. We also predict that trial-to-trial attractor switches in the model correspond to pupil dilation and P300 signals in participants, consistent with those signals providing a readout of neural latent state transitions. With this in mind, we will test whether the PVi-depleted model transitions match patient behavioral and neurophysiological (EEG/pupil) data better than the full PVi-normal model.

Potential pitfalls/alternative outcomes: The main pitfall with patient testing could be a failure to recruit the anticipated number of volunteers. To overcome this, PI Wolpe can leverage his clinical connections with the Cambridgeshire and Peterborough NHS Foundation Trust, as part of the Cambridge Psychosis Centre and active collaboration with Professor Paul Fletcher (collaboration letter attached to submission). If recruitment in Tel Aviv fails to reach the aimed sample size, we plan to widen patient recruitment to Cambridge Department of Psychiatry, where EEG and eyetracking (albeit different to the Nassar and Wolpe labs) are available. Moreover, the specific predictions made by our computational framework may not be observed in patients. Regardless of whether our model predictions are correct or refuted, given the richness of the behavioral and neurophysiological measures in the study, we expect the study to make significant contributions to our understanding of abnormal belief updating in patients with schizophrenia.

Prior NSF Support: Co-PI FeldmanHall has previously been the recipient of NSF funding (2123469: Cognitive maps as a framework for organizing relationships in large-scale real social networks) on which PI Nassar is a Co-PI. The award was made in Fall 2021 and thus is, at the time of writing, in early stages of data collection, however it has already yielded returns in terms of intellectual merit, including two papers on which Drs. Nassar and FeldmanHall are authors. One of these papers inspired discussions about the need for latent state inference in everyday social interactions that contributed directly to some of the ideas explored in this proposal (FeldmanHall & Nassar, 2021). The funding has also led to broader impacts, including the training of one graduate student and one post-doc, the latter of which will begin a tenure track faculty position at the University of Leiden in January 2023. Nonetheless, given that the project is following the formation and dynamics of social relationships of undergraduate students who just matriculated this semester at Brown University, and thus has only included two months of data collection, we expect that the intellectual merit and broader impacts will be greatly amplified in the coming years.

Coordination plan

Specific Roles of Collaborating Co-PIs

Co-PI Nassar is a computational neuroscientist at Brown University with expertise in learning and decision making. Co-PI Wolpe is a clinician scientist at Tel Aviv University with research expertise in motor and perceptual learning and clinical specialization in psychiatry. Co-PIs Nassar and Wolpe have been collaborating successfully for the past twelve months with several preliminary data figures resulting directly from this collaboration. Both researchers have successfully carried out multi-national collaborative projects in the past, which is highlighted in their biosketches. Co-I FeldmanHall is a cognitive neuroscientist at Brown University with expertise in social decision making who has a history of collaboration with Co-PI Nassar that includes the development of the iterative trust game that will be used in all Aims of the proposal. Co-I Petzschner is an expert in computational psychiatry as well as a leading expert in Bayesian models of perception and is actively engaged in collaborations with Co-PI Nassar and Co-I FeldmanHall across a range of topics.

The entire senior research team will collaborate on Aim 1 of the proposal, with each team member overseeing collection, analysis, and modeling of the belief updating task(s) in their area of expertise. Data collection will occur online using PI-Wolpe's bespoke online testing platform. Recruitment will be carried out via crowdsourcing platforms, such as Prolific, and thus can occur in tandem across both sites. The software developer based in Tel Aviv will provide ongoing support to the greater team, by helping to set up the task battery in the online platform; debugging errors when they arise; and maintaining the testing platform. Collaboration will be facilitated by a post-doc and research assistant at Brown who will work across labs to implement and analyze the belief updating battery. Both the post-doc and research assistant will have a primary appointment and space in the Nassar lab, but will be encouraged to meet in person frequently with Co-Is FeldmanHall and Petzschner regarding social and perceptual tasks, respectively. Group virtual project meetings that include all research personnel (US & Israel) will be held monthly to coordinate activities across all research personnel, including the graduate student and software developer in Tel Aviv who will be participating in collection and analysis of Aim 1 data, particularly for the performance belief and sensorimotor tasks that are in Co-PI Wolpe's area of expertise. Additional interaction across research sites will occur through lab meetings, which in both the Nassar and Wolpe labs are hybrid. Indeed, Co-PI Wolpe has attended the Nassar lab meeting regularly during the preparatory phase of this grant submission, and we anticipate a healthy two-way exchange of ideas through lab meetings over the course of the award, with all research personnel on both sides of the Atlantic invited to participate in lab meeting discussions of relevant data. Co-PI Nassar will lead the development of the Bayesian Hierarchical modeling framework that will be used to make out-of-task predictions for how an individual will update their beliefs in a new task and to identify the task-general constructs that make such predictions possible. These efforts will also include a post-doc at Brown, a graduate student at Tel-Aviv, and Co-I Petzschner.

Aim 2 of the proposal will involve data collection locally at Brown University. Co-PI Nassar will oversee the collection, analysis, and modeling of the data with Co-I FeldmanHall providing additional oversight with respect to the iterative trust game and associated fMRI data. Data collection will involve the Brown post-doc and research assistant, under the supervision of Co-PI Nassar. Modeling and analysis will be implemented by the Brown post-doc with direct mentoring from Co-PI Nassar and Co-I FeldmanHall. Since Aim 2 will be conducted at Brown during the same period where Aim 3 is conducted at the Tel Aviv University, virtual project meetings will occur monthly to facilitate crosstalk between those working on clinical, cognitive, and computational aspects of the research project.

Aim 3 of the proposal will involve face-to-face data collection of patients and healthy controls locally at Tel Aviv University. Co-PI Wolpe will oversee the collection, analysis, and modeling of the data with ongoing input and discussions from the US-based team led by Co-PI Nassar. Data collection will involve the Tel Aviv graduate student and software developer, under the supervision of Co-PI Wolpe. Co-PI Wolpe will closely support the recruitment and testing of patients, particularly in the initial phases of training of the new lab personnel. Modeling, analysis, and write-up will be done in close collaboration with the team in Brown,

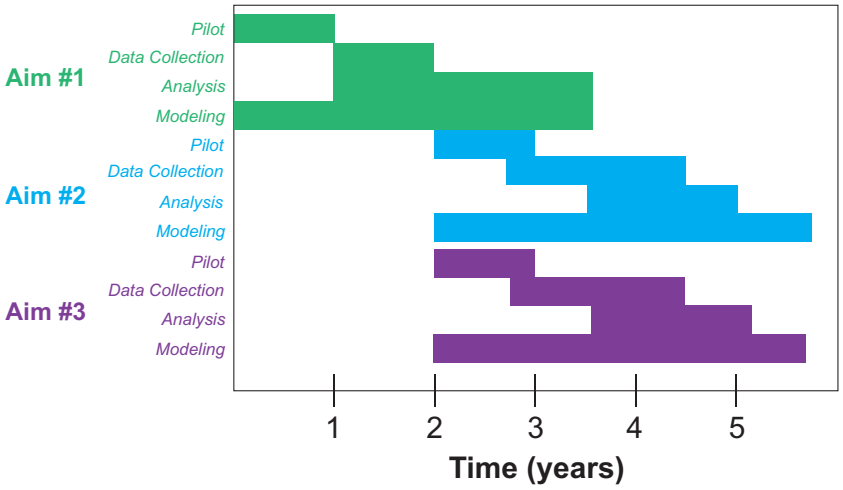
through regular and *ad hoc* meetings both online and face-to-face. We anticipate biannual face-to-face meetings across labs to facilitate individual training and attainment of a wide skillset across the team.

Coordination mechanisms enabling cross-organizational and cross-disciplinary scientific integration.

Cross-disciplinary integration will occur in several ways. As described above, hybrid lab meetings with either computational (Nassar lab) or clinical (Wolpe Lab) emphasis will broaden the training of research staff and allow them to see the project from different perspectives. Monthly virtual project meetings, which will include Co-Is FeldmanHall and Petzschner, will provide an additional channel for crosstalk across disciplines and labs. We also anticipate that one graduate students from the Tel Aviv group will spend time in the US to learn computational modeling from Co-PI Nassar and Co-I Petzschner.

Practical tools for data sharing will include GitHub for code and google drive for data and text (following all protocols for data sharing and protection of subjects described in the data management plan). Coordination of publications will occur through google docs or overleaf, depending on the preferences of the first author.

Timeline for project coordination



Large scale behavioral study data collection will precede and inform in-person MRS/fMRI/EEG/eyetracking studies. Modeling will be used to optimize task design during the pilot phase, simulate results during data analysis, and models will be revised after data collection to take newly observed data into account.