A corticostriatal learning mechanism linking excess striatal dopamine and auditory hallucinations

Kaushik Lakshminarasimhan[1,†], Justin Buck[1,†], and Guillermo Horga[1,2]

[1]Zuckerman Mind Brain Behavior Institute, Department of Neuroscience, Columbia University, New York, NY, USA,
[2]Department of Psychiatry, Columbia University, New York, NY, USA, [†]equal contribution

1 Abstract

Auditory hallucinations are associated with elevated striatal dopamine, but the underlying mechanisms are unclear. We investigated this by developing a circuit model in which corticostriatal plasticity in the ventral striatum is modulated by reward prediction error while that in the sensory-dorsal striatum is modulated by sensory prediction error derived from internal belief. Testing this in recordings from mice performing auditory signal detection revealed dissociable signatures of dopamine signaling in ventral and auditory striatum, suggesting parallel but distinct learning processes in these regions. This dichotomy was mirrored in human behavior in the same task where stimulus-history and reward-history altered future choices by selectively biasing confidence and policy respectively. Intriguingly, only the strength of the former correlated with hallucination propensity in the general population, an effect that was recapitulated by selectively enhancing dopamine levels in the model sensory-dorsal striatum. These findings identify plasticity mechanisms underlying learning of sensory expectations as the link between excess dopamine and hallucinations.

2 Introduction

Hallucinations are false percepts commonly experienced by patients with psychotic disorders. Unfortunately, our poor mechanistic understanding of these symptoms prevents us from improving upon treatments that are often ineffective and poorly tolerated (Sommer et al., 2012). To address this, a clear explanation of how changes in neural substrates drive differences in cognition and phenomenology is urgently needed.

To this end, a growing body of work has linked hallucinations to an altered ability to learn sensory statistics (Friston, 2005; Adams et al., 2013; Corlett et al., 2019). This is thought to disrupt perceptual inference: the combination of learned sensory expectations with sensory evidence to form a percept. Thus, if sensory expectations are overweighted or learned inappropriately, they could drive hallucinatory perception. In line with this hypothesis, in noisy sensory environments (e.g., signal detection tasks) patients with hallucinations consistently show exaggerated biases toward prior expectations regardless of the true sensory evidence (Powers et al., 2017; Cassidy et al., 2018). These results represent substantial progress in our understanding of the cognitive basis of hallucinations, but a complete mechanistic explanation should bridge neural and cognitive levels by accommodating both types of empirical results.

The most replicated neural finding in patients with hallucinations is elevated striatal dopamine function (Kesby et al., 2018; McCutcheon et al., 2019; Cassidy et al., 2019). This finding is challenging to interpret given that dopamine release in different striatal sub-regions is known to have distinct influences on cognition. Direct comparisons of dopamine release in the striatum in rodents show that in contrast to the the classical role of the ventral striatal dopamine in reward learning, dopamine in the auditory striatum is specifically sensitive to the sensory features of the environment (Menegas et al., 2017; Schmack et al., 2021; A. Chen et al., 2022). Intriguingly, the latter sub-region also receives dense input from sensory cortices (A. P. Chen et al., 2021) suggesting it may be a functional analogue of the human associative striatum, where dopamine excess in patients is most pronounced. While such findings indicate that sensory striatal dopamine plays a direct role in perception independent of reward and motor function, the precise computation it instantiates is unclear.

Here, we take a theory-driven approach to constructing a circuit-level computational model of hallucinations. Our central hypothesis is that sensory striatal dopamine signals violations in sensory expectations (i.e., sensory prediction errors). This inspires the corollary hypothesis that excess striatal dopamine in patients biases learning of sensory expectation thereby increasing the probability of experiencing false perceptions. We first strengthen the theoretical validity of the hypotheses by demonstrating that an optimized model learns to represent the nigrostriatal dopamine signal as a sensory prediction error. We then validate key predictions of this circuit model using dopamine recordings from behaving rodents. Finally, we simulate selective alterations in dopamine function and find that behavioral data in humans with varying degrees of hallucination propensity is consistent with model predictions of excess sensory striatal dopamine function. Together, the results suggest that a learning-based perpspective of heterogeneous striatal dopamine signaling can account simultaneously for hallucination-related biological, cognitive, and phenomenological phenotypes.

3 Results

Hallucination proneness is characterized by an increased tendency to exhibit false alarms with high confidence in perceptual tasks like signal detection. To investigate the mechanisms by which excess striatal dopamine leads to confident false alarms, we first take a computational view of the broader class of perceptual decision-making tasks. In such tasks, the state of the world is typically encoded in a series of noisy observations that unfold in time, and choosing an action yields a stochastic performance feedback or reward. One must therefore perform inference to estimate the underlying state from observations and then select an action based on the inferred state. The two processes impose unique constraints on learning. Optimal inference entails learning the statistics of incoming stimulus (i.e., the *sensory expectation*) while selecting optimal actions in the inferred state requires learning the outcome statistics (i.e., the *value*) to maximize reward (**Figure 1B**). Since neither stimulus nor reward statistics are stationary in the real-world, a normative approach to flexible perceptual decision-making should support learning of both types of statistics. Given that inference and action selection are cascaded processes, one could, in principle, bypass explicit learning of both the sensory expectation and value by relying on implicit trial-and-error learning based on performance feedback. However, previous studies have revealed engagement of both learning processes in tasks involving temporal and auditory judgments, suggesting that perceptual decision-making is a two-stage process (Sohn & Jazayeri, 2021; Loewenstein et al., 2021). What biological mechanisms subserve the two learning objectives?

From a neurobiological standpoint, dopamine is known to support learning by modulating plasticity in corticostriatal synapses (**Figure 1C** – left). However existing models, informed by reinforcement learning theories focus primarily on the role of dopamine in value learning. In such models, learning is mediated by reward prediction errors (rPE) signalled by the mesostriatal pathway i.e., midbrain dopamine neurons that project from the ventral tegmental area to the ventral striatum (VS). The role of dopamine signalling in the nigrostriatal pathway i.e., from the subtantia nigra to the dorsal striatum is less clear. Patients with schizophrenia show elevated dopamine levels specifically in striatal subregions targeted by this pathway, where dopamine synthesis is correlated with psychosis severity (Laruelle et al., 1996; Cassidy et al., 2018, 2019). Furthermore, studies in mice indicate that neurons and dopamine signals in the nonmotor subregions play a causal role in perception (Guo et al., 2018; Wang et al., 2018; A. Chen et al., 2022), and that dopamine stimulation induces hallucination-like percepts (Schmack et al., 2021). For simplicity, we refer to this subregion as the sensory striatum (SS) throughout the paper and use more specific terminology when referring to specific results from previous work. Based on the normative learning constraints outlined earlier, we hypothesized that rPE signals in VS dopamine facilitate learning of value while SS dopamine signals sensory prediction errors (sPE) – a specific type of prediction error appropriate for learning expectations about the stimulus.

Temporal difference reinforcement learning specifies the precise algebraic form of time-varying rPE signals but the composition of sPE signals remains ambiguous. A recent study shows that accurate posterior beliefs emerge in recurrent neural networks trained to compute the value of stimuli without explicit learning of stimulus statistics but the synaptic weight updates in this model rely on backpropagating error

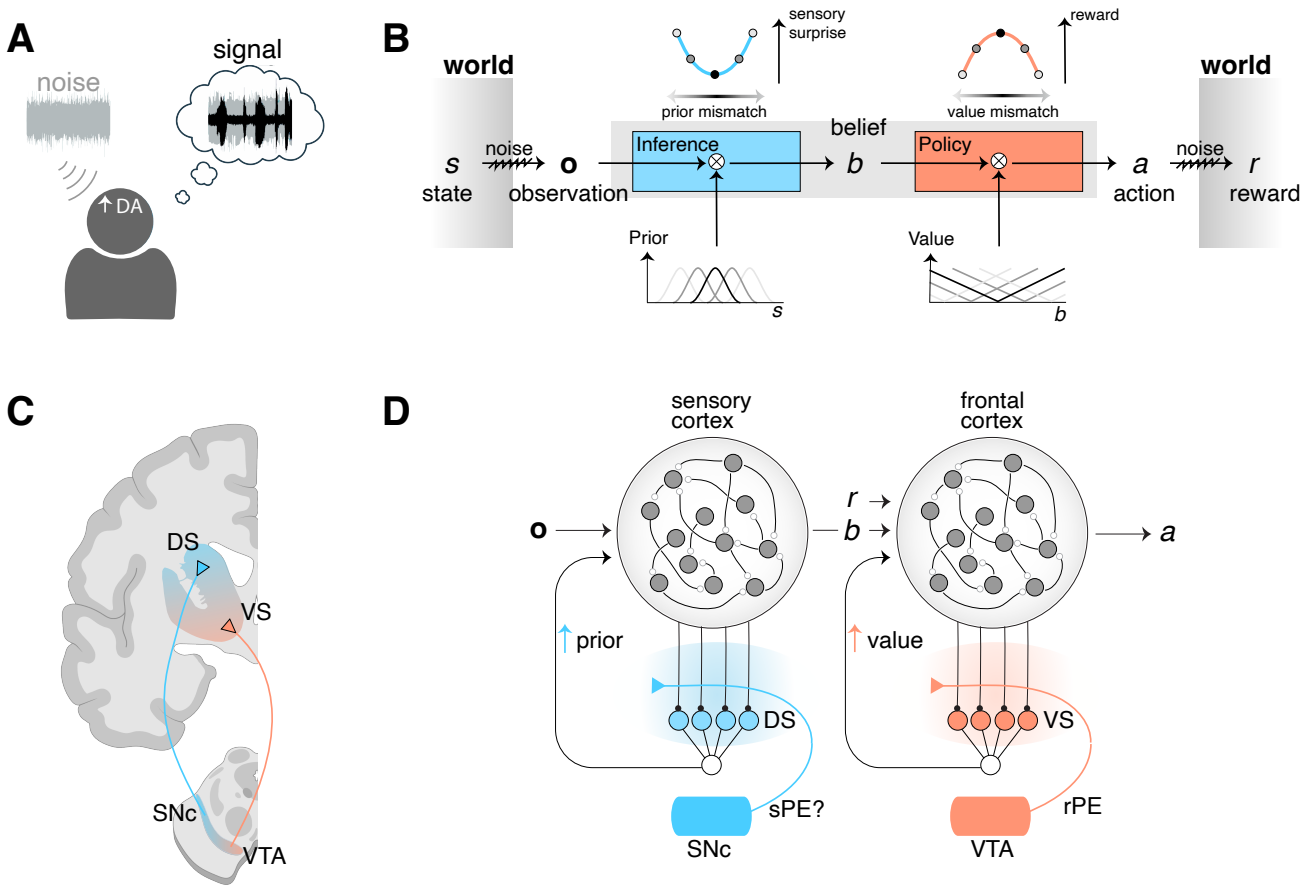

Figure 1: **Modeling framework for auditory hallucinations**. **A.** Excess striatal dopamine is linked to hallucinations, which manifests as an increase in confident false alarms in signal detection tasks. **B.** Perceptual decision-making entails inference and action selection stages that must fulfill distinct desiderata – minimizing sensory surprise and maximizing reward, objectives that can be met by learning the true expectation over states (i.e., prior) and rewards (i.e., value) respectively. **C.** Left: Distinct dopaminergic pathways target the ventral striatum and the sensorimotor areas of the dorsal striatum. Right: Circuit-level mechanisms of the computations in (B) can be implemented in a biologially plausible manner via plasticity mediated by distinct dopamergic prediction error signals in parallel corticostriatal loops. VS: ventral striatum, DS: Dorsal striatum, VTA: ventral tegmental area, SNc: Substantia nigra pars compacta, DA: dopamine, rPE: reward prediction error, sPE: sensory prediction error.

gradients through time and thus not biologically plausible (Hennig et al., 2023). One potential solution is provided by a class of models that use sensory temporal difference errors to learn a successor representation of the states (Gardner et al., 2018), but constructing such prediction errors requires knowledge of the stimulus identity which may not always be available to the animal. While the identity can be inferred from rewards in instrumental tasks, experiments suggest that relative rates of different stimuli can be learned even in the complete absence of performance feedback (Zylberberg et al., 2018). Therefore, we asked whether sPE could instead be computed without external supervision. To address this, we derived an error signal for accurately predicting the cumulative change in the internal belief induced by sensory input, on average. Unlike rPE signals which depend on reward, the resulting sPE signals depended

on the momentary changes in internal belief (Methods – Equation 1). To test whether this signal can mediate biologically plausible learning and to generate predictions at the neural level, we constructed a corticostriatal circuit model informed by neuroanatomical constraints (**Figure 1D** – right).

3.1 Modular corticostriatal model of perceptual decision-making

In the model, distinct midbrain nuclei—substantia nigra and ventral tegmental area—project to distinct striatal regions—SS and VS—to mediate plasticity in corticostriatal synapses according to a biologically plausible three-factor learning rule. According to this rule, the update to a particular synaptic weight W_{ij}^k from cortical neuron j to striatal neuron i in region $k \in \{\text{SS}, \text{VS}\}$ depends only on the pre- and postsynaptic activities, r_j and r_i, and a global prediction error signal δ_k conveyed by the midbrain dopamine inputs to that striatal region: $\Delta W_{ij}^k \propto r_i \phi'(r_j)\delta_k$, where $\phi(\cdot)$ denotes neuronal nonlinearity (Methods). SS and VS neurons receive inputs from different cortical regions—sensory and frontal—and their summed activities project back to those same regions. Sensory and frontal cortices are modeled as recurrent neural networks, optimized for perceptual inference and value-based decision-making respectively (Methods). Concretely, sensory observations are provided as external input to the model sensory cortex, which outputs a graded estimate of the posterior belief to the model frontal cortex, which in turn outputs a binary yes/no response. Critically, both cortical computations depend on learning in the striatum. Accurate perceptual inference relies on learning of sensory expectations in the SS while accurate value-based decision depends on value learning in the VS. We assume that plasticity in the corticostriatal synapses in SS is modulated by the sPE signal derived above, where the momentary change in internal belief was calculated by taking the time-derivative of the output of the sensory cortex. Plasticity in VS is modulated by a standard rPE signal following previous works.

To test this model, we varied stimulus and reward statistics in a signal detection task by manipulating the proportion of 'signal' trials and the fraction of rewarded correct 'yes' responses across blocks (**Figure 2A**; Methods). Although both manipulations initially disrupted performance, the model adapted to them by adjusting its sensory expectations and value across trials, ultimately restoring performance to baseline levels (**Figure 2B** – left). Concretely, adaptation to manipulations of stimulus and reward statistics was mediated by changes in the output of SS and VS respectively (**Figure 2B** – middle vs right). Examining dopamine signaling in these regions revealed a similar dissociation wherein stimulus and reward manipulations induced lasting changes in stimulus-induced dopamine transients within the model SS and VS respectively, albeit in opposite directions. Increasing the base rate of signal trials decreased the amplitude of signal-evoked SS dopamine after learning (**Figure 2C** – left) whereas increasing the reward associated with a correct 'yes' response increased signal-evoked VS dopamine (**Figure 2C** – right). This dichotomy in striatal dopamine signaling was mirrored in cortical computations where the two types of manipulations resulted in similar sensorimotor adaptation (stimulus → action; **Figure 2D** – left) but through distinct computational mechanisms. Manipulating stimulus statistics exclusively modified the transformation implemented by the sensory cortex i.e., perceptual inference (stimulus → posterior belief; **Figure 2D** – middle) while reward manipulation modified policy computation in the frontal cortex (belief → action; **Figure 2D** – right).

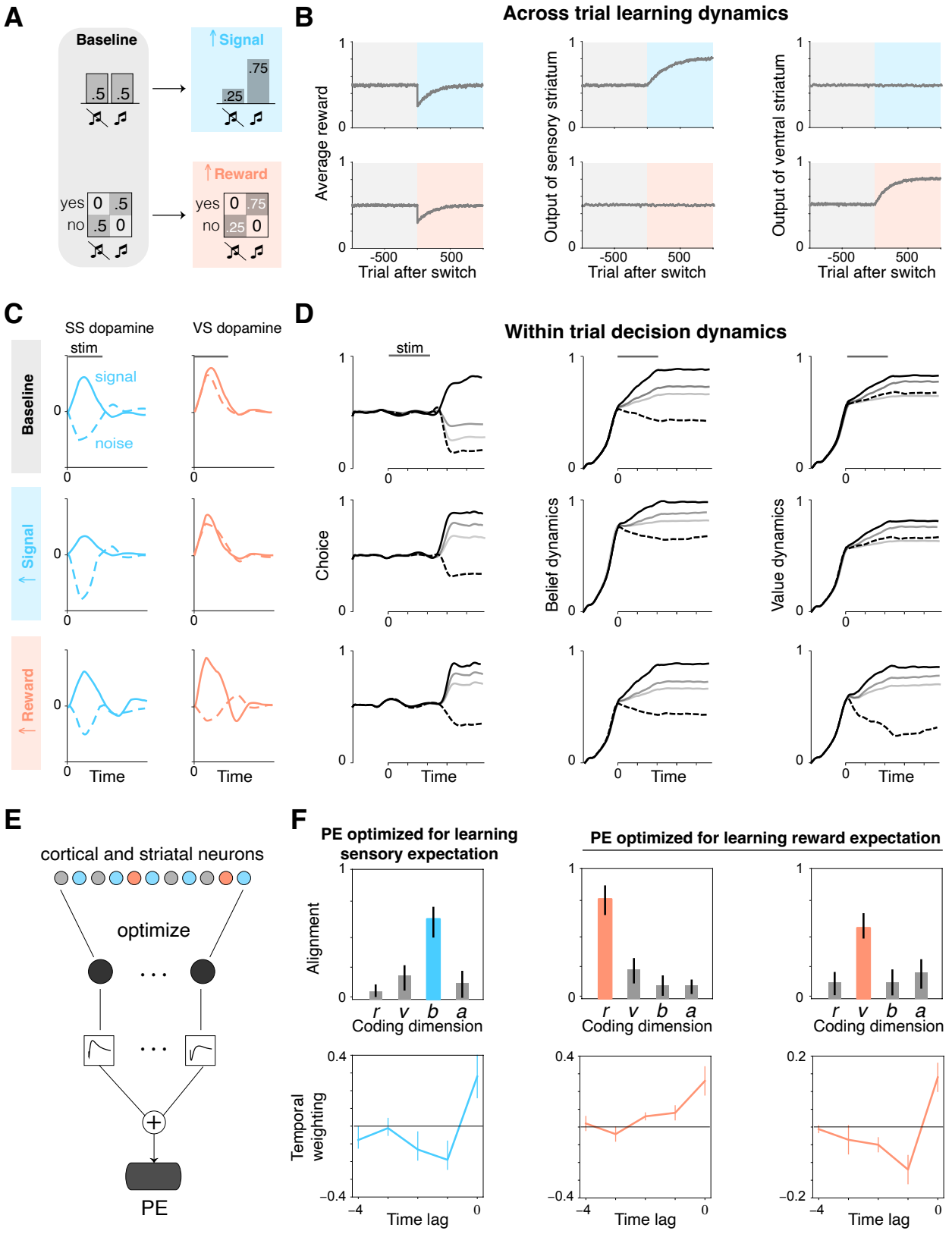

Figure 2: (Caption next page.)

Figure 2: **Distinct computational mechanisms of adaptation to signal and reward statistics**. **A.** Starting from a block of baseline trials (signal/noise stimuli presented in equal proportion and correct yes/no rewarded in equal proportion), stimulus and reward statistics were separately manipulated by increasing the fraction of signal trials (top) and increasing the probability of rewarding a correct 'yes' response (bottom) respectively. **B.** Evolution of average reward (left), sensory expectation reflected in the output of model SS (middle), and the reward expectation reflected in the output of model VS (right) across trials before and after manipulation of stimulus (top row) and reward (bottom row) statistics. **C.** Stimulus-evoked model dopamine transients in signal (solid lines) and noise (dashed lines) trials. Rows correspond to different conditions and columns correspond to different striatal sub-regions. **D.** Average within-trial dynamics of choice (left), belief (middle), and value (right) under different conditions (rows). Dashed line corresponds to noise trials while lines with gray hues correspond to signal trials, with darker hues denoting a higher signal-to-noise ratio. **E.** Prediction error (PE) was optimized by expressing it as a sum of multiple spatiotemporal filters. Each spatial filter corresponds to a weighted sum of the activity of units in the cortical and striatal regions and each temporal filter corresponds to a weighted sum of the output of the spatial filter across a moving window. **F.** Left panel: PE optimized for learning sensory expectation – magnitude of alignment of the optimized spatial weights with activity dimension encoding reward (r), value (v), belief (b), and action (a), and the corresponding temporal weights. Right two panels: Similar to the left panel, but showing weights optimized for learning reward expectation.

Above results suggest that sPE constructed from momentary changes in internal belief can support biologically plausible learning of sensory expectations, alongside learning of reward expectations from temporal difference rPE. We aimed to determine whether learning sensory expectations specifically requires the type of sPE described above, or if it could instead be guided by signals that do not rely on internal beliefs. To address this, we expressed the prediction error as an arbitrary spatiotemporal transformation of cortical and striatal activity and then used a technique called meta-learning (or learning-to-learn) to optimize the spatial and temporal kernels i.e., weights (**Figure 2E**; Methods). When we optimized the prediction error for learning sensory expectation, the resulting spatial weights were aligned primarily with the dimension of neural activity encoding the belief that the stimulus contained a signal and the temporal weights resembled a derivative operator (**Figure 2F** – left). Alignment with dimensions that encoded reward, value (i.e., expected reward), and action was much weaker. This suggests that sPE constructed using momentary changes in internal belief optimizes learning of sensory expectations. In contrast, a similar approach for optimizing reward expectation learning produced two spatiotemporal filters that together resembled the temporal difference rPE signal: one filter acted like a smoothing operator on reward signals, while the other filter extracted momentary changes in value (**Figure 2F** – middle and right).

3.2 Signatures of sensory prediction errors in sensory striatal dopamine

We asked whether model VS and SS dopamine signals can inform the interpretation of dopamine signals recorded in those areas across different tasks. We considered two studies in which dopamine recordings were performed both in the ventral and sensory striatum (specifically, the tail of striatum or TS) in mice allowing for a direct comparison of dopamine dynamics between the model and data. In the first of these studies, (Menegas et al., 2017) tracked dopamine dynamics across the VS and TS during associative learning. They tested whether repeated training affected the observed pattern in VS and TS dopamine

by introducing a new odor paired with a reward every day, and then measured dopamine activity while learning new odor-reward association (**Figure 3A** – top; Methods). They found that TS dopamine responded to novel odors, and this decreased across repeated exposures. In contrast, VS dopamine initially responded to rewards and only later on to odors in a manner that reflected the learned association between odors and rewards. Simulating the model under associative learning yielded a strikingly similar pattern where odor-induced VS and TS dopamine dynamics increased and decreased respectively across trials (**Figure 3A** – bottom). The computational role of dopamine in the model offers an interpretation for these experimental results. Because model VS dopamine encodes rPE, the trajectory of dopamine signals reflects the solution to the temporal credit assignment from rewards to odors. In contrast, TS dopamine encodes sPE and reflects the animal's surprise upon encountering an unexpected stimulus, which gradually reduces with repeated presentations as the animal comes to expect this stimulus.

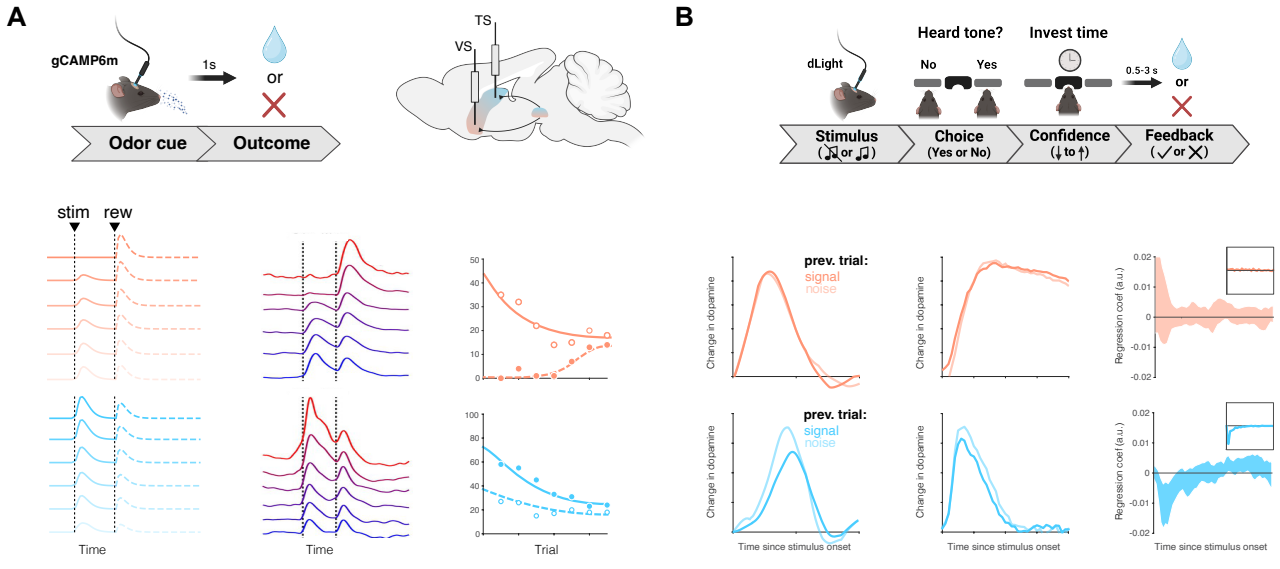

Figure 3: **Signatures of distinct prediction errors in associative learning and signal detection**. **A.** Schematic of the associative learning task. Mice were exposed to repeated pairings of a novel odor and reward delivery. **Left**: The evolution of model-predicted prediction error traces in the ventral (top) and dorsal (bottom) striatum across trials. **Middle**: The evolution of dopamine signals in the mouse VS and TS. **Right**: The amplitude of dopamine transients evoke by odor (solid circles) and reward (open circles) in VS and TS of mice. Model-predicted prediction errors in response to the odor (dashed line) and reward (solid line) in the ventral and dorsal striatum. **B.** Schematic of the signal detection task. Following stimulus presentation, mice reported their choice and confidence, estimated as the amount of time invested in the choice, and received feedback in the form of a reward. **Left**: Model-predicted stimulus-evoked prediction errors in the ventral (top) and dorsal (bottom) striatum, conditioned on the previous trial's stimulus. **Middle**: Stimulus-evoked dopamine transients in the mouse VS and TS. Bottom: Influence of the previous trial's stimulus on the dopamine dynamics, estimated by a multiple linear regression model. Inset shows model predictions. Error bars denote 95% confidence intervals.

In a different study, (Schmack et al., 2021) recorded dopamine dynamics in the same areas but during auditory signal detection (**Figure 3B** – top). In this task, mice initiated trials with a center nosepoke, followed by a nosepoke at a left or right port depending on whether it heard a tone. To look for signatures of rPE and sPE in the data, we first examined dopamine signals in the model trained to perform signal

detection. We found that the amplitude of signal-evoked model SS dopamine, but not VS dopamine, was modulated by the previous trial's stimulus identity. Specifically, encountering a signal trial increases the signal expectation, decreasing the magnitude of signal-evoked sPE encoded by SS dopamine in the subsequent trial. We found this precise pattern of modulation by stimulus history in the mouse TS. In contrast, mouse VS dopamine was not significantly modulated by stimulus history (**Figure 3B** – bottom). A regression analysis across the set of all trials confirmed that the previous trial's state (signal/noise) negatively modulated dopamine dynamics in TS but not VS across all mice (Mixed-effects linear regression: VS t= 0.6447; p = 0.5191 and TS t = -2.5896; p = 0.009615).

The above results show that dopamine dynamics in the mouse sensory but not ventral striatum resembles model sensory prediction errors. We demonstrated earlier that this prediction error signal is optimized for learning sensory expectations via plasticity mechanisms. If this is the case, then perturbing sensory striatal dopamine should have a qualitatively similar effect on behavior as that on the model performance. This leads to two specific predictions. First, the effects should be consistent with a bias in signal expectation rather than reward expectation. Second, the effect of perturbation, if any, should gradually build up over time. We tested both predictions by re-analyzing the data from optogenetic stimulation experiments carried out during the same paradigm. Briefly, dopamine terminals in the tail of striatum were chronically stimulated across blocks of fifty trials, interleaved with baseline blocks in which the laser was turned off (**Figure 4A**; Methods).

We found that, on average, stimulation significantly increased the rate of false alarms as reported previously (Schmack et al., 2021) (**Figure 4B**) as well as an increase in the hit rate (**Figure 4C**) (Mixed-effects logistic regression of choice: t = 3.6987; p = 0.0002175). Moreover, stimulation also increased the confidence that a signal was present in all trials (**Figure 4D**; Mixed-effects linear regression of signal confidence: t = 2.7602; p = 0.0057882), consistent with the hypothesis that dopamine stimulation in this circuit influences behavior by increasing signal rather than reward expectation. To test whether these effects were due to a learning rather than driving role of dopamine, we estimated how the above response measures evolved across blocks. We found that all three measures – false alarms, hits, signal confidence – exhibited a significant increase from baseline as a function of blocks within a stimulation session (**Figure 4B-E**; Mixed-effects logistic regressions predicting false alarms across session, t = 2.9776; p = 0.002930 and predicting hits t = 2.1078; p = 0.035133. Mixed-effects linear regressions predicting signal confidence across session t = 3.5781; p = 0.0003489). These effects were recapitulated by model simulations in which we scaled up the level of dopamine specifically in the dorsal (blue) but not the ventral striatum (orange). Taken together, these analyses suggest that dopamine-mediated plasticity mechanisms in the dorsal striatum play a role in learning signal expectation.

3.3 Signal detection engages a two-stage learning process in humans

In addition to supporting a plasticity-based role for dopamine in perception, the results from the optogenetic stimulation study validates the unique behavioral signatures of excess dopamine in the dorsal sensory regions. Notably however, this requires going beyond just measuring changes in false alarms

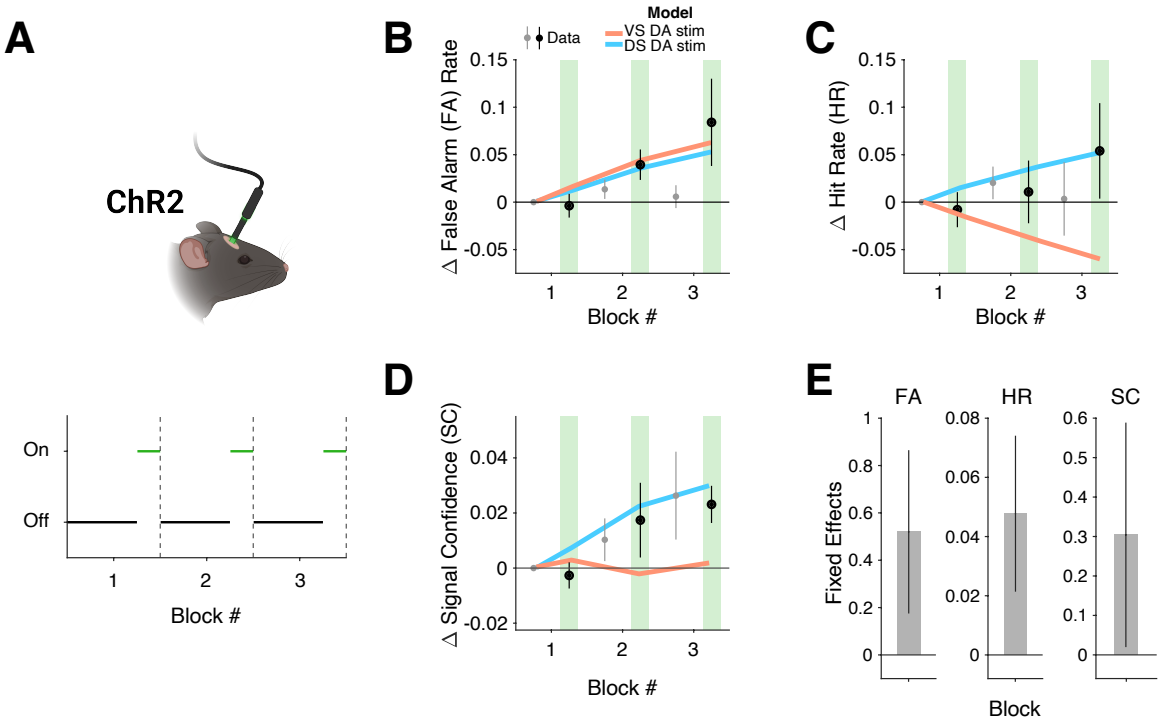

Figure 4: **Evolution of signal expectation from dopamine stimulation**. **A.** Schematic of the optogenetic stimulation protocol. Laser was ON or OFF throughout the respective blocks. **B.** Change in the rate of false alarms with respect to the baseline (i.e., the first block of the experimental session) in each block. **C.** Similar to B, but showing the change in the rate of hits. **D.** Similar to B, but showing the change in signal confidence. **E.** Fixed effects of the block number on the false alarm rate (FA), hit rate (HR), and signal confidence (SC).

commonly reported in previous work on patients with hallucinations.

Next, we asked whether the modular architecture in which stimulus and reward statistics are learned in distinct pathways can account for human behavior in perceptual decision-making tasks. Previous studies have typically investigated behavioral adaptation to either sensory or reward statistics but not both. To detect distinct signatures of reward and perceptual learning, we developed a novel signal detection task (**Figure 5A** – left). On each trial, participants experienced an ambiguous auditory stimulus (i.e., 1 KHz tone of varying signal to noise ratio) and reported whether they heard a tone as well as the confidence associated with their judgement. Participants then received feedback about whether a tone was present and whether their choice resulted in additional reward. To incentivize learning, we manipulated both whether signals are more likely to be present or absent (i.e., sensory prior) and whether a 'yes' choice is likely to result in more or less reward (i.e., action value) (**Figure 5A** – right).

The model predicts that the two types of manipulations will drive distinct patterns of behaviors in this task. First, since both perceptual and reward learning pathways influence choice via feedback to the cortex, the model predicts that choices would shift in accordance to the previous trial's stimulus (**Figure 5B** – left) and reward (**Figure 5C** – left). Second, since only the perceptual learning pathway (SS) influences belief estimates by conveying the prior expectation, the model predicts that confidence reports

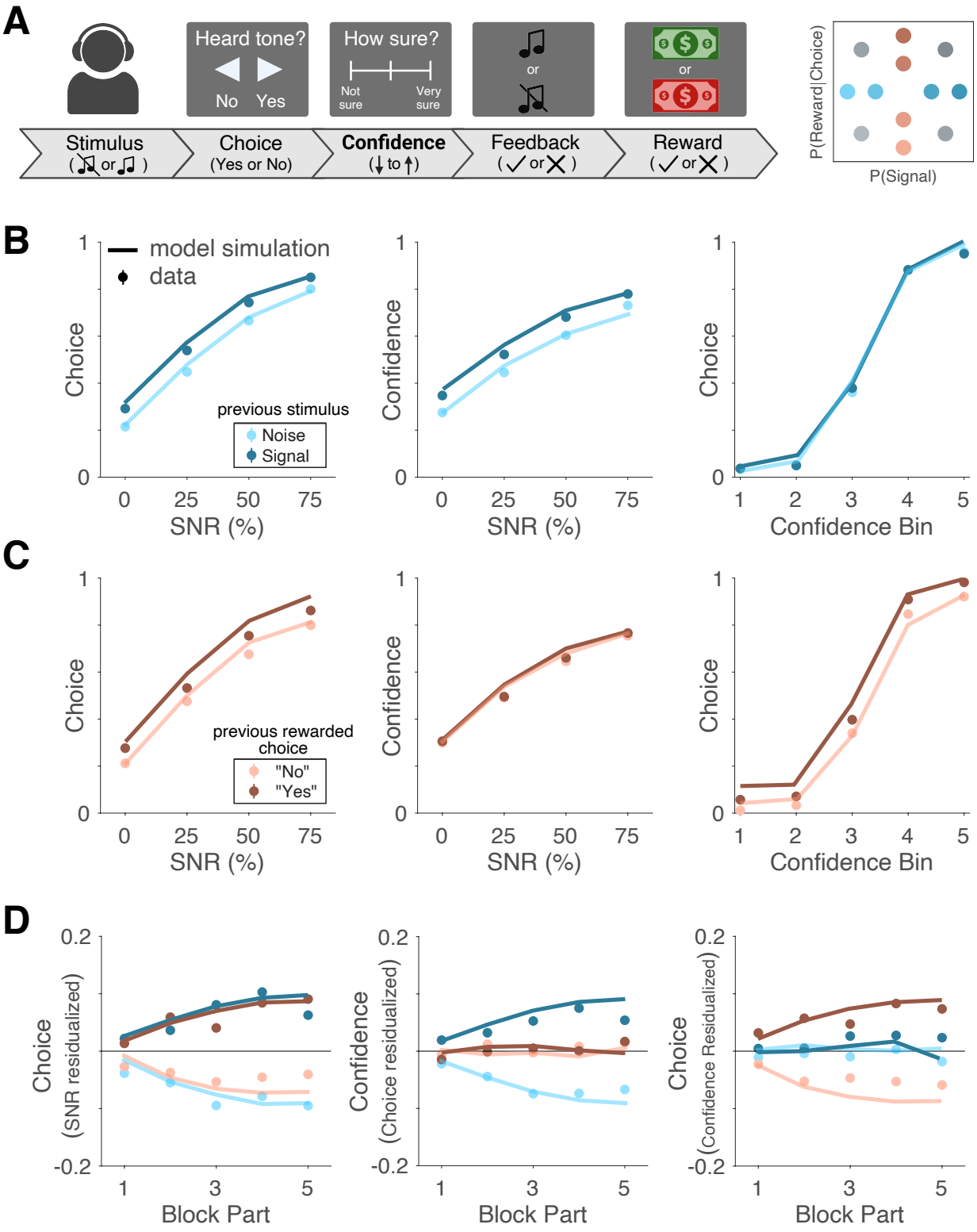

Figure 5: (Caption next page.)

Figure 5: **Distinct behavioral signatures of perceptual and reward learning**. **A.** Schematic of the task. **Left**: Human participants made a binary judgement about whether they heard a tone, reported the associated confidence on a continuous scale, and received feedback. **Right**: The fraction of signal trials and the fraction of rewarded yes responses were manipulated across blocks. **B.** Effect of signal manipulation. **Left**: The fraction of yes responses as a function of the signal to noise ratio (SNR). **Middle**: Confidence as a function of SNR. **Right**: Participants' policy, quantified as the fraction of yes responses as a function of confidence. **C.** Similar to B, but showing the effect of reward manipulation. **D.** The evolution of choice, confidence, and policy across trials of the block for different manipulations.

should be shifted in alignment with the previous stimulus but not reward feedback (**Figure 5B,C** – middle). Third, only the reward learning pathway (VS) influences the policy, defined as the action taken in a given belief state and estimated as the fraction of yes choices binned by signal confidence for each participant. The model predicts that policy should shift in congruence with the previous reward but not stimulus (**Figure 5B,C** – right). The data is consistent with all of these predictions, suggesting that human participants use distinct mechanisms to adapt to changes in sensory and reward statistics.

In the model, adaptation to statistics is achieved by dopamine-mediated plasticity mechanisms. To further confirm that the human participants engaged mechanisms that relied on a gradual learning process, we examined the evolution of all three variables – choice, confidence, and policy – across trials. We found that all quantities shifted gradually, plateauing towards the end of the block, suggesting the adaptation is an outcome of a slow learning process (**Figure 5D**).

3.4 Hallucination propensity is linked to increased sensory prediction errors

We now turn to the key question that motivated this study. What is the computational mechanism underlying the robust empirical link between hallucination severity and elevated striatal dopamine release? To address this, we combine insights gained from our analysis of the physiology of dopamine signaling in mouse striatum and history-dependent choices seen in human behavior. Recall that stimulating dopamine terminals carrying signals that bear signatures of a sensory prediction error in mice led to a slow build-up of perceptual confidence over the course of the experimental session. In humans, both stimulus and reward history biased future choices to a comparable extent but only the former also biased the perceptual confidence underlying those choices. Based on these findings, we reasoned that elevated dopamine might cause hallucinations primarily through its role in encoding sensory prediction errors. If this is the case, then participants with increased hallucination proneness (i.e., CAPS score) would also exhibit greater stimulus history effects.

To test this, we re-analyzed the data from human participants and asked whether the behavioral effects of stimulus history interacted with the participants' CAPS score. We found that participants with a high CAPS score indeed exhibited a significantly greater tendency to report hearing a signal (including on noise trials i.e., false alarms) following a signal trial, relative to those with a low CAPS score (**Figure 6A,C** – left; Mixed-effects logistic regression: $t = 2.1751$; $p = 0.029629$). Critically, this differential effect of stimulus history on the choices made by participants from the two groups was mirrored by a similar

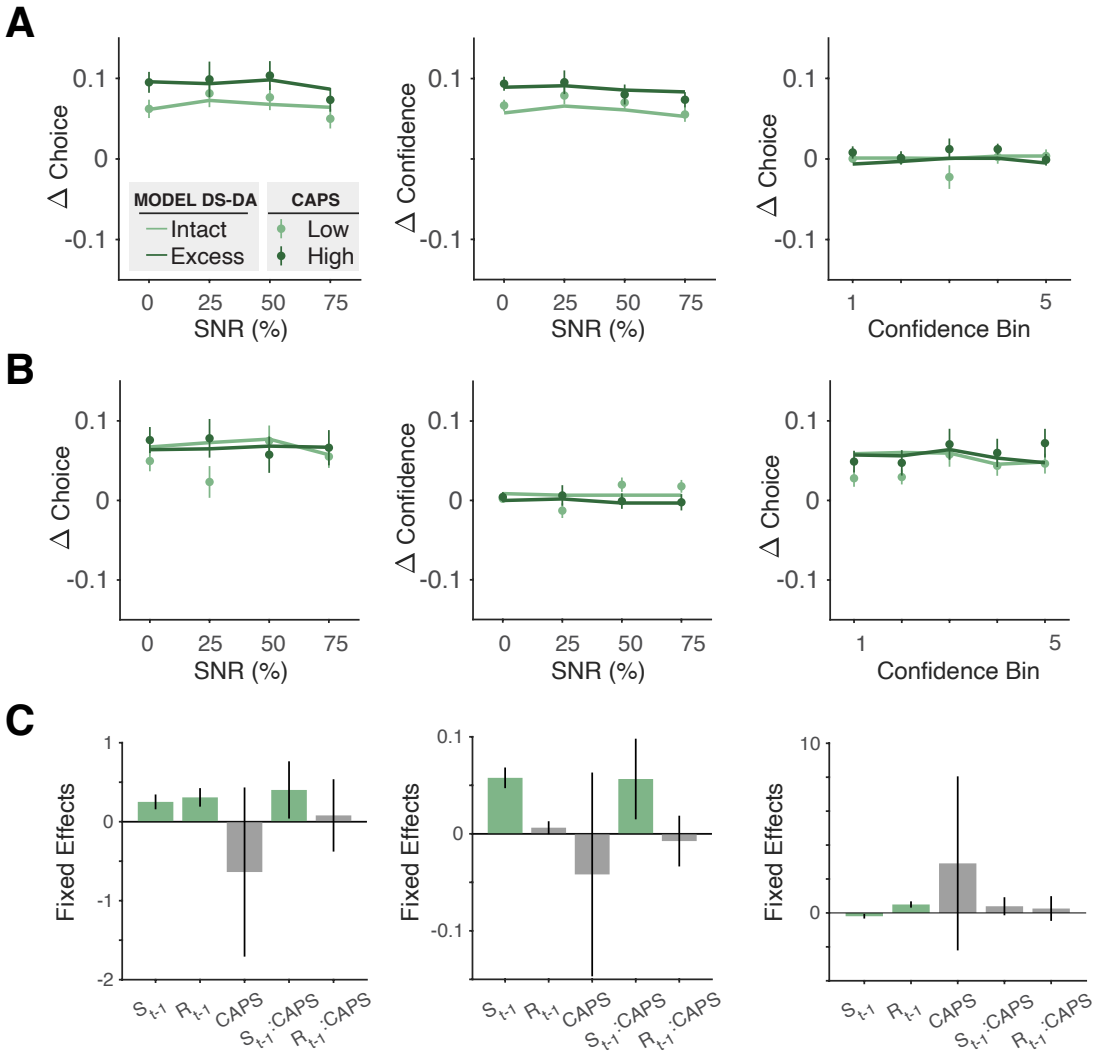

Figure 6: **Hallucination proneness exaggerates stimulus history effects**. **A. Left**: The change in the rate at which participants report hearing a signal as a function of signal to noise ratio, after a signal trial, relative to after a noise trial. **Middle**: Similar to the left panel, but showing the change in the reported perceptual confidence. **Right**: Similar to the left panel, but showing the change in policy i.e., rate at which participants report hearing a signal as a function of confidence. Participants are groups based on their CAPS score was low (light green) or high (dark green). Solid lines show model simulations in which dorsal striatal dopamine was either intact (line green) or scaled up (dark green; Methods). **B.** Similar to A, but showing the change in response measures due to the previous trial's reward feedback. **C.** Coefficients of a fixed effects model, showing the effect of previous trial's stimulus (s_{t-1}), reward (r_{t-1}), hallucination proneness (CAPS), and the interactions ($s_{t-1} \times$ CAPS, $r_{t-1} \times$ CAPS) on choice (left), confidence (middle) and policy (right).

effect on perceptual confidence which was significantly greater when making false alarms (**Figure 6A,C** – middle; Mixed-effects linear regression: t = 2.6652; p = 0.007695), but we did not observe a similar effect on the behavioral policy (**Figure 6A,C** – right; Mixed-effects logistic regression: t = 1.4485; p = 0.1475). These results are readily explained by adding an offset to the dopamine signals specifically in the

model SS (Figure 6A – solid lines), suggesting that hallucination proneness arises from elevated sensory prediction errors in the SS.

In contrast, reward history did not reveal a consistent difference between groups in the pattern of choices (**Figure 6B** – left) nor in the pattern of perceptual confidence (**Figure 6B** – middle), and this was recapitulated by simulations in which dopamine in the model VS faithfully conveyed reward prediction errors in both groups (**Figure 6B** – solid lines). These results suggest that hallucination-like percepts are consistent with the effects of an elevated SPE dopamine signaling in the SS.

4 Discussion

We developed an anatomically constrained, normative circuit model of corticostriatal loops in which dopaminergic prediction error signals in the ventral and sensory striatum are optimized to support value learning and perceptual learning respectively. Dopamine dynamics in the mouse tail of the striatum mirrored sensory prediction errors essential for perceptual learning, and dopamine stimulation in this region gradually adjusted perceptual estimates indicative of a plasticity-based mechanism. In an auditory detection task, human behavior showed both value and perceptual learning, but only perceptual learning correlated with hallucination proneness. These findings reveal a concrete computational mechanism underpinning the association between excess striatal dopamine and auditory hallucinations.

Although the complex polygenic etiology of psychosis points to a dysfunction in a broad range of synaptic mechanisms involving dopamine, NMDA, GABA etc., almost all effective antipsychotic medications exclusively target dopamine receptors and dopamine stimulants induce psychotic symptoms in healthy individuals (Curran et al., 2004; Sommer et al., 2012; Lally & MacCabe, 2015). Unsurprisingly, some of the earliest theories of psychosis focused on dopamine (Seeman & Kapur, 2000; Kapur, 2003). Reflecting the dominant computational views at the time, these theories appealed to the role of dopamine in reward learning and motivational salience. However, the problems with these early theories were twofold. First, they did not directly model how excess dopamine leads to psychotic symptoms like hallucinations, instead relying on paradigms like associative and motor learning for empirical validation (Walter et al., 2009; Gradin et al., 2011). Second, they were not neurobiologically grounded in a circuit model and therefore carried limited clinical relevance. Consequently, the 'dopamine hypothesis' has been supplanted by theoretical frameworks like predictive coding that focus on pathological inference, instantiated in circuit models as an imbalance between excitation and inhibition (Kehrer, 2008; Keller & Sterzer, 2024). However, an emphasis on inferential mechanisms does not preclude a role for dopamine. In fact, a growing body of basic research using modern molecular tools points to anatomically segregated and physiologically distinct dopamine signaling pathways, including a role for the sensory striatum and dopamine therein in perception (Wang et al., 2018; Guo et al., 2018; Schmack et al., 2021; A. Chen et al., 2022). This development has paralleled proposals that extend the canonical, reward-centric computational theories by rethinking dopamine signaling as a generalized prediction error (Gardner et al., 2018). Informed by these empirical and theoretical results, we developed a corticostriatal model in which dopamine contributes to the inferential process by enabling learning of sensory expectations in the sensory striatum.

4.1 Novelty and implications

The modeling approach represents an integration of normative and mechanistic considerations and is particularly appealing from the perspective of preclinical translational research since it allows for testing both granular predictions about neural dynamics and coarse-level behavioral predictions. The proposed link between dopamine and perceptual learning has major implications for understanding the neural basis of auditory hallucinations. First, we identified signatures of sensory prediction errors in the sensory striatum but not in the ventral striatum, a result that paralleled the significant interaction between hallucination proneness and stimulus-history but not reward history in humans. This reconciles findings from PET imaging studies that localize elevated extracellular dopamine and synthesis capacity to the human dorsal striatum (Howes et al., 2009; Kegeles et al., 2010; Mizrahi et al., 2012) rather than limbic striatum as anticipated by earlier theories of dopamine and psychosis. Second, a specialized role for dopamine in this sub-region provides a potential explanation for why most anti-psychotic drugs, which are broadly antidopaminergic, have wide-ranging cognitive and motor side-effects (Huhn et al., 2019). Lastly, our analysis of the time course of evolution of perceptual bias in response to dopamine stimulation was consistent with a plasticity-based mechanism. A role for sensory striatal dopamine in perceptual learning rather than driving decisions potentially reconciles a large body of clinical work demonstrating slow recovery timescales following antipsychotic medication. Additionally, this finding suggests that pharmacological agents inhibiting corticostriatal plasticity could be explored as a potential treatment option and may underlie the efficacy of mascarinic agonists in treating psychosis (Mirza et al., 2003; Kaul et al., 2024).

An advantage of sPE-gated corticostriatal plasticity is that it enables post-synaptic neurons i.e., neurons in the sensory striatum to develop action-independent sensory representations, allowing for flexible reuse of these signals in different contexts. Action-independent coding has indeed been reported recently in the sensory striatum of rodents (Guo et al., 2018) and is consistent with earlier studies indicating that optogenetic stimulation of neurons in this region biases perception rather than choice (Wang et al., 2018). While the model postulates sPE-based learning in sensory neurons in the striatum in general, we validated it using data from the rodent tail of striatum. Previous studies demonstrated that dopamine signal in this region is consistent with other types of prediction errors such as action prediction errors (Greenstreet et al., 2022) and threat prediction errors (Menegas et al., 2018; Akiti et al., 2022). We note that these proposals are not mutually exclusive as they pertain to different task dimensions – sensory input, motor command, and outcome valence – and dopamine dynamics could multiplex these signals to enable learning along different dimensions. While the precise computational consequences of overestimating threat prediction errors have not yet been investigated, we speculate that this could lead to negative symptoms of schizophrenia such as avoidance, and similarly, overestimating action prediction errors might accelerate the acquisition of stimulus-response associations leading to compulsivity (Lakshminarasimhan, 2024). However, neither threat nor action prediction errors provide a plausible link between excess striatal dopamine and psychosis, especially perceptual abnormalities like hallucinations. In contrast, our finding that hallucination proneness significantly heightens sensitivity to stimulus-history in humans supports the interpretation that hallucinations arise from elevated sensory prediction errors. The proposed model

of the sensory striatum also differs from alternative models employing different types of outcome-specific prediction errors (Gardner et al., 2018; Lee, Sagiv, Engelhard, Witten, & Daw, 2024) in that it allows for learning exclusively from internal beliefs i.e., in the absence of actions and/or outcomes. This is in line with empirical work showing that humans can learn stimulus statistics in the absence of feedback (Zylberberg et al., 2018; Loewenstein et al., 2021).

4.2 Predictions and extensions

Our analyses of striatal dopamine revealed signatures of sensory prediction errors in the tail of the rodent striatum, but the model makes concrete predictions linking dopamine to the learning of sensory expectations, that could be validated using signal detection paradigms with variable signal probability. First, stimulus-induced dopamine levels should decrease as signal probability increases, since higher probability makes signals less surprising. Second, this decrease in dopamine should be accompanied by an increase in cortical activity within the dimension that encodes belief, driven by input from the basal ganglia prior to stimulus onset. Third, discrepancies between signal probability decoded from the initial cortical state and the true signal probability should be correlated with deviations of the sensory striatal dopamine from the optimal sensory prediction error signal.

One challenge in testing the model predictions is the need to precisely target recordings to the corticostriatal circuit involved in learning and inference of the variable of interest. While the loop involving the primary auditory cortex is an appropriate choice in simple auditory signal detection tasks, future extensions of the model should expand this approach to learn complex tasks using architectures that incorporate hierarchical processing of rich sensory inputs, informed by detailed wiring diagrams of the cortico-basal-ganglia loop (Foster et al., 2021). This will prove crucial to testing whether the pathophysiology of complex hallucinations in patients with psychosis is limited to dopamine signals that encode prediction errors about specific sensory features or if extends more broadly. A related assumption was our treatment of dopamine as a scalar prediction error signal. While this is justified and supported by empirical work in the reward domain, sensory inputs inherently comprise multiple feature dimensions, each of which requires a separate learning channel. More theoretical work is needed to explore the scalability of the proposed model to settings with naturalistic inputs using vector-valued sensory prediction errors similar to those used in reinforcement learning (Lee et al., 2024). Our emphasis of dopamine as a prediction error signal for learning does not preclude complementary roles that stem from sources of heterogeneity not considered here. (Collins & Frank, 2014). Previous studies in the reward domain have proposed separate roles for both D1- and D2-expressing striatal neurons (Collins & Frank, 2014) and for tonic and phasic components of dopamine (Niv et al., 2007) in learning and driving behavior. This spatiotemporal heterogeneity has recently been shown to have interesting consequences for learning due to differential dopamine affinities of D1 and D2 receptors (Pinto & Uchida, 2023). Incorporating such features into the proposed account of perceptual learning will be crucial for understanding the specific subcellular mechanisms by which excess striatal dopamine leads to hallucinations and for developing safer treatments.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*. doi: 10.3389/fpsyt.2013.00047

Akiti, K., Tsutsui-Kimura, I., Xie, Y., Mathis, A., Markowitz, J. E., Anyoha, R., ... Watabe-Uchida, M. (2022). Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron*, *110*. doi: 10.1016/j.neuron.2022.08.022

Cassidy, C. M., Balsam, P. D., Weinstein, J. J., Rosengard, R. J., Slifstein, M., Daw, N. D., ... Horga, G. (2018). A perceptual inference mechanism for hallucinations linked to striatal dopamine. *Current Biology*. doi: 10.1016/j.cub.2017.12.059

Cassidy, C. M., Zucca, F. A., Girgis, R. R., Baker, S. C., Weinstein, J. J., Sharp, M. E., ... Horga, G. (2019). Neuromelanin-sensitive mri as a noninvasive proxy measure of dopamine function in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *116*. doi: 10.1073/pnas.1807983116

Chen, A., Malgady, J., Chen, L., Shi, K., Cheng, E., Plotkin, J., ... Xiong, Q. (2022). Nigrostriatal dopamine pathway regulates auditory discrimination behavior. *Nature Communications*, *13*.

Chen, A. P., Chen, L., Kim, T. A., & Xiong, Q. (2021, 6). Integrating the roles of midbrain dopamine circuits in behavior and neuropsychiatric disease. *Biomedicines*, *9*, 647. doi: 10.3390/biomedicines9060647

Collins, A. G., & Frank, M. J. (2014). Opponent actor learning (opal): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*. doi: 10.1037/a0037015

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). *Hallucinations and strong priors* (Vol. 23). doi: 10.1016/j.tics.2018.12.001

Curran, C., Byrappa, N., & McBride, A. (2004). Stimulant psychosis: Systematic review. *British Journal of Psychiatry*, *185*. doi: 10.1192/bjp.185.3.196

Foster, N. N., Barry, J., Korobkova, L., Garcia, L., Gao, L., Becerra, M., ... Dong, H. W. (2021). The mouse cortico–basal ganglia–thalamic network. *Nature*, *598*. doi: 10.1038/s41586-021-03993-3

Friston, K. J. (2005). Hallucinations and perceptual inference. *Behavioral and Brain Sciences*, *28*. doi: 10.1017/S0140525X05290131

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, *285*, 20181645. doi: 10.1098/rspb.2018.1645

Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., ... Steele, J. D. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain*, *134*. doi: 10.1093/brain/awr059

Greenstreet, F., Vergara, H. M., Pati, S., Schwarz, L., Wisdom, M., Marbach, F., ... Stephenson-Jones, M. (2022, 1). Action prediction error: a value-free dopaminergic teaching signal that drives stable learning. *bioRxiv*, 2022.09.12.507572. doi: 10.1101/2022.09.12.507572

Guo, L., Walker, W. I., Ponvert, N. D., Penix, P. L., & Jaramillo, S. (2018). Stable representation of sounds in the posterior striatum during flexible auditory decisions. *Nature Communications*, *9*. doi:

10.1038/s41467-018-03994-3

Hennig, J. A., Pinto, S. A., Yamaguchi, T., Linderman, S. W., Uchida, N., & Gershman, S. J. (2023). Emergence of belief-like representations through reinforcement learning. *PLoS Computational Biology*, *19*. doi: 10.1371/journal.pcbi.1011067

Howes, O. D., Montgomery, A. J., Asselin, M. C., Murray, R. M., Valli, I., Tabraham, P., ... Grasby, P. M. (2009). Elevated striatal dopamine function linked to prodromal signs of schizophenia. *Archives of General Psychiatry*, *66*. doi: 10.1001/archgenpsychiatry.2008.514

Huhn, M., Nikolakopoulou, A., Schneider-Thoma, J., Krause, M., Samara, M., Peter, N., ... Leucht, S. (2019). Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. *The Lancet*, *394*. doi: 10.1016/S0140-6736(19)31135-3

Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry*, *160*. doi: 10.1176/appi.ajp.160.1.13

Kaul, I., Sawchak, S., Correll, C. U., Kakar, R., Breier, A., Zhu, H., ... Brannan, S. K. (2024). Efficacy and safety of the muscarinic receptor agonist karxt (xanomeline–trospium) in schizophrenia (emergent-2) in the usa: results from a randomised, double-blind, placebo-controlled, flexible-dose phase 3 trial. *The Lancet*, *403*. doi: 10.1016/S0140-6736(23)02190-6

Kegeles, L. S., Abi-Dargham, A., Frankle, W. G., Gil, R., Cooper, T. B., Slifstein, M., ... Laruelle, M. (2010). Increased synaptic dopamine function in associative regions of the striatum in schizophrenia. *Archives of General Psychiatry*, *67*. doi: 10.1001/archgenpsychiatry.2010.10

Kehrer, C. (2008). Altered excitatory-inhibitory balance in the nmda-hypofunction model of schizophrenia. *Frontiers in Molecular Neuroscience*, *1*. doi: 10.3389/neuro.02.006.2008

Keller, G. B., & Sterzer, P. (2024). Predictive processing: A circuit approach to psychosis. *Annual Review of Neuroscience*, *47*. doi: 10.1146/annurev-neuro-100223-121214

Kesby, J. P., Eyles, D. W., McGrath, J. J., & Scott, J. G. (2018). Dopamine, psychosis and schizophrenia: The widening gap between basic and clinical neuroscience. *Translational Psychiatry*. doi: 10.1038/s41398-017-0071-9

Lakshminarasimhan, K. (2024). A computational principle of habit formation. *bioRxiv*. doi: 10.1101/2024.10.12.618033

Lally, J., & MacCabe, J. H. (2015). Antipsychotic medication in schizophrenia: A review. *British Medical Bulletin*, *114*. doi: 10.1093/bmb/ldv017

Laruelle, M., Abi-Dargham, A., Dyck, C. H. V., Gil, R., D'Souza, C. D., Erdos, J., ... Innis, R. B. (1996). Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. *Proceedings of the National Academy of Sciences of the United States of America*, *93*. doi: 10.1073/pnas.93.17.9235

Lee, R. S., Sagiv, Y., Engelhard, B., Witten, I. B., & Daw, N. D. (2024). A feature-specific prediction error model explains dopaminergic heterogeneity. *Nature Neuroscience*, *27*, 1574-1586. doi: 10.1038/s41593-024-01689-1

Loewenstein, Y., Raviv, O., & Ahissar, M. (2021). Dissecting the roles of supervised and unsupervised learning in perceptual discrimination judgments. *Journal of Neuroscience*, *41*. doi:

10.1523/JNEUROSCI.0757-20.2020

McCutcheon, R. A., Abi-Dargham, A., & Howes, O. D. (2019, 3). Schizophrenia, dopamine and the striatum: From biology to symptoms. *Trends in Neurosciences*, *42*, 205-220. doi: 10.1016/j.tins .2018.12.004

Menegas, W., Akiti, K., Amo, R., Uchida, N., & Watabe-Uchida, M. (2018, 10). Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nature Neuroscience*, *21*, 1421-1430. doi: 10.1038/s41593-018-0222-1

Menegas, W., Babayan, B. M., Uchida, N., & Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife*, *6*, e21886. doi: 10.7554/eLife.21886.001

Mirza, N. R., Peters, D., & Sparks, R. G. (2003). Xanomeline and the antipsychotic potential of muscarinic receptor subtype selective agonists. *CNS Drug Reviews*, *9*. doi: 10.1111/j.1527-3458.2003.tb00247.x

Mizrahi, R., Addington, J., Rusjan, P. M., Suridjan, I., Ng, A., Boileau, I., ... Wilson, A. A. (2012). Increased stress-induced dopamine release in psychosis. *Biological Psychiatry*, *71*. doi: 10.1016/ j.biopsych.2011.10.009

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, *191*. doi: 10.1007/s00213-006-0502-4

Pinto, S. R., & Uchida, N. (2023, 1). Tonic dopamine and biases in value learning linked through a biologically inspired reinforcement learning model. *bioRxiv*, 2023.11.10.566580. doi: 10.1101/ 2023.11.10.566580

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*. doi: 10.1126/science.aan3458

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*. doi: 10.1126/science.abf4740

Seeman, P., & Kapur, S. (2000). Schizophrenia: More dopamine, more d2 receptors. *Proceedings of the National Academy of Sciences of the United States of America*, *97*. doi: 10.1073/pnas.97.14.7673

Sohn, H., & Jazayeri, M. (2021). Validating model-based bayesian integration using prior–cost metamers. *Proceedings of the National Academy of Sciences of the United States of America*, *118*. doi: 10.1073/ pnas.2021531118

Sommer, I. E., Slotema, C. W., Daskalakis, Z. J., Derks, E. M., Blom, J. D., & Gaag, M. V. D. (2012). The treatment of hallucinations in schizophrenia spectrum disorders. *Schizophrenia Bulletin*, *38*. doi: 10.1093/schbul/sbs034

Walter, H., Kammerer, H., Frasch, K., Spitzer, M., & Abler, B. (2009). Altered reward functions in patients on atypical antipsychotic medication in line with the revised dopamine hypothesis of schizophrenia. *Psychopharmacology*, *206*. doi: 10.1007/s00213-009-1586-4

Wang, L., Rangarajan, K. V., Gerfen, C. R., & Krauzlis, R. J. (2018, 3). Activation of striatal neurons causes a perceptual decision bias during visual change detection in mice. *Neuron*, *97*, 1369-1381.e5. doi: 10.1016/j.neuron.2018.01.049

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. *Neuron*, *99*, 1083-1097.e6. doi: https://doi.org/10.1016/ j.neuron.2018.07.035