Translational Psychiatry

https://doi.org/10.1038/s41398-025-03773-x

Article in Press

Refining mechanistic models of hallucinations for enhanced translatability

Received: 26 February 2025

Revised: 3 November 2025

Accepted: 14 November 2025

Cite this article as: Buck, J., Iigaya, K., Horga, G. Refining mechanistic models of hallucinations for enhanced translatability. *Transl Psychiatry* (2025). https://doi.org/10.1038/s41398-025-03773-x

Justin Buck, Kiyohito Iigaya & Guillermo Horga

Refining Mechanistic Models of Hallucinations for Enhanced Translatability

Justin Buck M.A.[1,2*], Kiyohito Iigaya Ph.D.[1,3], Guillermo Horga M.D. Ph.D.[1,3]

Department of Psychiatry[1] and Department of Neuroscience[2], Columbia University
Department of Psychiatry, New York State Psychiatric Institute[3]

*Correspondence: justin.buck@columbia.edu & guillermo.horga@nyspi.columbia.edu
1051 Riverside Drive, Suite 6100
New York, NY 10032
301-525-9310

Running Title: Refining Mechanistic Models of Hallucinations

Summary

Over the past two decades, foundational work has provided key insights into the cognitive and neural basis of hallucinations. This progress has led to the development of several families of theories, each of which propose that hallucinations arise from distinct cognitive mechanisms. Since these cognitive mechanisms likely map onto separate circuit-level implementations, arbitrating between them is critical to advance our understanding of hallucination pathophysiology and guide the development of novel targeted therapeutics. However, several obstacles have hindered this progress, including the under-specification of theories and inadequate comparative testing. To overcome these challenges, and following best practices in cognitive computational neuroscience, theories should **1)** articulate computational and biological details at a level that allows the generation of precise, testable predictions, and **2)** be evaluated using experiments designed to emphasize their unique signatures to facilitate falsification. To illustrate this general approach, we demonstrate how theory-driven computational models constrained by well-replicated findings across basic, preclinical, and clinical neuroscience can provide a principled means to prioritize falsifiable mechanistic theories of auditory hallucinations. We then discuss how these models can be used to inform the development of richer behavioral paradigms and analytical approaches that enable direct comparisons between competing theories. Overall, we propose a general strategy for the specification and falsification of candidate mechanisms underlying (auditory) hallucinations – a critical prerequisite for refining hallucination theories and advancing their translational potential.

Introduction

Mechanistic theories are an invaluable tool for clinical translation. By specifying how core features of hallucinations arise, such models can guide the identification of neurobiological treatment targets. Several promising mechanistic theories of hallucinations are currently under investigation in the field[1–13]. Arbitrating between these theories is critically important for translation because they likely map onto distinct neurobiological substrates, each of which could represent a distinct treatment target[14]. Progress in this direction requires theories that are described with enough detail to be rigorously tested by experiments that are designed to facilitate falsification[15–17].

A historical example illustrates potential barriers to achieving this goal. In 1543, Nicolaus Copernicus proposed a *heliocentric* (sun-centered) model of the solar system, challenging the dominant *geocentric* (earth-centered) model. This new model offered simpler explanations for puzzling phenomena like retrograde planetary motion, but lacked key details, which led to worse quantitative predictions than the geocentric model – which had been finely tuned to fit observations. Progress required both refinement of the heliocentric model and the development of decisive tests. In the 17th century, Johannes Kepler's discovery of elliptical orbits and Isaac Newton's laws of motion and gravity transformed the heliocentric model, enabling more accurate predictions and the generation of new, testable hypotheses. Crucially, experiments that directly pitted the two theories against each other on a level playing field (e.g., Galileo's observation of the phases of Venus), provided compelling evidence for the heliocentric model, ultimately leading to its universal acceptance.

Similarly, despite important advances in theoretical accounts of hallucinations[1–4,7,11,13], we argue that the field would benefit from models with greater specificity and mechanistic detail that can generate more concrete, testable cognitive and biological predictions. Then, experiments designed to separate these predictions could be used to refine and rule out models[16]. In this vein, here we build from seminal research in the computational psychiatry of hallucinations[2,4,7,14,15] and

apply established best practices in cognitive/computational neuroscience[16,18] to develop a pipeline for the description, prioritization, and testing of hallucination theories.

Theories of hallucinations

Two families of theories have been particularly influential in the mechanistic study of hallucinations: *predictive-processing* and *salience-based* theories.

Predictive-processing theories of hallucinations posit that hallucinations arise from alterations in perceptual inference[3,6,11,13,19–21]. One prominent instantiation, the 'strong-prior' hypothesis, proposes that sensory expectations are overweighted relative to sensory evidence, which leads to false perceptions[1,6]. In general, these theories emphasize the role of environmental uncertainty in shaping expectations[22,23] and dopamine's role in signaling this uncertainty[24–27]. These theories are supported by evidence of altered statistical learning in patients with schizophrenia[20,21,28] and have been formalized using Bayesian models of perceptual inference[3,6].

Salience-based (also called 'aberrant salience') theories posit that hallucinations arise from altered attribution of significance to irrelevant or random stimuli[7,8,29,30]. These theories were originally inspired by the therapeutic efficacy of antidopaminergic drugs[31] and seminal research on dopamine's role in attributing motivational salience[32], but have evolved to incorporate a wide variety mechanisms. Among these is dopamine's role in signaling reward prediction errors (rPEs)[33–35], through which aberrant learning of stimulus value could indirectly distort perceptual experience[2,30,36]. Broadly consistent with this perspective, alterations in reward processing have been identified in patients with schizophrenia[2,19,30,37,38], although specific relationships with hallucinations are less clear. Attempts have also been made to incorporate salience into predictive processing perspectives through related principles like surprise[3,28], but specific neural computations and their biological implementation are still unclear.

A translatable mechanistic hypothesis should specify how an altered circuit computation leads to hallucinations[14] (Figure 1). In other words, the hypothesis should explain how false perceptions arise from an altered cognitive process implemented by a specific biological substrate. Do these hypotheses meet this standard? The 'strong-prior' hypothesis is often described in relatively concrete mathematical terms but typically lacks a precise biological implementation (including in our own previous work[1]). The 'aberrant salience' hypothesis specifies striatal dopamine as a key biological substrate but its cognitive/computational basis has largely been described in abstract terms (but see[2]). Thus, both theories require further development to make precise cognitive and neural predictions that are crucial for translation to clinical applications.

Additionally, these hypotheses have predominantly been tested using separate experimental frameworks. While the 'strong-prior' hypothesis is often evaluated using sensory expectation paradigms[20,21,39–41], the 'aberrant salience' hypothesis is often tested using reward learning tasks[29,37,38,42]. These divergent methods make it difficult to determine whether empirical results reflect the validity of a theory or limitations of an experimental design. To fairly evaluate these and other viable theories, a unified experiment that directly pits these theories against each other would be ideal. Such an experiment would emphasize the qualitative differences between the theories, promoting falsification and refinement.

In sum, we see room for progress in the description and testing of mechanistic theories of hallucinations. In what follows, we endeavor to illustrate a general strategy by which detailed mechanistic hypotheses can be developed and tested with experiments that evaluate them on a level playing field.

Constraining a mechanistic theory of hallucinations

We aim to develop theory-driven computational models of hallucinations grounded in fundamental neural and cognitive mechanisms. Specifically, we focus on generative models that explain how intraindividual and interindividual variability in hallucinations arises. These models can be empirically validated by testing whether the behavioral signatures they predict, such as patterns of misperceptions in a cognitive task, systematically relate to hallucination. In the context of this paper, we are interested in models that account for variability in hallucination level, reflecting the overall intensity and frequency of hallucinations. These models are not intended to account for the specific content of hallucinations.

To achieve this goal, we follow a two-step approach (Figure 2). First, we will draw on established findings in cognitive neuroscience to construct a general computational model. This model will specify key details about relevant cognitive mechanisms and their circuit implementation. Second, we will evaluate whether specific alterations in elements of this general model capture behavioral patterns robustly linked to hallucinations. We will refer to empirical findings we use to narrow the space of potential models as *constraints*[43] (Table 1).

Constraints from the Basic Neuroscience Literature

Hallucinations are internal perceptual experiences that occur in the absence of corresponding external stimuli, and their subjective experience is often comparable to true percepts in terms of their phenomenological (e.g., acoustic) features[44]. Furthermore, substantial evidence indicates that perceptual disturbances, including hallucinations, exist on a continuum across the general population and clinical groups[13,21,45–48]. Together, these results support the building of hallucination theories upon basic theories of perceptual decision-making[6,49].

Perceptual decisions are often made under uncertainty. For example, in a subway station, a sound that resembles your name being called out may be obscured by ambient noise. You can

resolve this uncertainty by incorporating your learned expectations to decide if your name was actually called[22,50]. Specifically, how likely it is that someone would call your name (i.e., sensory expectations) and the desirability of concluding your name was called (i.e., reward expectations). Empirically, expectations (often inferred from sensory and reward histories[13,24,51,52]) bias choices (Constraint A.1). If you were planning to meet someone, both expectations could bias you toward deciding that your name was called, even if no one called your name. Since reward expectations can bias decisions without reflecting a true percept, additional reports such as confidence are useful for ascertaining the determinants of a particular choice[53–55]. In decision-making tasks, confidence reports are also biased by expectations (Constraint A.1) and are commensurate with an observer's subjective probability of being correct (i.e., confidence is higher on correct trials and this effect scales with stimulus difficulty[13,51,52]; Constraint A.2).

At the circuit level, expectation learning is thought to be facilitated by striatal dopamine (Constraints A.3). Specifically, *mesolimbic* dopamine signals reward prediction errors (rPEs)[33–35], which represent the difference between received and expected reward, prompting an update of reward expectations. More recent work in rodents suggests that *nigrostriatal* dopamine encodes distinct, reward-independent information[56,57] relevant to perception[13,58,59]. Consistently, the human striatum is also sensitive to sensory statistics[60] and involved in sensory learning[61]. While the precise computation instantiated by nigrostriatal dopamine is still an area of active research, this evidence supports constraining reward and sensory processing to distinct dopaminergic circuits[13,56–58,62–64] (Constraint A.4; but see[65–68]). Overall, these data support a model whereby dopamine-mediated prediction errors in distinct pathways are used to learn stimulus probability (i.e., how likely it was that my name was called) and decision value (i.e., the desirability of concluding my name was called) which are incorporated into decisions and confidence (Figure 2).

Constraints from the Clinical Neuroscience Literature

In the example above, incorrectly concluding your name was called would in principle be analogous to a hallucination – the two have interchangeable definitions as perceptual reports of a stimulus in its objective absence. Consistent with this conceptual link, higher hallucination propensity correlates with increased reports of experiencing stimuli in signal detection tasks (i.e., false alarms and hits; Constraint A.1) and higher confidence in these decisions (Constraint A.2) in both patients with hallucinations[21,46,69] and the general population[13,46,48,70,71] (Figure S1). Importantly, not all false alarms reflect true hallucinations. However, *high-confidence* false alarms are thought to capture them because they represent a deliberate commitment to reporting a genuine percept, consistent with the typical conviction of experienced hallucinations in psychosis. These results are well-replicated across clinical and non-clinical populations, research groups, and methodological variations, suggesting that a mechanistic explanation for increased high-confidence false alarms is a key ingredient for any computational model of hallucinations. Note that we focus here on auditory hallucinations given their relevance to idiopathic psychotic disorders, although similar cognitive alterations in vision[39–41] suggest that our discussion will also be relevant to perceptual disturbances in other modalities.

At the neural level, converging evidence from multiple lines of work has implicated excess striatal dopamine in the development of hallucinations (reviewed in greater depth elsewhere[6]). Briefly, molecular neuroimaging studies measuring striatal dopamine function have repeatedly shown a specific positive relationship between striatal dopamine function (e.g., dopamine synthesis and release capacity) and the intensity of positive symptoms, including hallucinations[20,72–76]. Consistently, pharmacological studies have demonstrated a dose-dependent relationship between pro-dopaminergic drugs and psychotic symptoms[77–80] and the therapeutic benefits of antidopaminergic drugs on psychotic symptoms are definitively established[31] and depend on

striatal dopamine-receptor blockade[81]. Recent results from the preclinical literature have also causally implicated stimulation of sensory-striatal dopamine in high-confidence false perception[13].

This evidence has been taken to suggest that excess striatal dopamine is likely *sufficient* for the generation of hallucinations – even if it is not necessary[6,82]. Indeed, other neurotransmitters, including glutamate[83–85] and acetylcholine[86–88], have also been implicated in hallucinations, but their mechanistic roles in perceptual decision-making and psychosis are much less clear. Here, we will use the extensive body of research into dopamine function to provide initial – and likely imperative – neurobiological constraints for hallucination theories. Specifically, viable theories should link increased striatal dopamine to elevated rates of high-confidence false alarms (Constraint B.3). We would argue that theories that do not satisfy these constraints (B.1-B.3) should be deprioritized by virtue of their limited explanatory power to account for critical empirical findings.

Decision-making model

Having defined the relevant constraints, we will now proceed with the first step in the model development process: constructing a computational model of decision-making (Figure 2). Computational models are particularly useful for formalizing and testing mechanistic hypotheses[89]. They offer simple and interpretable explanations for complex phenotypes and can generate specific predictions that can be compared against empirical data[89–91]. We do not intend to make strong claims about a specific model architecture but rather demonstrate how a reasonable decision-making model can be used as a scaffold for developing and testing theories of hallucinations.

Perceptual Inference

Consider the subway example where you are trying to decide whether your name was called. Following standard models of perception[6,13] your percept μ_{Po} can be represented as a combination of the external sensory evidence (or likelihood μ_{Lik}) and your internal sensory expectations ν_S (i.e., the prior probability that your name was called). We can write this as a sum of log-odds (or logits).

$$\text{logit}(\mu_{Po}) = \text{logit}(\nu_S) + \text{logit}(\mu_{Lik}) \qquad \textit{Eq. 1}$$

μ_{Lik} is the internal representation of the stimulus s and is drawn from a cumulative normal distribution ϕ with mean b and standard deviation σ:

$$\mu_{Lik} = \phi(s; b, \sigma) \qquad \textit{Eq. 2}$$

Choice

Your decision θ is influenced by both your percept μ_{Po} and your reward expectation (i.e. the desirability of concluding your name was called). We represent reward expectation ν_R as the probability of a reward following a "Yes" choice ($\theta = 1$), with the probability of reward for a no choice ($\theta = 0$) being complementary. This simplifies the choice computation and enables more

direct comparisons with sensory expectations due to the parallel structure. The decision with the larger expected value is made[22]:

$$\theta = \begin{cases} 1, & \text{if } [\text{logit}(\nu_R) + \text{logit}(\mu_{Po})] \geq 0 \\ 0, & \text{otherwise} \end{cases} \qquad \textit{Eq. 3}$$

Note that this decision rule is equivalent to a signal detection framing in which sensory evidence is compared to a decision criterion that is informed by expectations (Supplemental Methods).

Confidence

As we discussed above, behavioral reports like confidence can be used to dissociate the relative influences of perception and reward on decisions[53–55]. Following similar models[13,22,24], perceptual confidence C_P is derived from perceptual belief μ_{Po} while reward confidence C_R is derived from reward expectations v_R. Decision confidence C_D integrates both into a single metric using a logit sum:

$$C_P = \mu_{Po} \qquad \textit{Eq. 4}$$

$$C_R = \begin{cases} v_R, & if\ \theta = 1 \\ 1 - v_R, & if\ \theta = 0 \end{cases} \qquad \textit{Eq. 5}$$

$$\mathrm{logit}(C_D) = \mathrm{logit}(C_P) + \mathrm{logit}(C_R) \qquad \textit{Eq. 6}$$

Reward Learning

So far, we have considered a single decision, but sensory and reward expectations are dynamic and are learned through experience with environment. Reward expectations are learned through reward prediction errors δ_R^t using the delta rule[33–36,51,92] at the current time point t (Constraint A.3):

$$\delta_R^t = (R^t - C_R^t) \qquad \textit{Eq. 7}$$

where R^t is the reward outcome (i.e., the choice was or was not desirable). For simplicity, we will consider this to be a binary outcome. This reward prediction error is then scaled by a learning rate α_R and used to update your reward expectation:

$$v_R^{t+1} = \begin{cases} v_R^t + \alpha_R \cdot \delta_R^t, if\ \theta = 1 \\ v_R^t - \alpha_R \cdot \delta_R^t, if\ \theta = 0 \end{cases} \qquad \textit{Eq. 8}$$

Sensory Learning

Sensory expectations are learned through distinct error signals (Constraints A.4). Sensory prediction errors δ_S^t are computed using a delta rule [13]:

$$\delta_S^t = (F^t - C_P^t) \qquad \textit{Eq. 9}$$

where F^t is sensory feedback (i.e., your name was or was not called). For simplicity, we will consider a case where veridical feedback is available but more sophisticated learning rules can operate without feedback [93]. This sensory prediction error is then scaled by a learning rate α_S and used to update the observer's sensory expectation v_S:

$$v_S^{t+1} = v_S^t + \alpha_S \cdot \delta_S^t \qquad \textit{Eq. 10}$$

Each time step, sensory and reward expectations will shift in the direction of the true outcome (i.e., sensory feedback or reward) and the magnitude of the shift will depend on how unexpected the outcome was.

Overall, the model uses learned expectations (v_S and v_R) and sensory evidence (μ_{Lik}) to decide if a signal was present or absent on each trial. This structure results in choices and confidence that are sensitive to stimulus and reward history (Constraint A.1) and the degree of sensory evidence (Constraint A.2) which can be seen in simulated choice and confidence reports (Figures S2-3).

Prioritizing hallucination models

Having defined a plausible decision-making model, we now proceed to the second step of the model development process: identifying alterations in the model parameters that reproduce hallucination-related behavioral patterns (Constraints B.1 & B.2; Figure 2). Rather than exhaustively outlining all potential models, we illustrate how a constraint-based strategy can be used to prioritize the most promising candidates, inspired by similar efforts in other fields like genomics[94].

Assessing Candidate Models

To evaluate a candidate model, we simulate behavior on a signal detection task (see Supplemental Methods). We consider parameter alterations that could plausibly arise from excess striatal dopamine (Constraint B.3) and conclude that an alteration meets prioritization criteria if it increases both signal decision bias (Constraint B.1) and false alarm confidence (Constraint B.2; Figures 3, S4). As illustrative test cases, we focus on models corresponding to the 'strong-prior' and 'aberrant salience' hypotheses.

'Strong Prior': The 'strong prior' hypothesis proposes that increased prior weighting biases perception away from sensory evidence, leading to hallucinations. A strong prior is in principle analogous to a reduced learning rate in our model (i.e., smaller α; see Supplemental Methods Eqs. S5-S7) because the learning rate governs the degree to which new evidence is used to update beliefs. Therefore observers with low learning rates maintain their prior beliefs even when faced with contradictory evidence. We find that models with slow sensory or reward learning did not pass our prioritization criteria (Figure 3).

However, we note that patients with hallucinations display increased signal decisions in auditory detection tasks with stable statistics (e.g., probability of signal is maintained at 0.5 throughout the task[13,46,48,69]) and voice hearers show increased ability to detect corrupted speech without prior

exposure[95]. These findings suggest that a candidate alteration could reflect a pre-existing bias (or a prior acquired based on experiences outside the task) which can be operationalized in our framework by biasing initial sensory expectations ($v_S^{t=0} > 0.5$). But if an observer with such initial bias learns normally, their expectations would adjust over time, inconsistent with empirical data. When biased initial sensory expectations are combined with a reduced sensory learning rate (consistent with a type of pre-existing 'strong prior'), these observers maintain their biased sensory expectations despite contradictory evidence. Models with an initial bias and reduced learning in either the sensory or reward domains satisfy our prioritization criteria (Figure 3; Table 2). While the 'strong prior' hypothesis does not necessarily prescribe an initial bias in expectations[1], such a bias is nonetheless consistent with this hypothesis and seems necessary to account for empirical data.

'Aberrant Salience': Reward learning-based implementations of the 'aberrant salience' hypothesis propose that striatal dopamine indirectly alters perception by disrupting the learning of the value associated with stimuli through changes in rPEs[30,36]. This hypothesis has previously been operationalized by assuming that positive reward prediction errors are of larger absolute magnitude than their negative counterparts in psychosis ($|\delta_R^+| > |\delta_R^-|$)[2,36]. Consistent with this notion, genotypes known to influence striatal dopamine[96] and pharmacological dopamine agonism[97,98] have been linked to asymmetric learning biases in humans. This mechanistic interpretation can be approximated in our modeling framework with biased learning from positive reward predictions ($\alpha_R^+ > \alpha_R^-$). This model does not pass our prioritization criteria (Table 2). In contrast, a model with biased learning from sensory predictions ($\alpha_S^+ > \alpha_S^-$) does pass or prioritization criteria. However, since this model would develop exaggerated sensory expectations, it relates more closely to predictive-processing theories.

In general, our theoretical results support deprioritizing models that foreground standard value learning rPE-based mechanisms. Consistent with this, previous empirical attempts have generally

failed to produce evidence linking altered value learning rPE signals and hallucinations[2,30,38,99,100]. That said, models that propose a more nuanced role for rPEs, or those focused on cortical gating mechanisms[2], could still be considered. For example, the value associated with certain environmental states (e.g., signal presence versus absence) can bias attentional[101] or sensory processes[59,65–67,102] to influence perceptual decisions. Additionally, dopamine signals may represent complex combinations of reward and sensory features[57,103–105] (e.g., threat prediction-errors[106]), so models of hallucinations that consider reward-sensory crosstalk may deserve further consideration. Speaking to this, one possible model that satisfies our prioritization criteria learns more rapidly when signal trials are rewarded (hybrid bias model; Table 2).

So far, we have illustrated a strategy to develop and prioritize biologically and computationally detailed models of hallucinations (Table 2). This tentative taxonomy relates to, but does not map perfectly onto, prevalent hypotheses: only certain concrete operationalizations of a given hypothesis satisfy our a priori criteria. We believe that this exercise thus highlights the importance of concretely specifying generative models and comparing their predictions against empirical data.

Testing prioritized models

Now that we have prioritized a set of candidate models, we will evaluate whether they can be empirically distinguished in a hypothetical experiment. This is a nontrivial challenge because all prioritized models have shared behavioral patterns (Constraints B.2 & B.3). As with comparisons between heliocentric and geocentric models, the most compelling tests elicit qualitatively distinct signatures from each candidate model.

Here, we focus on signal detection tasks due to the robust relationships previously identified with hallucination level. However, in principle, other tasks that probe sensory and reward learning (e.g., conditioning paradigms) could also be considered as long as they permit reliable estimation of learning from behavior.

Standard Tasks Are Limited in Their Ability to Distinguish Prioritized Models

In standard signal-detection tasks in the psychosis literature, sensory statistics are often kept stable, so participants can perform well without learning. Furthermore, participants are usually rewarded for perceptual accuracy or are not explicitly rewarded for performance. So even if participants do learn, reward and sensory feedback are highly correlated which makes it challenging to determine if choices are driven by sensory or reward expectations. This dissociation is particularly important to evaluate theories that consider interactions between reward and sensory domains, like the hybrid bias model (Table 2). One approach to reveal the limitations of a task once we have a set of prioritized computational models is based on model simulations and recovery analyses that treat simulated data as a ground truth[18]. To distinguish between these models, at minimum key model parameters – the parameter(s) for a given model that drive increased high-confidence false alarms – must be recoverable in simulations. For example, to test initial bias models, participants' initial expectations and learning rates for each

domain ($\nu_S^{t=0}, \nu_R^{t=0}, \alpha_S, \alpha_R$) must be estimated. However, using a standard task, parameters are poorly recovered (Figure S6; Supplemental Methods).

Designing a Task that Separates Model Variants

Given limitations of existing paradigms and to illustrate actionable next steps of our model prioritization scheme, we next set out to develop a novel task that would be better at empirically separating prioritized models. To address abovementioned concerns, we propose a task where the probability of a signal trial and the probability that a "yes" response yields additional reward are varied independently in a blocked structure (Figure 4). Data from an ongoing study suggests that this task is feasible in human participants and show that task manipulations modulate behaviors in line with our model predictions[62].

This structure incentivizes simultaneous sensory and reward learning and decorrelates perceptual correctness from reward which should enable appropriate estimation of learning dynamics. If so, a recovery analysis should be able to demonstrate improvements, as our simulations show (Figure S7). Finally, a sufficiently rich task design should allow different models to make *qualitatively* distinct or even opposing predictions apparent in model-agnostic analyses of raw behavioral data. For example, the asymmetric sensory learning and initial sensory bias models make opposing predictions about how signal probability influences the evolution of perceptual confidence over the course of a block (Figure S8). By combining several of these metrics, unique 'fingerprints' of these models can be identified and models that do not exhibit signatures observed in the data can be falsified (Figure 4; Figure S8).

In sum, once a set of models have been prioritized, we recommend testing them using a paradigm that emphasizes their unique signatures[18]. Our simulations suggest that existing paradigms fall short of this goal, but that computational models can be used to design new experiments that isolate the unique model signatures and rule out models inconsistent with empirical results.

Discussion

Exciting advances in the development of hallucination theories have taken place in recent years[1–3,5,6,8–13,36]. Several distinct mechanistic hypotheses of hallucinations are under active investigation, but progress in arbitrating between them has been limited by challenging operationalizations of some hypotheses and a dearth of experimental frameworks designed to directly compare and falsify them. Falsification is particularly important given the continued appeal of influential theories like 'aberrant salience'[7,8]. The face value of this hypothesis has made it a popularly invoked explanation for a wide array of disparate results[107]. Critically, our intention is not to question the value of this hypothesis or the evidence supporting altered reward processing in psychosis. Indeed, one of our prioritized models (hybrid bias) is broadly consistent with the 'aberrant salience' hypothesis. In more general terms, we argue that it seems counterproductive to continue debating theories that do not map onto a unique circuit computation (Figure 1), and that it may be more fruitful to test sets of specific operationalizations of these theories that do.

To this end, here we described a step-by-step strategy for specifying, prioritizing, and testing hallucination theories. We first developed a constrained model of decision-making which provided a base architecture to formulate a series of potential models of hallucinations (Figure 2). This architecture was informed by generally accepted models of reward learning and perceptual decision-making which have been extensively validated by convergent translational data[13,22,33,35,51,62]. While there are other ways to instantiate such a model (e.g., alternative learning rules), predictions would likely be qualitatively similar (see Figure S5). We then prioritized models that capture key empirical data in the hallucination literature (Figure 3). Since these prioritized models each map onto a unique circuit computation, we see them as a starting point for a new, falsifiable taxonomy of mechanistic hallucination theories (Table 2). This taxonomy can be used to identify the most promising subtypes of prominent theories in the field. For example, our simulations suggest that the initial sensory bias model should be favored over the slow sensory

learning model to operationalize the 'strong prior' hypothesis. Importantly, this general pipeline can be readily extended to operationalize other leading cognitive (e.g., related to volatility[12,28] and other aspects of inference[1,20,108]) and neurobiological (e.g., the role of acetylcholine in modulating top-down predictions[88,109]) hypotheses in the literature. Finally, we demonstrated that a key strength of prioritizing models in this manner is that it enables the design of experiments that emphasize their qualitative differences, which promotes model refinement and falsification[16,18]. Moreover, since these experimental paradigms are optimized to test multiple hypotheses simultaneously, they avoid reliance on a single a priori hypothesis and are well-suited to uncover distinct mechanisms for hallucinations across subpopulations.

Overall, we believe that progress in our mechanistic understanding of hallucinations will require testing a broader set of model families selected in a principled manner together with a carefully defined criteria for falsification. Identifying a more precise underlying model with a distinct neurobiological substrate could ultimately facilitate the development of more targeted treatments with increased efficacy and tolerability that improve patients' quality of life.

Acknowledgements

Funding was provided by the National Institute of Mental Health under award numbers R01MH114965 to G.H. and F31MH134617 to J.B. and the BBRF NARSAD Young Investigator Grant to K.I. The authors would like to thank members of the Horga and Iigaya labs for helpful discussions. The brain image in Figure 1 was created using Biorender (Created in BioRender. Horga, G. (2025) https://BioRender.com/tyeefsf).

A previous version of this article was published as a preprint on PsyArXiv: https://osf.io/preprints/psyarxiv/jpzfw.

Author Contributions

Conceptualization: JB and GH

Simulations: JB

Drafting of the manuscript: JB, KI, and GH

Supervision: KI and GH

Manuscript revision and approval of final version: JB, KI, and GH

Financial Disclosures

The authors report no relevant conflicts of interest.

References

1 Corlett PR, Horga G, Fletcher PC, Alderson-Day B, Schmack K, Powers AR. Hallucinations and Strong Priors. *Trends in Cognitive Sciences* 2019; **23**: 114–127.

2 Maia TV, Frank MJ. An Integrative Perspective on the Role of Dopamine in Schizophrenia. *Biological Psychiatry* 2017; **81**: 52–66.

3 Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L *et al.* The Predictive Coding Account of Psychosis. *Biological Psychiatry* 2018; **84**: 634–643.

4 Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci* 2009; **10**: 48–58.

5 Sheldon AD, Kafadar E, Fisher V, Greenwald MS, Aitken F, Negreira AM *et al.* Perceptual pathways to hallucinogenesis. *Schizophrenia Research* 2022; **245**: 77–89.

6 Horga G, Abi-Dargham A. An integrative framework for perceptual disturbances in psychosis. *Nat Rev Neurosci* 2019; **20**: 763–778.

7 Kapur S. Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia. *AJP* 2003; **160**: 13–23.

8 Heinz A, Schlagenhauf F. Dopaminergic Dysfunction in Schizophrenia: Salience Attribution Revisited. *Schizophrenia Bulletin* 2010; **36**: 472–485.

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

10 Thakkar KN, Mathalon DH, Ford JM. Reconciling competing mechanisms posited to underlie auditory verbal hallucinations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2020; **376**: 20190702.

11 Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The Computational Anatomy of Psychosis. *Front Psychiatry* 2013; **4**. doi:10.3389/fpsyt.2013.00047.

12 Katthagen T, Fromm S, Wieland L, Schlagenhauf F. Models of Dynamic Belief Updating in Psychosis—A Review Across Different Computational Approaches. *Front Psychiatry* 2022; **13**. doi:10.3389/fpsyt.2022.814111.

13 Schmack K, Bosc M, Ott T, Sturgill JF, Kepecs A. Striatal dopamine mediates hallucination-like perception in mice. *Science* 2021; **372**: eabf4740.

14 Wang X-J, Krystal JH. Computational Psychiatry. *Neuron* 2014; **84**: 638–654.

15 Teufel C, Fletcher PC. The promises and pitfalls of applying computational models to neurological and psychiatric disorders. *Brain* 2016; **139**: 2600–2608.

16 Palminteri S, Wyart V, Koechlin E. The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences* 2017; **21**: 425–433.

17 Miłkowski M, Litwin P. Testable or bust: theoretical lessons for predictive processing. *Synthese* 2022; **200**: 462.

18 Wilson RC, Collins AG. Ten simple rules for the computational modeling of behavioral data. *eLife* 2019; **8**: e49547.

19 Murray GK, Corlett PR, Clark L, Pessiglione M, Blackwell AD, Honey G *et al.* How dopamine dysregulation leads to psychotic symptoms? Abnormal mesolimbic and mesostriatal prediction error signalling in psychosis. *Mol Psychiatry* 2008; **13**: 239–239.

20 Cassidy CM, Balsam PD, Weinstein JJ, Rosengard RJ, Slifstein M, Daw ND *et al.* A Perceptual Inference Mechanism for Hallucinations Linked to Striatal Dopamine. *Current Biology* 2018; **28**: 503-514.e4.

21 Powers AR, Mathys C, Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science* 2017; **357**: 596–600.

22 Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience* 2008; **8**: 429–453.

23 Nassar MR, Wilson RC, Heasly B, Gold JI. An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *J Neurosci* 2010; **30**: 12366–12378.

24 Lak A, Okun M, Moss MM, Gurnani H, Farrell K, Wells MJ *et al.* Dopaminergic and Prefrontal Basis of Learning from Sensory Confidence and Reward Value. *Neuron* 2020; **105**: 700-711.e6.

25 Andreou C, Bozikas VP, Luedtke T, Moritz S. Associations between visual perception accuracy and confidence in a dopaminergic manipulation study. *Front Psychol* 2015; **6**. doi:10.3389/fpsyg.2015.00414.

26 Lou HC, Skewes JC, Thomsen KR, Overgaard M, Lau HC, Mouridsen K *et al.* Dopaminergic stimulation enhances confidence and accuracy in seeing rapidly presented words. *Journal of Vision* 2011; **11**: 15.

27 Fiorillo CD, Tobler PN, Schultz W. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science* 2003; **299**: 1898–1902.

28 Nassar MR, Waltz JA, Albrecht MA, Gold JM, Frank MJ. All or nothing belief updating in patients with schizophrenia reduces precision and flexibility of beliefs. *Brain* 2021; **144**: 1013–1029.

29 Winton-Brown TT, Fusar-Poli P, Ungless MA, Howes OD. Dopaminergic basis of salience dysregulation in psychosis. *Trends in Neurosciences* 2014; **37**: 85–94.

30 Jensen J, Willeit M, Zipursky RB, Savina I, Smith AJ, Menon M *et al.* The Formation of Abnormal Associations in Schizophrenia: Neural and Behavioral Evidence. *Neuropsychopharmacol* 2008; **33**: 473–479.

31 Emsley R, Rabinowitz J, Medori R. Time Course for Antipsychotic Treatment Response in First-Episode Schizophrenia. *AJP* 2006; **163**: 743–745.

32 Berridge KC, Robinson TE. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* 1998; **28**: 309–369.

33 Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 2006; **442**: 1042–1045.

34 Pagnoni G, Zink CF, Montague PR, Berns GS. Activity in human ventral striatum locked to errors of reward prediction. *Nat Neurosci* 2002; **5**: 97–98.

35 Schultz W, Dayan P, Montague PR. A Neural Substrate of Prediction and Reward. *Science* 1997; **275**: 1593–1599.

36 Smith AJ, Li M, Becker S, Kapur S. Linking Animal Models of Psychosis to Computational Models of Dopamine Function. *Neuropsychopharmacol* 2007; **32**: 54–66.

37 Roiser JP, Howes OD, Chaddock CA, Joyce EM, McGuire P. Neural and Behavioral Correlates of Aberrant Salience in Individuals at Risk for Psychosis. *Schizophrenia Bulletin* 2013; **39**: 1328–1336.

38 Roiser JP, Stephan KE, Ouden HEM den, Barnes TRE, Friston KJ, Joyce EM. Do patients with schizophrenia exhibit aberrant salience? *Psychological Medicine* 2009; **39**: 199–209.

39 Teufel C, Subramaniam N, Dobler V, Perez J, Finnemann J, Mehta PR *et al.* Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proc Natl Acad Sci U S A* 2015; **112**: 13401–13406.

40 Zarkali A, Adams RA, Psarras S, Leyland L-A, Rees G, Weil RS. Increased weighting on prior knowledge in Lewy body-associated visual hallucinations. *Brain Communications* 2019; **1**: fcz007.

41 Bansal S, Bae G-Y, Robinson BM, Hahn B, Waltz J, Erickson M *et al.* Association Between Failures in Perceptual Updating and the Severity of Psychosis in Schizophrenia. *JAMA Psychiatry* 2022; **79**: 169–177.

42 Boehme R, Deserno L, Gleich T, Katthagen T, Pankow A, Behr J *et al.* Aberrant Salience Is Related to Reduced Reinforcement Learning Signals and Elevated Dopamine Synthesis Capacity in Healthy Adults. *J Neurosci* 2015; **35**: 10103–10111.

43 Kriegeskorte N, Douglas PK. Cognitive computational neuroscience. *Nat Neurosci* 2018; **21**: 1148–1160.

44 Larøi F, Sommer IE, Blom JD, Fernyhough C, ffytche DH, Hugdahl K *et al.* The Characteristic Features of Auditory Verbal Hallucinations in Clinical and Nonclinical Groups: State-of-the-Art Overview and Future Directions. *Schizophrenia Bulletin* 2012; **38**: 724–733.

45 Baumeister D, Sedgwick O, Howes O, Peters E. Auditory verbal hallucinations and continuum models of psychosis: A systematic review of the healthy voice-hearer literature. *Clinical Psychology Review* 2017; **51**: 125–141.

46 Moseley P, Alderson-Day B, Common S, Dodgson G, Lee R, Mitrenga K *et al.* Continuities and Discontinuities in the Cognitive Mechanisms Associated With Clinical and Nonclinical Auditory Verbal Hallucinations. *Clinical Psychological Science* 2022; **10**: 752–766.

47 Verdoux H, van Os J. Psychotic symptoms in non-clinical populations and the continuum of psychosis. *Schizophrenia Research* 2002; **54**: 59–65.

48 Moseley P, Aleman A, Allen P, Bell V, Bless J, Bortolon C *et al.* Correlates of Hallucinatory Experiences in the General Population: An International Multisite Replication Study. *Psychol Sci* 2021; **32**: 1024–1037.

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

50 Gold JI, Shadlen MN. The Neural Basis of Decision Making. *Annual Review of Neuroscience* 2007; **30**: 535–574.

51 Lak A, Hueske E, Hirokawa J, Masset P, Ott T, Urai AE *et al.* Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *eLife* 2020; **9**: e49834.

52 Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Current Biology* 2017; **27**: 821–832.

53 Mihali A, Broeker M, Ragalmuto FDM, Horga G. Introspective inference counteracts perceptual distortion. *Nat Commun* 2023; **14**: 7826.

54 Maldonado Moscoso PA, Cicchini GM, Arrighi R, Burr DC. Adaptation to hand-tapping affects sensory processing of numerosity directly: evidence from reaction times and confidence. *Proceedings of the Royal Society B: Biological Sciences* 2020; **287**: 20200801.

55 Gallagher RM, Suddendorf T, Arnold DH. Confidence as a diagnostic tool for perceptual aftereffects. *Sci Rep* 2019; **9**: 7124.

56 Menegas W, Babayan BM, Uchida N, Watabe-Uchida M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife* 2017; **6**: e21886.

57 Menegas W, Akiti K, Amo R, Uchida N, Watabe-Uchida M. Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nat Neurosci* 2018; **21**: 1421–1430.

58 Chen APF, Malgady JM, Chen L, Shi KW, Cheng E, Plotkin JL *et al.* Nigrostriatal dopamine pathway regulates auditory discrimination behavior. *Nat Commun* 2022; **13**: 5942.

59 Xiong Q, Znamenskiy P, Zador AM. Selective corticostriatal plasticity during acquisition of an auditory discrimination task. *Nature* 2015; **521**: 348–351.

60 Vilares I, Howard JD, Fernandes HL, Gottfried JA, Kording KP. Differential Representations of Prior and Likelihood Uncertainty in the Human Brain. *Current Biology* 2012; **22**: 1641–1648.

61 Feng G, Yi HG, Chandrasekaran B. The Role of the Human Auditory Corticostriatal Network in Speech Learning. *Cerebral Cortex* 2019; **29**: 4077–4089.

62 Lakshminarasimhan K, Buck J, Kellendonk C, Horga G. A corticostriatal learning mechanism linking excess striatal dopamine and auditory hallucinations. 2025; : 2025.03.18.643990.

63 Riley B, Gould E, Lloyd J, Hallum LE, Vlajkovic S, Todd K *et al.* Dopamine transmission in the tail striatum: Regional variation and contribution of dopamine clearance mechanisms. *Journal of Neurochemistry* 2024; **168**: 251–268.

64 Valjent E, Gangarossa G. The Tail of the Striatum: From Anatomy to Connectivity and Function. *Trends in Neurosciences* 2021; **44**: 203–214.

65 Kahnt T, Grueschow M, Speck O, Haynes J-D. Perceptual Learning and Decision-Making in Human Medial Frontal Cortex. *Neuron* 2011; **70**: 549–559.

66 Guggenmos M, Wilbertz G, Hebart MN, Sterzer P. Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* 2016; **5**: e13388.

67 Arsenault JT, Vanduffel W. Ventral midbrain stimulation induces perceptual learning and cortical plasticity in primates. *Nat Commun* 2019; **10**: 3591.

68 Sharpe MJ, Chang CY, Liu MA, Batchelor HM, Mueller LE, Jones JL *et al.* Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat Neurosci* 2017; **20**: 735–742.

69 Bentall RP, Slade PD. Reality testing and auditory hallucinations: A signal detection analysis. *British Journal of Clinical Psychology* 1985; **24**: 159–169.

70 Kafadar E, Fisher VL, Quagan B, Hammer A, Jaeger H, Mourgues C *et al.* Conditioned Hallucinations and Prior Overweighting Are State-Sensitive Markers of Hallucination Susceptibility. *Biological Psychiatry* 2022; **92**: 772–780.

71 Boer JN de, Linszen MMJ, Vries J de, Schutte MJL, Begemann MJH, Heringa SM *et al.* Auditory hallucinations, top-down processing and language perception: a general population study. *Psychological Medicine* 2019; **49**: 2772–2780.

72 Howes OD, Bose SK, Turkheimer F, Valli I, Egerton A, Valmaggia LR *et al.* Dopamine Synthesis Capacity Before Onset of Psychosis: A Prospective [18F]-DOPA PET Imaging Study. *AJP* 2011; **168**: 1311–1317.

73 Howes O, Bose S, Turkheimer F, Valli I, Egerton A, Stahl D *et al.* Progressive increase in striatal dopamine synthesis capacity as patients develop psychosis: a PET study. *Mol Psychiatry* 2011; **16**: 885–886.

74 Cassidy CM, Zucca FA, Girgis RR, Baker SC, Weinstein JJ, Sharp ME *et al.* Neuromelanin-sensitive MRI as a noninvasive proxy measure of dopamine function in the human brain. *Proceedings of the National Academy of Sciences* 2019; **116**: 5108–5117.

75 Laruelle M, Abi-Dargham A, van Dyck CH, Gil R, D'Souza CD, Erdos J *et al.* Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. *Proc Natl Acad Sci U S A* 1996; **93**: 9235–9240.

76 Abi-Dargham A, Gil R, Krystal J, Baldwin RM, Seibyl JP, Bowers M *et al.* Increased Striatal Dopamine Transmission in Schizophrenia: Confirmation in a Second Cohort. *AJP* 1998; **155**: 761–767.

77 McKetin R, Lubman DI, Baker AL, Dawe S, Ali RL. Dose-related psychotic symptoms in chronic methamphetamine users: evidence from a prospective longitudinal study. *JAMA Psychiatry* 2013; **70**: 319–324.

78 McKetin R, McLaren J, Lubman DI, Hides L. The prevalence of psychotic symptoms among methamphetamine users. *Addiction* 2006; **101**: 1473–1478.

79 Beaulieu-Boire I, Lang AE. Behavioral effects of levodopa. *Movement Disorders* 2015; **30**: 90–102.

80 Moskovitz C, Moses H, Klawans HL. Levodopa-induced psychosis: a kindling phenomenon. *Am J Psychiatry* 1978; **135**: 669–675.

81 Yun S, Yang B, Anair JD, Martin MM, Fleps SW, Pamukcu A *et al.* Antipsychotic drug efficacy correlates with the modulation of D1 rather than D2 receptor-expressing striatal projection neurons. *Nat Neurosci* 2023; **26**: 1417–1428.

82 Tost H, Alam T, Meyer-Lindenberg A. Dopamine and psychosis: Theory, pathomechanisms and intermediate phenotypes. *Neuroscience & Biobehavioral Reviews* 2010; **34**: 689–700.

83 Leptourgos P, Bansal S, Dutterer J, Culbreth A, Powers A III, Suthaharan P *et al.* Relating Glutamate, Conditioned, and Clinical Hallucinations via 1H-MR Spectroscopy. *Schizophrenia Bulletin* 2022; **48**: 912–920.

84 Jardri R, Hugdahl K, Hughes M, Brunelin J, Waters F, Alderson-Day B *et al.* Are Hallucinations Due to an Imbalance Between Excitatory and Inhibitory Influences on the Brain? *Schizophrenia Bulletin* 2016; **42**: 1124–1134.

85 Adams RA, Pinotsis D, Tsirlis K, Unruh L, Mahajan A, Horas AM *et al.* Computational Modeling of Electroencephalography and Functional Magnetic Resonance Imaging Paradigms Indicates a Consistent Loss of Pyramidal Cell Synaptic Gain in Schizophrenia. *Biological Psychiatry* 2022; **91**: 202–215.

86 Gritton HJ, Howe WM, Mallory CS, Hetrick VL, Berke JD, Sarter M. Cortical cholinergic signaling controls the detection of cues. *Proceedings of the National Academy of Sciences* 2016; **113**: E1089–E1097.

87 Perry EK, Perry RH. Acetylcholine and Hallucinations - Disease-Related Compared to Drug-Induced Alterations in Human Consciousness. *Brain and Cognition* 1995; **28**: 240–258.

88 Warburton DM, Wesnes K, Edwards J, Larrad D. Scopolamine and the sensory conditioning of hallucinations. *Neuropsychobiology* 1985; **14**: 198–202.

89 Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci* 2016; **19**: 404–413.

90 Schlagenhauf F, Huys QJM, Deserno L, Rapp MA, Beck A, Heinze H-J *et al.* Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage* 2014; **89**: 171–180.

91 Huys QJ, Pizzagalli DA, Bogdan R, Dayan P. Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biology of Mood & Anxiety Disorders* 2013; **3**: 12.

92 Rescorla R, Wagner A. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical Conditioning II: Current Research and Theory*. 1972.

93 Zylberberg A, Wolpert DM, Shadlen MN. Counterfactual Reasoning Underlies the Learning of Priors in Decision Making. *Neuron* 2018; **99**: 1083-1097.e6.

94 Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *Am J Hum Genet* 2010; **86**: 6–22.

95 Alderson-Day B, Lima CF, Evans S, Krishnan S, Shanmugalingam P, Fernyhough C *et al.* Distinct processing of ambiguous speech in people with non-clinical auditory verbal hallucinations. *Brain* 2017; **140**: 2475–2489.

96 Doll BB, Hutchison KE, Frank MJ. Dopaminergic Genes Predict Individual Differences in Susceptibility to Confirmation Bias. *J Neurosci* 2011; **31**: 6188–6198.

97 Pagnier GJ, Asaad WF, Frank MJ. Double dissociation of dopamine and subthalamic nucleus stimulation on effortful cost/benefit decision making. *Current Biology* 2024; **34**: 655-660.e3.

98 Voon V, Pessiglione M, Brezing C, Gallea C, Fernandez HH, Dolan RJ *et al.* Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron* 2010; **65**: 135–142.

99 Waltz JA, Schweitzer JB, Gold JM, Kurup PK, Ross TJ, Jo Salmeron B *et al.* Patients with Schizophrenia have a Reduced Neural Response to Both Unpredictable and Predictable Primary Reinforcers. *Neuropsychopharmacol* 2009; **34**: 1567–1577.

100 Deserno L, Boehme R, Heinz A, Schlagenhauf F. Reinforcement Learning and Dopamine in Schizophrenia: Dimensions of Symptoms or Specific Features of a Disease Group? *Front Psychiatry* 2013; **4**. doi:10.3389/fpsyt.2013.00172.

101 Asutay E, Västfjäll D. Auditory attentional selection is biased by reward cues. *Sci Rep* 2016; **6**: 36989.

102 David SV, Fritz JB, Shamma SA. Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences* 2012; **109**: 2144–2149.

103 Takahashi YK, Batchelor HM, Liu B, Khanna A, Morales M, Schoenbaum G. Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. *Neuron* 2017; **95**: 1395-1405.e3.

104 Sharpe MJ, Batchelor HM, Mueller LE, Yun Chang C, Maes EJP, Niv Y *et al.* Dopamine transients do not act as model-free prediction errors during associative learning. *Nat Commun* 2020; **11**: 106.

105 Stalnaker TA, Howard JD, Takahashi YK, Gershman SJ, Kahnt T, Schoenbaum G. Dopamine neuron ensembles signal the content of sensory prediction errors. *eLife* 2019; **8**: e49315.

106 Akiti K, Tsutsui-Kimura I, Xie Y, Mathis A, Markowitz JE, Anyoha R *et al.* Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron* 2022; **110**: 3789-3804.e9.

107 Jensen J, Kapur S. Salience and psychosis: moving from theory to practise: A commentary on: 'Do patients with schizophrenia exhibit aberrant salience?' by Roiser et al. (2008). *Psychological Medicine* 2009; **39**: 197–198.

108 Corlett PR, Fraser KM. 20 Years of Aberrant Salience in Psychosis: What Have We Learned? *AJP* 2025; : appi.ajp.20240556.

109 Bao S, Chan VT, Merzenich MM. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature* 2001; **412**: 79–83.

110 Molina JL, Joshi YB, Nungaray JA, Thomas ML, Sprock J, Clayson PE *et al.* Central auditory processing deficits in schizophrenia: Effects of auditory-based cognitive training. *Schizophrenia Research* 2021; **236**: 135–141.

111 Javitt DC. Sensory Processing in Schizophrenia: Neither Simple nor Intact. *Schizophrenia Bulletin* 2009; **35**: 1059–1064.

112 Rolls ET, Loh M, Deco G, Winterer G. Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nat Rev Neurosci* 2008; **9**: 696–709.

113 Nagy H, Levy-Gigi E, Somlai Z, Takáts A, Bereczki D, Kéri S. The Effect of Dopamine Agonists on Adaptive and Aberrant Salience in Parkinson's Disease. *Neuropsychopharmacol* 2012; **37**: 950–958.

114 Frank MJ, O'Reilly RC. A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience* 2006; **120**: 497–517.

115 Pinto SR, Uchida N. Tonic dopamine and biases in value learning linked through a biologically inspired reinforcement learning model. *bioRxiv* 2023; : 2023.11.10.566580.

Figure Legends

Figure 1. Bridging changes in cognition and psychiatric symptoms with explicit circuit mechanisms. Translatable theories should explain how a cognitive process is implemented by a biological circuit and how alterations in this mechanism drive altered computations and the symptom of interest. For example, extensive evidence has linked both excess striatal dopamine transmission and shifted decision criteria to the development and level of hallucinations. However, the underlying cognitive and circuit mechanisms remain unclear. Theories that propose circuit mechanisms mediating the relationship between excess striatal dopamine, biased decision criteria, and hallucinations are essential for translation. For example, excess mesostriatal or nigrostriatal dopamine may correspond to altered reward prediction error (rPE) or sensory prediction error (sPE) signaling, respectively. Both options could bias expectation learning and influence perception but would represent distinct treatment targets.

Figure 2. Operationalizing mechanistic theories of hallucinations using computational modeling. We will use a two-step process to specify detailed models of hallucinations (top). First, we will apply constraints from the basic neuroscience literature (group A) to build a plausible decision-making model. Second, we will use constraints from the human clinical literature (group B) to prioritize candidate models of hallucinations. Group A constraints led to the development of a decision-making model that updates sensory and reward expectations based on experience and integrates these expectations with current sensory evidence to determine if a signal was present or absent on each trial (bottom).

Figure 3. Prioritizing candidate hallucination model variants. Candidate models of hallucinations can be prioritized by their ability to account for well-established empirical findings. Given the robust relationship between hallucination propensity and increased signal decision bias and false alarm confidence in signal detection tasks, we expect a candidate alteration to also drive these behaviors (left). To evaluate these criteria, we simulated observers across a range of

parameter alteration magnitudes. If *both* signal decision bias and false alarm confidence scale with a parameter alteration magnitude, we prioritize that model. For example, we find that as learning rate decreases, (i.e., slow learning; Eqs. 8 & 10) neither signal decision bias nor false alarm confidence increase, so this model is not prioritized. In contrast, if a reduced learning rate is combined with a pre-existing bias (i.e., initial bias), both signal decision bias and false alarm confidence increase.

Figure 4. Dissociating prioritized hallucination models. Standard tasks rarely incentivize the development of both sensory and reward expectations which may be critical for determining relevant hallucination mechanisms. In contrast, an alternative task in which the probability of a signal and the probability of reward for a signal decision are systematically varied would incentivize learning of both sensory and reward expectations and the incorporation of those expectations into responses (left). This task also enables the isolation of qualitatively distinct behavioral signatures of each model which can faciliate falsification (right).

Tables

Constraint Type	Decision-Making Model Constraints	Hallucination Model Constraints
Behavioral	**Choice and confidence are sensitive to:** A1. Reward and stimulus history A2. Choice difficulty	**Hallucinations scale with:** B1. Signal decision bias B2. False alarm confidence
Neural	**Striatal dopamine encodes:** A3. Teaching (e.g., prediction-error) signals A4. Anatomically distinct information	**Hallucinations scale with:** B3. Striatal dopamine transmission

Table 1. Empirical constraints on decision-making and hallucination models

Model	Cognitive Evidence	Link to Dopamine (B.3)		Parameter Alteration	B.1	B.2	Priority?
Sensory Noise	Altered sensory processing in SCZ[110,111]	DA can modulate cortical noise[112]		$\uparrow \sigma$	✗	✓	No
Fast Learning	Increased learning for irrelevant stimuli in SCZ[30] and in PD patients given DA agonists[113]	DA involved in signaling decision uncertainty in animals[27,52] and humans[25,26]	S	$\uparrow \alpha_S$	✗	✓	No
			R	$\uparrow \alpha_R$	✗	✓	No
Slow Learning	Decreased statistical learning in SCZ[1,6,20,21]	Decreased learning correlates with striatal DA release and hallucination level[20]	S	$\downarrow \alpha_S$	✗	✗	No
			R	$\downarrow \alpha_R$	✗	✗	No
Initial Bias	Baseline signal decision bias correlates with hallucination proneness[46,48,69] and improved detection of corrupted speech in NCVH[95]	DA agonism induces Go bias and reduces learning from positive outcomes[114]	S	$\uparrow v_S^{t=0} \downarrow \alpha_S$	✓	✓	Yes
			R	$\uparrow v_R^{t=0} \downarrow \alpha_R$	✓	✓	Yes
Asymmetric Learning	Learning asymmetries linked to striatal DA genotypes[96] and DA agonism in PD[97,98]	Models of increased tonic dopamine release can drive asymmetric learning[115]	S	$\alpha_S^{F=1} > \alpha_S^{F=0}$	✓	✓	Yes
			R	$\alpha_R^{R=1} > \alpha_R^{R=0}$	✗	✓	No
Hybrid Bias	Altered learning of stimulus-reward associations in SCZ[30] and reward schedules influence sensory processing[59,102]	DA can represent complex combinations of variables[103–105] (e.g., threatening stimuli[57])		$\uparrow \alpha^{F=1 \cap R=1}$	✓	✓	Yes

Table 2. Evaluation of potential models according to group B constraints SCZ=schizophrenia; DA = dopamine; NCVH = non-clinical voice-hearer; PD=Parkinson's disease; S=sensory; R=reward. See Figure S4 for group B constraint evaluation simulations.