eLife

Reviewed Preprint
v1 • June 6, 2025
Not revised

Reviewed Preprint
v2 • January 13, 2026
Revised by authors

For correspondence:
colin.bredenberg@mila.quebec

* indicates equal contribution

Competing interests: No competing interests declared

Funding: See page 32

Reviewing editor: Anna C Schapiro, University of Pennsylvania, United States

© 2025, Bredenberg et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

The oneirogen hypothesis: modeling the hallucinatory effects of classical psychedelics in terms of replay-dependent plasticity mechanisms

Colin Bredenberg[1,2] ✉, **Fabrice Normandin**[1], **Blake Richards**[1,3,*], **Guillaume Lajoie**[1,2,*]

[1]Mila - Quebec AI Institute, Montreal, Canada • [2]University of Montreal, Montreal, Canada • [3]McGill University, Montreal, Canada

eLife Assessment

This paper provides a **useful** new theory of the hallucinatory effects of 5-HT2A psychedelics. The authors present **convincing** evidence that a computational model trained with the Wake-Sleep algorithm can reproduce some features of hallucinations by varying the strength of top-down connections in the model, though it is not clear that this model applies to 5-HT2A hallucinogens in particular. The work will be of interest to researchers studying hallucinations or offline activity and plasticity more broadly.

https://doi.org/10.7554/eLife.105968.2.sa3

Abstract

Classical psychedelics induce complex visual hallucinations in humans, generating percepts that are coherent at a low level, but which have surreal, dream-like qualities at a high level. While there are many hypotheses as to how classical psychedelics could induce these effects, there are no concrete mechanistic models that capture the variety of observed effects in humans, while remaining consistent with the known pharmacological effects of classical psychedelics on neural circuits. In this work, we propose the "oneirogen hypothesis", which posits that the perceptual effects of classical psychedelics are a result of their pharmacological actions inducing neural activity states that truly are more similar to dream-like states. We simulate classical psychedelics' effects via manipulating neural network models trained on perceptual tasks with the Wake-Sleep algorithm. This established machine learning algorithm leverages two activity phases: a perceptual phase (wake) where sensory inputs are encoded, and a generative phase (dream) where the network internally generates activity consistent with stimulus-evoked responses. We simulate the action of psychedelics by partially shifting the model to the 'Sleep' state, which entails a greater influence of top-down connections, in line with the impact of psychedelics on apical dendrites. The effects resulting from this manipulation capture a number of experimentally observed phenomena including the emergence of hallucinations, increases in stimulus-conditioned variability, and large increases in synaptic plasticity. We further provide a number of testable predictions which could be used to validate or invalidate our oneirogen hypothesis.

Introduction

Classical psychedelics—including psilocybin, mescaline, DMT, and LSD—are a family of hallucinogenic compounds with a common mechanism of action: they are agonists for the 5-HT2a serotonin receptor commonly expressed on the apical dendrites of cortical pyramidal neurons [1] and on parvalbumin (PV) interneurons [2]. These drugs induce numerous effects in human subjects, including: complex visual, auditory, and tactile hallucinations; intense spiritual

experiences; long-lasting alterations in mood; changes in personality; and increases in synaptic plasticity [3, 4, 5]. Recently, they have been explored clinically as potential treatments for depression and anxiety [6], as well as PTSD [7].

The 5-HT2a receptor plays a critical role in psychedelic-induced hallucinations. Indeed, behavioral measures of hallucinatory drug effects are induced selectively by cellular membrane-permeable 5-HT2a agonists [8], and perceptual effects of classical psychedelics are largely eliminated by blocking 5-HT2a receptors in the cortex [9, 8] (though 5-HT2a agonists with mixed receptor selectivity are in some cases characterized by primarily non-hallucinatory effects [10, 11]). However, very little is understood about *why* highly structured hallucinations and changes in synaptic plasticity emerge from activating cortical 5-HT2a receptors: to explain this, it is necessary to develop mechanistic theories that are capable of linking changes in neuron-level properties (receptor agonism) to changes in perception and behavior. Psychedelic drug users and therapists have long noted the 'dream-like' qualities of psychedelic drug hallucinations, which are realistic but untethered from the external world; this observation leads naturally to speculation that these drugs are 'oneirogens,' or dream-manifesting compounds [12]. However, beyond perceptual phenomenology (and some evidence pointing to the effects of psychedelics on sleep cycles [13, 14, 15]), we lack a mechanistic proposal that could explain the similarity between dreams and psychedelic drug experiences. Here, we articulate the 'oneirogen hypothesis', which describes one such potential mechanistic explanation. We propose that classical psychedelics induce a dream-like state by shifting the balance between bottom-up pathways transmitting sensory information and top-down pathways ordinarily used to create replay sequences in the brain. Replay sequences have been shown to be important for learning during sleep [16, 17, 18, 19, 20]: we propose that mechanisms supporting replaydependent learning during sleep are key to explaining the increases in plasticity caused by psychedelic drug administration. In total, our model of the functional effect of psychedelics on pyramidal neurons could provide a explanation for the perceptual psychedelic experience in terms of learning mechanisms for consolidation during sleep [21], and cortical 'replay' phenomena [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32].

To explore the oneirogen hypothesis concretely, we use the aptly named Wake-Sleep algorithm [33], which has historically been used to train artificial neural networks (ANNs) that possess both a bottomup "recognition" pathway and a top-down "generative" pathway to learn a representation of incoming sensory data. It enables unsupervised learning in ANNs by alternating between periods of "waking perception" (wherein bottom-up recognition pathways drive activity) and "dreaming sequences" (wherein top-down generative pathways drive activity). With these alternate periods of distinct activity, connectivity parameters in each pathway are adjusted to match the activity of the opposite pathway. This way, the top-down pathway learns to generate activity consistent with that induced by sensory inputs, and the bottom-up pathway learns better representations thanks to generated activity.

In this work, we show that within a neural network trained via Wake-Sleep, it is possible to model the action of classical psychedelics (i.e. 5-HT2a receptor agonism) by shifting the balance during the wake state from the bottom-up pathways to the top-down pathways, thereby making the 'wake' network states more 'dream-like'. Specifically, we model the effects of classical psychedelics by manipulating the relative influence of top-down and bottom-up connections in neural networks trained with the Wake-Sleep algorithm on images. Doing so, we capture a number of effects observed in experiments on individuals under the influence of psychedelics, including: the emergence of closed-eye hallucinations, increases in stimulus-conditioned variability, and large increases in synaptic plasticity. This data suggests that the oneirogen hypothesis may indeed help to explain why 5-HT2a agonists have the functional effects that they do. We subsequently identify several testable predictions that could be used to further validate the oneirogen hypothesis.

Results

Mapping the Wake-Sleep algorithm onto cortical architecture

The Wake-Sleep algorithm allows ANNs to optimize a global, unsupervised objective function for sensory representation learning— the Evidence Lower Bound (ELBO)—through local synaptic modifications to a bottom-up recognition pathway and a top-down generative pathway. As a precursor to the variational autoencoder [34, 35], the Wake-Sleep algorithm provides a mechanism for learning a probabilistic latent representation **r** responding to incoming sensory stimuli **s**, which obeys representational characteristics that are ideal for a neural system (e.g. sparsity and metabolic efficiency [36], compression and coding efficiency [37, 38], or disentanglement [39, 40]). To do this, Wake-Sleep optimizes the ELBO through an approximation of the Expectation Maximization (EM) algorithm [41] to train the two pathways (Figure 1a) (For readers who are unfamiliar with the Wake-Sleep algorithm, a tutorial can be found here [42]).

Notably, the Wake-Sleep algorithm requires two phases of activity (i.e. "Wake" and "Sleep"), where the network phase is controlled by a global state variable $\alpha \in [0, 1]$ that regulates the balance between the bottom-up and top-down pathways. In the Wake phase ($\alpha = 0$), the network processes real sensory stimuli drawn from the environment, and network activity is sampled based on the bottom-up inputs (corresponding to the approximate inference distribution). In the Sleep phase ($\alpha = 1$), the network internally samples neural activity from its generative model, which then produces generated activity in the stimulus layer **s**. We use this structure of the Wake-Sleep algorithm as a concrete model to express oneirogen hypothesis. Specifically, we use changes to the value of α as a means of modeling a 5-HT2a agonist-induced shift to a more dream-like state, as we detail below.

Within the Wake-Sleep algorithm, neurons alternate between 'Wake' and 'Sleep' modes, where activity during each mode is dominated by the bottom-up and top-down pathways, respectively. We can determine the neural activity for a given intermediate layer l with the following equation:

$$\mathbf{r}^{(l)} = f\left(h(\mathbf{r}), \mu(\mathbf{r}), \alpha\right) + f(\sigma_b, \sigma_p, \alpha)\boldsymbol{\eta}, \qquad (1)$$

where $h(\mathbf{r})$ defines bottom-up input, $\mu(\mathbf{r})$ defines top-down input, $f(h, \mu, \alpha)$ is any interpolation function such that $f(h, \mu, 0) = h$ and $f(h, \mu, 1) = \mu$, σ_b and σ_p define the bottom-up and top-down activity standard deviations, and $\boldsymbol{\eta} \sim \mathcal{N}(0, 1)$ adds random noise to the neural activity (see Methods for more detail). Here, for notational conciseness we treat **r** as a concatenated vector of all $\mathbf{r}^{(l)}$ vectors from each layer. This equation means that α controls whether bottom-up inputs or top-down inputs control the dynamics of individual neural units.

Thus, as α moves from a value of 0 to a value of 1, the activity of the neurons shifts from being driven by the bottom-up recognition pathway to being driven by the top-down generative pathway. How could this occur in the brain? Realistically, each neuron in cortex would have its own α variable defining the relative influence of top-down and bottom-up inputs on its spiking activity; here, for simplicity, we will assign the entire network a single α value reflecting the 'mean' relative top-down/bottom-up influence averaged across neurons, as determined by the network state (Wake, Sleep, dose-dependent psychedelic administration). In the cortex, excitatory pyramidal neurons receive inputs from distinct sources: inputs that are from 'higher order' cortical areas target the apical dendrites, whereas inputs that are from 'lower order' cortical or sensory subcortical areas target the basal dendrites [43]. Thus, we can capture the core idea behind the oneirogen hypothesis using the Wake-Sleep algorithm, by postulating that the bottom-up basal synapses are predominantly driving neural activity during the Wake phase (when α is low), while top-down apical synapses are predominantly driving neural activity during the Sleep phase (when α is high; Figure 1) [44]; this is in agreement with several recent theoretical studies that have proposed that apical dendrites could serve as a site for integrating top-down learning signals [45, 46, 47, 48, 49, 50], particularly those which propose that the top-down signal corresponds to a predictive or generative model of neural activity [51, 52]. This proposed change in α does indeed appear to occur during both slow-wave (SW) [53, 54] and rapid eye movement (REM) [55, 56, 57] sleep, where apical dendritic inputs have been observed to exert increased

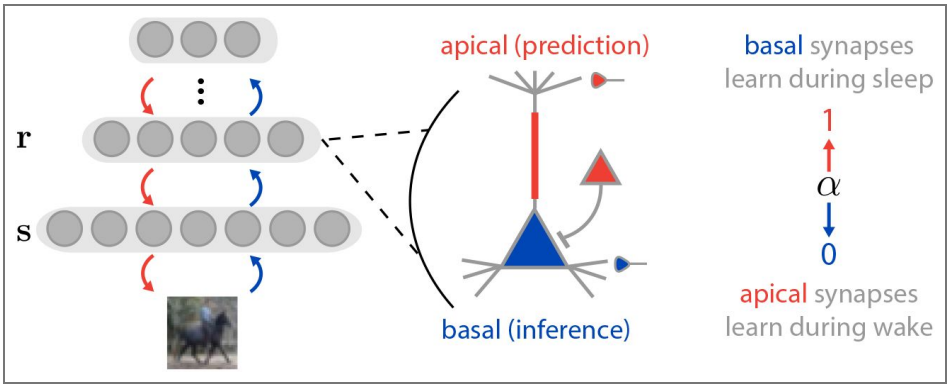


Figure 1. Mapping the Wake-Sleep algorithm onto cortical architecture.

Left: Network architecture. We model early sensory processing in the cortex with a multilayer network, **r**, receiving stimuli **s**. Center: individual pyramidal neurons receive top-down inputs (red) at the apical dendritic compartment, and bottom-up inputs at the basal dendritic compartment (blue). 5-HT2a receptors are expressed on the apical dendritic shaft (red bar), and on PV interneurons (red triangle); both sites may play a role in gating basal input. Right: Over the course of Wake-Sleep training, basal inputs dominate activity during the Wake phase ($\alpha = 0$) and are used to train apical synapses, whereas apical inputs dominate activity during the Sleep phase ($\alpha = 1$) and are used to train basal synapses.

influence on neural activity that is critical for plasticity induction and consolidation of learned behaviors; during REM sleep, this increased influence has been shown to be mediated by potentiation of basal dendrite-targeting PV inhibitory interneurons [57].

Next we ask: can we model the effects of classical psychedelics in terms of changes in α? Notably, 5-HT2a receptors are expressed in the apical dendrites of pyramidal neurons [1] and PV interneurons [2], and have an excitatory effect that positively modulates glutamatergic transmission due to apical dendritic inputs [58, 59]; further, classical psychedelic administration has been shown to have an inhibitory effect on glutamatergic transmission due to basal dendritic inputs [60]. These data suggest that 5-HT2a agonists could have a push-pull effect on cortical pyramidal neurons, increasing the relative influence of apical dendrites and decreasing the relative influence of basal dendrites [61] in much the same way as has been observed during SW and REM sleep. Hence, we can model these effects by increasing the α value in a Wake-Sleep trained network, and then ask whether the networks exhibit other phenomena that match the known impact of classical psychedelics on neural activity. We note that with this mapping of the Wake-Sleep algorithm to models of basal and apical processing, synaptic modifications at both apical and basal synapses correspond to minimizing a local prediction-error between top-down and bottom-up inputs (see Methods).

Modeling Hallucinations

To see whether a transition from waking to a more dream-like state would induce hallucinatory effects in our model, we trained multilayer neural networks with branched dendritic arbors (see Methods) on the MNIST digits dataset [62] using the Wake-Sleep algorithm and subsequently simulated hallucinatory activity by varying α (see Methods; Eq. (8)). We could visualize the effects of our simulated psychedelic with snapshots of the stimulus layer $\mathbf{s}$ at a fixed point in time for various values of α (Figure 2; see Supplemental Materials for videos). As α increased, we observed that network activity gradually deformed away from the ground-truth stimulus in a highly structured way, adding strokes to the original digit that were not originally present. At the highest values of α tested, we found that network states were wholly divorced from the ground-truth stimulus, but retained many characteristics of the MNIST digits on which the network was trained (e.g. smooth strokes and the rough form of digits). These results emphasize that hallucinations induced by a shift to a more dream-like state in these models are heavily influenced by the training dataset, which for an animal would correspond to the statistics of the sensory environment in which it learns its sensory representation. To emphasize this point, we further trained our networks on the CIFAR10 natural images dataset [63] (Figure 2c), to provide an example of a more naturalistic training dataset. In this case, our model was not powerful enough to reproduce realistic natural images—instead, we found that our modeled hallucinatory activity corresponded to 'ripple' effects, which are similar to the 'breathing' and 'rippling' phenomena reported by psychedelic drug users at low doses [3].

These simulations were produced with a complex, multicompartmental neuron model; however, we found similar results with two alternative network architectures, one with within-layer recurrence (Supplemental Figure S1a) and one which used a simpler single compartment neuron model (Supplemental Figure S1b). We found that our single compartment model produced qualitatively less realistic generated images than the multicompartment and recurrent models, justifying our use of the more complex models (Supplemental Figure S2). To demonstrate the importance of a learned top-down pathway to produce complex, structured hallucinations in the earliest layers of our network, we generated model hallucinations from two control networks: an untrained model and a trained network where psychedelic activity was alternatively modeled by a simple increase in the variance of individual neurons (we will refer to this latter control as the noise-based hallucination protocol). We found that hallucinations under these control conditions resembled additive white noise, rather than structured digit-like shapes (Supplemental Figure S1c-d).

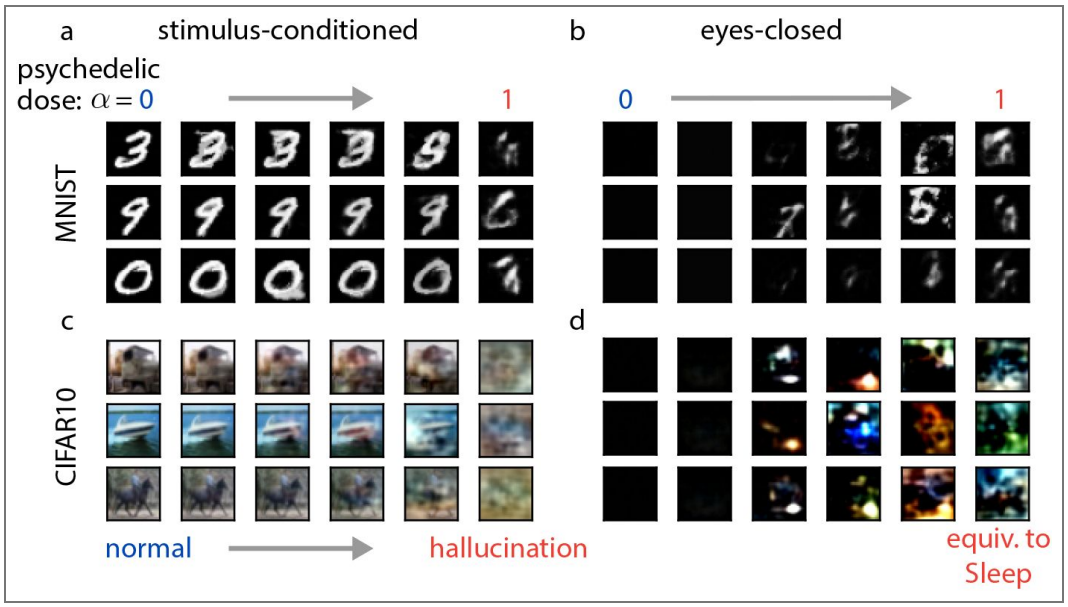


Figure 2. Visualizing the effects of psychedelics in the model.

We model the effects of classical psychedelics by progressively increasing α from 0 to 1 in our model, where $\alpha = 1$ is equivalent to the Sleep phase. We visualize the effects of psychedelics on the network representation by inspecting the stimulus layer **s. a)** Example stimulus-layer activity (rows) in response to an MNIST digit presentation as psychedelic dose increases (columns, left to right). **b)** Same as (a) but for 'eyes-closed' conditions where an entirely black image is presented. **c-d)** Same as (a-b), but for the CIFAR10 dataset.

Psychedelic drug users also report observing the emergence of hallucinations while their eyes are closed [3]. Interestingly, we found that our model recapitulated these phenomena: as α increased, networks trained on MNIST gradually began revealing increasingly complex and digit-like patterns (Figure 2b), whereas CIFAR10-trained networks again predominantly produced 'ripple' hallucinations (Figure 2d).

Effects of psychedelics on single neurons

Having recapitulated hallucinatory phenomena in stimulus space, we next explored how our proposed mechanism affected neural activity in our network model, in order to establish markers that could be used to experimentally validate or invalidate the oneirogen hypothesis. To start, we investigated the effects of learning and psychedelic drug administration on the activity of single neurons in the model. As noted previously, the learning algorithm used here trains synapses so that top-down inputs to apical dendritic compartments match bottom-up inputs to basal dendritic compartments. As a consequence, we observed that after training, inputs to apical and basal dendritic compartments were much more correlated on the same neuron than they were for random neurons (Figure 3a), which was not observed in untrained models (Supplemental Figure S3a). This form of strongly correlated tuning has been observed in both cortex and the hippocampus [64, 65].

There are many indicators that psychedelic drug administration in humans and animals can induce marked, long-lasting changes in behavior, as well as large increases in synaptic plasticity [4, 66, 67, 8, 5]. In Wake-Sleep learning, apical synapses learn during the Wake phase, whereas basal synapses learn during the Sleep phase—thus, plasticity at apical synapses is gated by $(1-\alpha)$, whereas plasticity at basal synapses is gated by α (see Methods). However, learning is still theoretically possible without this explicit gating, though it may be noisier and less efficient; furthermore, it is conceivable that classical psychedelics could increase the relative influence of apical inputs on the activity of a neuron without affecting this gating mechanism. As a consequence, we modeled the dose-dependent effects of psychedelics on plasticity both with and without gating (Figure 3b). Consistent with recent experimental results, for intermediate doses we found large increases in plasticity at both apical and basal synapses under *both* conditions, where plasticity was measured as a mean change in normalized synaptic strength across weight parameters in our network (see Methods). In our model, we found that the total evoked plasticity peaked at roughly $\alpha = 0.5$; we further found that if gating was affected by psychedelics, apical plasticity would eventually be quenched at very high drug doses. We also found that plasticity induced by psychedelic drug administration gradually became unaligned from the weight updates that would have occurred in the absence of the drug (Figure 3c), indicating that these results were not simply due to modulation of the effective learning rate of the underlying plasticity. Rather, as has been suggested by other theoretical studies [68], plasticity in the model likely increased because aberrant hallucinatory activity pulled the learning mechanism out of a local optimum in which plasticity was minimal, producing much more plasticity across the network. Importantly, we observed these increases in plasticity in all network architectures and training datasets we explored, including for our noise-based hallucination protocol (Supplemental S4), demonstrating that changes in apical dendritic influence within a WakeSleep learning framework are sufficient, but *not necessary* to induce increases in synaptic plasticity: for trained networks, it would seem that even simple increases in neural variability can have similar effects.

Effects of psychedelics on neural variability

Having observed that increasing our modeled drug dosage caused heightened fluctuations and deviations from the ground-truth stimulus in the sensory layer of our network (Figure 2), we next investigated whether variability was affected at the level of individual neurons in higher layers of the model. Indeed, we found that for a fixed stimulus, neural variability increased markedly as the simulated psychedelic drug dose increased (Figure 4a). This result is consistent with the data supporting the Entropic Brain Theory [69, 70, 71, 72], in which neural activity in resting state fMRI recordings becomes increasingly 'entropic' (i.e. variable) under the influence of

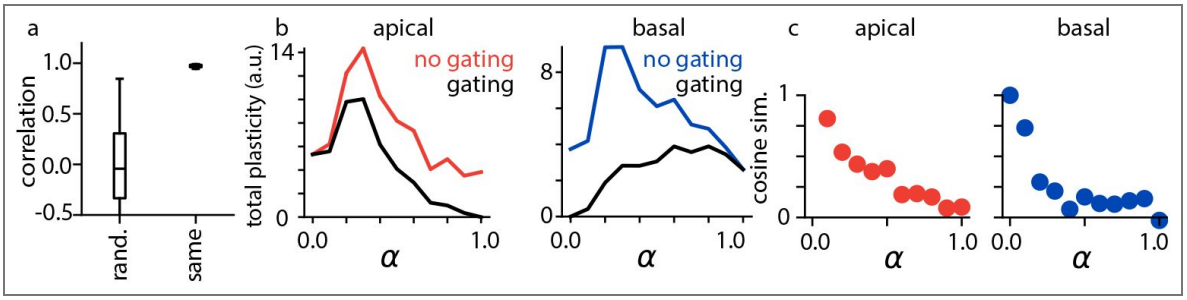


Figure 3. Effects of psychedelics on single model neurons.

a) Correlations between the apical and basal dendritic compartments of either the same network neuron or between randomly selected neurons. **b)** Total plasticity for apical (left) and basal (right) synapses as *α* increases in the model when plasticity is either gated or not gated by *α*. Error bars indicate +/-1 s.e.m. **c)** Cosine similarity between plasticity induced under psychedelic conditions compared to baseline for apical (left) and basal (right) synapses.

psychedelics; however, it is important to note that our noise-based hallucination protocol also produced these effects (Supplemental Figure S5a). Though most experimental data supporting the Entropic Brain Theory is taken from recordings with relatively poor spatial resolution, averaging activity over large cortical areas, our model predicts that this increase in variability should be reflected at the level of individual neurons; this increase in variability after psychedelic administration has been recently observed in auditory cortical neurons for active mice [73], but whether this phenomenon is general across tasks and cortical areas remains to be seen. We further found that this increase in variability corresponded to a decrease in ability to identify the stimulus being presented to the network: we trained a classifier to identify which MNIST digit was presented to our networks on Wake neural activity (see Methods), and found that the accuracy of our classifier decreased (Figure 4b) while the output variability of the classifier increased (Figure 4c) in response to drug administration.

Within our model, this increase in variability is quite sensible: in the ordinary Wake state, neural activity is constrained to correspond to the singular sensory stimulus being presented, whereas during Sleep states, neural activity is completely unconstrained by any particular sensory stimulus, reflecting instead the full distribution of *possible* sensory stimuli. As increasing α in our model interpolates between Wake and Sleep states, we can expect intermediate values of α to produce network states which are less constrained by the particular sensory stimulus being presented, reflected in increased neural variability.

Network-level effects of psychedelics

We next investigated the effects of psychedelics on networklevel and inter-areal dynamics within our model. We first identified an important negative result: the pairwise correlation structure between neurons was largely preserved across psychedelic doses (Figure 5a-b), as was the effective dimensionality of population activity (Figure 5c). This was sensible, because a network that has been well-trained with the Wake-Sleep algorithm will have the same marginal distribution of network states in the Wake mode as in the Sleep mode—thus, pairwise correlations between neurons should also not differ (as measures of the second order moments of the marginal distribution). We found empirically that even for intermediate values of α in which activity is a mixture of Wake and Sleep modes, these correlations are largely unchanged; in contrast, we observed large changes in correlation structure for untrained networks, and increases in effective dimensionality for both untrained networks and for our simple noise-based hallucination protocol, suggesting that these results are more specific to our trained models in which hallucinations are caused by an increase in apical dendritic influence (Supplemental Figure S6a-b). Interestingly, these results are consistent with a recent study that has shown only minimal functional connectivity and effective dimensionality changes in task-engaged humans being presented audiovisual stimuli under the influence of psilocybin [72].

However, though the pairwise correlations between single neurons are largely preserved, the causal influence between lower and higher layers of our model network changes considerably both during hallucination and Sleep modes. Because psychedelic drug administration increases influence of apical dendritic inputs on neural activity in our model, we found that silencing apical dendritic activity reduced acrossstimulus neural variability more as the psychedelic drug dose increases (Figure 5d). Further, we found that as α increased, inactivating the deepest network layer induced a large reduction in variability in the stimulus layer relative to baseline (Figure 5e), revealing that within our model, increases in top-down influence are responsible for much of the observed stimulus-conditioned variability at larger drug doses. These inactivations had no impact on neural variability in our noise-based hallucination protocol, but were observed for all network architectures and datasets that we tested in which hallucinations were caused by an increase in apical dendritic influence (Supplemental Figure S6), suggesting that these results are quite specific to our model. Further, these inactivations have not yet been performed in animals, and consequently constitute a critical testable prediction of our model.

Figure 4. **Effects of psychedelics on neural variability.**

a) Stimulus-conditioned variability for neurons in the network as α increases, as compared to variability in neural activity across stimuli (rightmost bar). Error bars indicate +/-1 s.e.m. **b)** Proportion correct for a classifier trained to detect the label of presented MNIST digits as α increases. **c)** Variability in the logit outputs of the trained classifier as α increases.

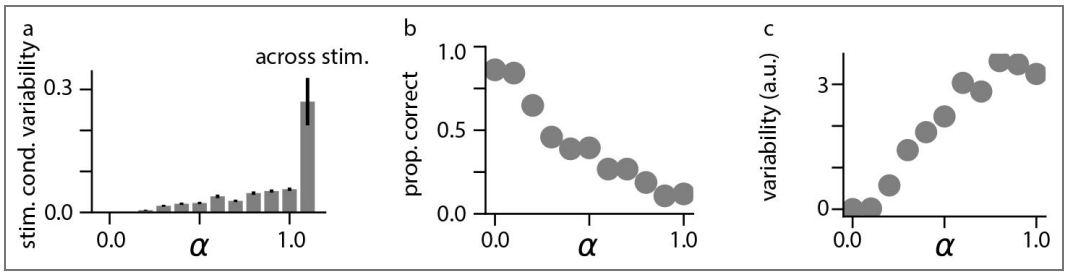
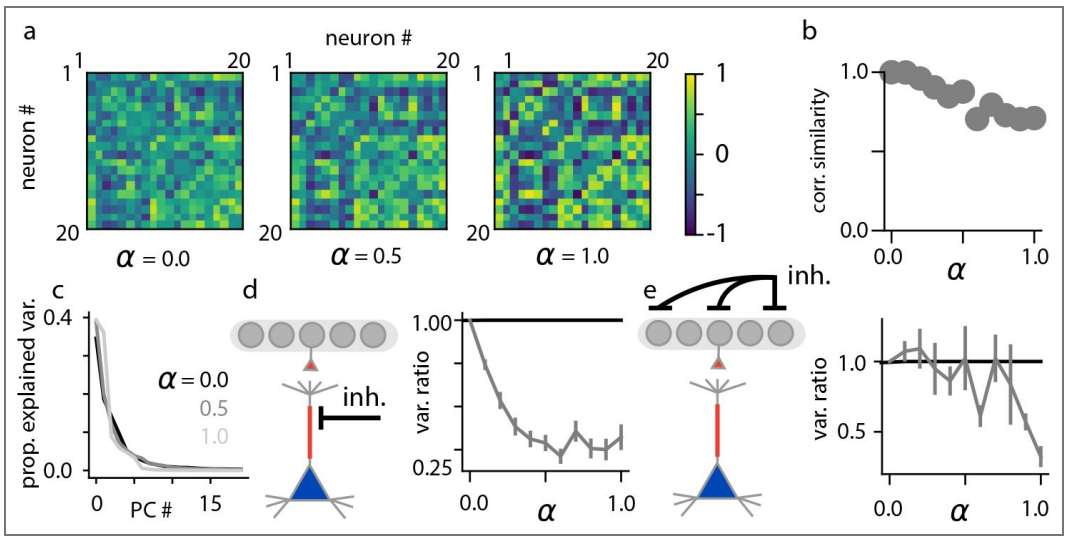


Figure 5. **Network-level effects of psychedelics.**

a) Pairwise correlation matrices computed for neurons in layer 2 across stimuli for $\alpha = 0$ (left), $\alpha = 0.5$ (center), and $\alpha = 1.0$ (right). **b)** Correlation similarity metric between the pairwise correlation matrices of the network in the absence of hallucination ($\alpha = 0$) as compared to hallucinating network states ($\alpha > 0$). **c)** Proportion of explained variability as a function of principal component (PC) number for $\alpha \in \{0, 0.5, 1\}$. **d)** Ratio of across-stimulus variance in individual stimulus layer neurons when the apical dendrites have been inactivated, versus baseline conditions across different α values. **e)** Ratio of across-stimulus variance in individual neurons in the stimulus layer when neurons at the deepest network layer have been inactivated, versus baseline conditions across different α values. Error bars indicate +/-1 s.e.m.



Modeling hallucinations in large-scale pretrained networks

While our trained model is capable of capturing several effects of classical psychedelics, it also has a clear limitation: our top-down generative model does not have sufficient expressive power to induce complex hallucinations of naturalistic stimuli, producing instead 'ripples,' or 'breathing' effects that preserve lower-order statistical features of the input data (Figure 5b). While psychedelic drug users do report these phenomena, they also report observing much more complex hallucinations, including of people, animals, and scenes [74, 75].

Generative models trained through backpropagation have been much more successful in producing more complex generated sensory stimuli [35, 34, 76], and furthermore, hierarchical variational autoencoder models have a nearly identical top-down/bottom-up model architecture as our Wake-Sleep-trained networks [77, 78]. Therefore, to see whether our proposed mechanism would induce complex, structured hallucinations in more powerful models, we induced hallucinations in Very Deep Variational Autoencoder (VDVAE) models [79] that were pretrained through backpropagation on a large natural images dataset, Tiny ImageNet [80], and a large corpus of human faces, FFHQ-256 [81]. These models have a few key differences compared to our Wake-Sleep-trained models: 1) they are trained through backpropagation, which is well-known to be biologically implausible [82]; 2) they exploit parameter sharing across spatial positions in convolutional layers for increased data efficiency during training, at the expense of further biological realism [83]; 3) the 'Wake' stage inference process of these models incorporates inputs from both bottom-up *and* top-down sources, which both improves performance [77] and is more biologically realistic [84, 43]; 4) the models are trained on more complex, higher-resolution datasets. Finally, to induce more 'abstract' hallucinations, we increased the α parameter in these models selectively for higher layers of the network, whereas for the Wake-Sleep-trained models we increased α evenly across layers (see Methods). Combined, these differences make for an effective model of high-level hallucination effects, at the expense of some biological realism.

We found that hallucinations generated by these pretrained models were much richer and more complex: increasing α in the Tiny ImageNet VDVAE caused the emergence of textural patterns and geometric shapes, while the FFHQ-256 VDVAE caused increasingly bizarre changes in facial features (Fig. 6). Both models were also capable of reproducing closed-eyes hallucinations (Fig. S7), where the content of these hallucinations was shaped by their respective training datasets.

To investigate the nature of hallucinations generated by the Tiny ImageNet VDVAE, we examined the Laplacian pyramid of decoded hallucination images at varying α values (Fig. S8a). Essentially, a Laplacian pyramid decomposes an image into levels of decreasing resolution features, with each level encoding the residual produced by downsampling to the next-lowest resolution (level 0 corresponds to the base 64x64 pixel image, while level 5 corresponds to a 4x4 reduced-resolution set of features). We found that low-level pyramid features varied considerably at low α levels, while high-level pyramid features did not begin to vary until higher α doses (Fig. S8b). This suggests that hallucinations within our model obey a fine-to-coarse structure, where low-dose hallucinations are confined to high-frequency, spatially localized changes, and progressively increasing doses begin to cause variations in more global image features.

Lastly, we were able to replicate our previous network-level results on the Tiny ImageNet VDVAE. We found that increasing psychedelic dose α caused an increase in stimulus-conditioned variance within the model (Fig. S8c), and that across-stimulus correlation structure between network units was largely preserved across doses (Fig. S8d). Further, we found that the ratio of beforeand after-inactivation across-stimulus variance decreased as the psychedelic dose α increased (though somewhat paradoxically, inactivation caused an *increase* in variance for $\alpha = 0$, likely due to the influence of top-down inputs during inference for this model). Combined, these results show that key testable predictions from our Wake-Sleep-trained model are preserved in the VDVAE, while this latter model is capable of producing some of the more complex hallucinations characteristic of psychedelic experience.

Figure 6. **Visualizing the effects of psychedelics in pretrained VDVAE models.**

Decoded outputs of a pretrained VDVAE model trained on Tiny ImageNet (Top) and FFHQ-256 (Bottom) based on hallucinations generated in the top 35 layers of the model. Image samples vary along rows, and hallucination intensity, parameterized by α, increases along columns.

Discussion

Experimental results captured by our model

In this study, we have examined a hypothetical mechanism explaining how the 5-HT2a receptor agonism of classical psychedelics could induce the highly structured hallucinations reported by people who have consumed these drugs. Specifically, we have explored the 'oneirogen hypothesis', which postulates that 5-HT2a agonists have the effects that they do because they shift the neocortex to a more dream-like state, wherein activity is more strongly driven by top-down inputs to apical dendrites than normally occurs during waking. To provide a concrete model to explore the 'oneirogen hypothesis' we used the classic Wake-Sleep algorithm, which learns by toggling between a Wake phase, where activity is driven by bottom-up sensory inputs, and a Sleep phase, where activity is driven by top-down generative signals. We modeled the 'oneirogen hypothesis' by simulating psychedelic administration as an increase in a neuronal state variable (α) that switches neural activity between these two phases, such that the simulated psychedelic caused the network to enter a state somewhere between the Wake and Sleep phases, making activity during the Wake phase less tied to actual sensory inputs by increasing the relative influence of the top-down, apical compartment in the models (depending on the "dosage"). This formulation is consistent with anatomical wiring data [43], as well as several recent theoretical studies which propose a specialized learning role for top-down projections to the apical dendrites of pyramidal neurons [45, 46, 47, 48, 49, 50]. It is also consistent with the known cellular mechanism of action of classical psychedelics [1, 58, 59, 9] and experiments that demonstrate a reduced responsivity to bottom-up stimuli in cortex after psychedelic drug administration [85, 86, 87]. Using this model, we were able to produce both stimulus-conditioned and "closed-eye" hallucinations that are consistent with the low-level effects reported by psychedelic drug users [3], and we were also able to recapitulate the large increases in plasticity observed at both apical and basal synapses at moderate psychedelic doses [4].

Our model uses a particular functional form of synaptic plasticity at both apical and basal synapses, reminiscent of the classical delta rule [88], which seeks to minimize a prediction error between inputs in apical and basal synapses. There are many theoretical models of learning that propose similar forms of plasticity [46, 47, 51], so while this plasticity is a necessary prediction of our model, it is not *sufficient* to validate it. Experimentally, plasticity dynamics which could, theoretically, minimize such a prediction error have been observed in cortex [89, 90]; we found that plasticity rules of this kind induce strong correlations between inputs to the apical and basal dendritic compartments of pyramidal neurons, which has been observed in both the hippocampus and cortex [64, 65]. Psychedelic administration within our model induced large increases in plasticity, which has also been observed experimentally [4, 5]. Within our model, this plasticity should not be interpreted as 'learning,' since it arises from aberrant network activity and does not necessarily produce behavioral or perceptual improvements; it is likely closer to 'noise,' that may still be useful for helping neural networks escape from local minima in the loss optimization landscape for synaptic weights, with possible implications for individuals suffering from post-traumatic stress disorder, early life trauma, or the negative effects of sensory deprivation. Further work will be required to analyze the relationship within our model between psychedelic dosage, usage frequency, and the long-term stability of learned representations in neural networks.

Interestingly, we also found that increasing the influence of apical dendrites in the model increased stimulus-conditioned variability in our individual neurons. In cortex, this effect has recently been shown at the level of single auditory neurons [73]; further, there have been numerous studies reporting similar increases in asynchronous variability [71] (or, analogously, sample entropy [70] and Lempel-Ziv complexity [91]) in resting-state human brain recordings, previously modeled using Entropic Brain Theory. This theory proposes that many of the effects of classical psychedelics on perception and learning can be explained in terms of increases in variability induced by drug administration (e.g. the increase in variability could introduce novel patterns of thinking, or perturb learning to allow it to break out of 'local minima'). Our results are broadly consistent with this perspective, to which we have added explanatory layers that are both

normative and mechanistic [92, 93]: namely, we speculate that this variability under ordinary conditions results from an ethologically important mechanism underlying generative replay for unsupervised learning during sleep or quiescence, and we propose that mechanistically this increase in variability is caused by the increased influence of top-down synapses that are not tied to incoming sensory stimuli. Alternatively, such entropy increases could be caused by increases in attention or selfreflective thought, as supported by recent studies showing that task engagement significantly attenuates psychedelic-induced entropy increases [72]; though our model does not include cognitive or attention components, such an interpretation is potentially consistent with and complementary to our framework.

Testable predictions

While our results are broadly consistent with existing experimental evidence, there are many unconfirmed aspects of our model which could be tested to validate or invalidate it (summarized in Table 1 ↗). As mentioned in the previous section, our model predicts that *single neurons* should increase variability in response to psychedelic drug administration in any cortical area affected by psychedelic drugs, an effect that has not yet been investigated systematically throughout cortex or across task conditions. Second, we propose that psychedelic drugs should not push network dynamics into wildly different operating regimes than normal wakefulness, beyond any differences observed between wakefulness and replay (dreams) during sleep. In particular, we found that our simulated psychedelic drug administration did not perturb pairwise correlations between neurons within local circuits when averaged across an ecologically representative set of stimuli.

Within our model, psychedelic drug administration is expected to increase the relative influence of top-down projections. This prediction appears to be supported by slice experiments [58, 59, 60], but to our knowledge this change in functional connectivity has not yet been shown via *in vivo* manipulations. This could be explored experimentally in several ways: first, we have shown that apical dendrite-targeted silencing experiments can identify the amount of influence apical dendritic inputs exert on neuronal dynamics; second, we have shown that increases in top-down influence can in principle be identified with interareal silencing experiments. We caution that interpreting results in this second vein may be difficult, as establishing a clean distinction between a 'higher order' and 'lower order' cortical area may be much more difficult in a densely recurrent system such as the brain, compared to our simplified and fully observable network model.

Interestingly, if psychedelic drugs are genuinely co-opting circuitry ordinarily reserved for generative replay during periods of offline quiescence or sleep, we would expect that the same changes in functional connectivity observed during psychedelic drug administration would also occur during periods of replay. Replay been observed and dreams have been documented during both SW [23, 25] and REM [31, 32] sleep, with REM dreams exhibiting greater degrees of bizarreness, possibly indicating a more 'generative' form of replay [96]. During SW sleep, increased top-down influence has been observed from secondary motor cortex to primary somatosensory cortex [54], and from hippocampus to prefrontal cortex [25]; however, it should be noted that increased hippocampal-to-prefrontal functional coupling was not observed after classical psychedelic administration [97]. During REM sleep, increased top-down influence (or apical dendritic influence) has been observed in prefrontal, visual [56], and motor [55] cortices, with some top-down inputs originating from higher-order thalamic nuclei [57, 98]; similarly, multiple non-invasive imaging studies have observed increases in top-down functional coupling from higher-order thalamic nuclei after psychedelic administration [99, 100]. Therefore, increases in top-down coupling appear broadly consistent between REM sleep and classical psychedelic administration, while psychedelic states appear inconsistent with the hippocampal-cortical coupling during SW sleep; this latter result could potentially be explained in terms of a recent complimentary learning systems model [101], in which SW sleep is responsible for orchestrating hippocampus-cortex-coupled *episodic* replay while REM sleep is responsible for orchestrating hippocampus-cortex-decoupled *generative* replay, but more experiments and theoretical work will likely be necessary to fully characterize this additional complexity. Given these data, it seems as

Testable Predictions	OH	EC	REBUS	DD
1. Psychedelic administration increases stimulus-conditioned variability of neurons.	✓	✓	✓	✓
2. Psychedelic administration preserves pairwise across-stimulus correlations between neurons.	✓	✗	✗	✗
3. Silencing apical dendritic compartments decreases neural variability more after psychedelic administration.	✓	n/a	n/a	n/a
4. Silencing higher-order cortical areas affects lower-order cortical activity more after psychedelic administration.	✓	✗	✗	✓
5. Psychedelic drug effects are mediated by the same circuitry responsible for inducing generative replay dynamics in cortex.	✓	✗	✗	✗

Table 1. Summarizing testable predictions of the 'oneirogen hypothesis'.

Models: OH - oneirogen hypothesis; EC - Ermentrout & Cowan [94]; REBUS - Relaxed Beliefs Under Psychedelics [69]; DD - DeepDream [95]. Key: ✓- model is consistent with the prediction; ✓-model is inconsistent with the prediction; n/a - model is neither inconsistent nor consistent with the prediction.

though REM sleep replay is a moderately stronger candidate for sharing a mechanism of action with classical psychedelics, though it remains possible that replay events during SW sleep occur via a similarly shared mechanism.

To summarize, though we have provided a candidate explanation for several of the hallucinatory effects of psychedelic drugs with a model that displays a strong correspondence with existing empirical evidence, our model rests on a number of testable assumptions. Our goal here has been to articulate these assumptions as clearly as possible, to facilitate experimental efforts to test them.

Comparisons to alternative models

Here, we review prominent existing hypotheses as to how psychedelic drugs could induce hallucinations in neural networks and compare to our model (summarized in Table 1). The first alternative proposed that incredibly complex, geometric patterns formed by DMT administration could be attributed to pattern-formation effects in visual cortex caused by a disruption of the balance between excitation and inhibition in locally-coupled topographic recurrent neural networks [94, 102]. Our work differs from this approach in several respects. First, rather than disrupting E-I balance, we propose that psychedelics increase the relative influence of apical dendrites and top-down projections on the dynamics of neural activity. Second, though their model is able to generate geometric patterns, it is not able to generate patterns that are statistically related to the features of the sensory environment (e.g. MNIST digits). Lastly, for simplicity we avoided including topographic (or convolutional) recurrent connectivity in our model; however, it would be a very fruitful direction for future research to extend our work to generative modeling of temporal video sequences, as in [103, 104]. With such a development, it is conceivable that our model could directly generalize these pattern formation-based approaches.

Perhaps more closely related to our model is the 'relaxed beliefs under psychedelics' (REBUS) model, which proposes to explain the effects of classical psychedelics in terms of predictive coding theory [69]. Similar to the Wake-Sleep algorithm, predictive coding theory [105] models sensory representation learning with neural dynamics and local synaptic modifications that collectively optimize an ELBO objective function. However, at a mechanistic level, there are numerous differences, the most easily distinguishable feature being that the Wake-Sleep algorithm *requires* periods of offline 'generative replay' to train bottom-up synapses in its network, whereas predictive coding learning occurs concomitantly with stimulus presentation. Furthermore, the REBUS model of psychedelic effects is described at a computational level, in terms of a decrease in the 'precision-weighting of top-down priors.' While it is more difficult to map the REBUS model directly onto cortical microcircuitry, and the hallucinatory effects of such a model have, to our knowledge, not been directly analyzed, it has been shown that the proposed mechanism causes an *increase* in bottom-up information flow between cortical areas [106], in direct contrast to the effects that we have shown in our model (Figure 5c-d); there is some evidence supporting this idea [107], but noninvasive imaging studies are inconsistent on this question, with many studies showing by contrast an increase in top-down functional connectivity caused by classical psychedelic administration [99, 100], and with invasive recordings showing a decrease in the influence of bottom-up inputs [85, 86, 87]. Because interareal causal influence can be difficult to analyze statistically due to dense recurrent connectivity (i.e. correlation does not imply causation), we stress that it would be more effective to distinguish between the REBUS model and our 'oneirogen hypothesis' by performing direct interventions on inputs to the apical and basal dendritic compartments of pyramidal neurons in cortex, and by exploring whether psychedelic drugs affect the same circuitry that induces 'generative replay' during periods of sleep and quiescence. More consistent with our model, a recent non-mechanistic approach based on the DeepDream algorithm has been used to generate realistic hallucinations via increased influence from a top-down learning signal [95]; however, this model proposes no relationship between psychedelics and replay during sleep.

Lastly, it should be noted that the Wake-Sleep algorithm and our choice of network architecture constitute one particular model within a family of related models, all of which satisfy our key criteria for a good model of the 'oneirogen hypothesis,' namely that 1) the model has well-defined top-down and bottom-up pathways, 2) it learns a generative model of incoming sensory inputs, and 3) it uses periods of offline replay for learning through local synaptic plasticity. For example, in the Supplemental Materials, we have replicated all of our essential results for two alternative network architectures, also learned via the Wake-Sleep algorithm: one model uses within-layer recurrence to improve generative performance, while the other model uses a simpler single compartment neuron model. Furthermore, the closely related Contrastive Divergence learning algorithm for Boltzmann Machines [108] also involves alternations between Wake and generative Sleep phases, learns through local synaptic plasticity, and has been used to model hallucination disorders like Charles Bonnet Syndrome [109], though Boltzmann machines are computationally more cumbersome to train and require more non-biological network features than the Wake-Sleep algorithm. We feel as though it is important to recognize that models that satisfy these three criteria are more similar than they are different, and that it may be quite difficult to experimentally distinguish between them.

Limitations

While our model is capable of capturing several effects of classical psychedelics, it also has several clear limitations. Firstly, while we have been able to model complex hallucination phenomena with backpropagation-trained networks, hallucinations generated by Wake-Sleep-trained networks were generally simpler, likely because the Wake-Sleep algorithm is well-known to be a less effective representation learning and generative modeling algorithm than backpropagation [35], despite its superior biological realism. This suggests that while it is quite possible for generative modeling approaches to produce complex hallucinations through non-biological means, algorithmic or architectural improvements may be necessary in order to make the performance of the more plausible Wake-Sleep algorithm closer to that achieved by state-of-the-art models.

Our model also oversimplifies several aspects of biology. In particular, we do not use neurons that respect Dale's law [110, 111], and the majority of our efforts to map the Wake-Sleep algorithm onto biology focus on excitatory pyramidal neurons. Furthermore, though we do observe that neural dynamics can tolerate a significant amount of top-down input before disrupting perception, experiments and theoretical studies have shown that inputs to apical dendrites of pyramidal neurons do play an important role in waking perception [43, 98, 112], and are not *just* learning signals. We focused on clear distinctions between basally-driven Wake modes and apically-driven Sleep modes during training for computational efficiency reasons, and also due to the fact that parameter sharing across inference and generative networks in the Wake-Sleep algorithm is theoretically under-explored (though it is supported in closely related predictive coding approaches [105] and Boltzmann machines [108]). Future elaborations on our model could incorporate feedback control [113], attention [114], or multimodal sensory inputs [115] into top-down projections; such inputs could help explore how psychedelic hallucinations interact with attentional or feedback control systems in the brain, and have been shown to interact constructively with top-down learning signals in prior models [116, 117, 118]. Our use of VDVAEs is a positive step in this direction, but ideally such network architectures would be made compatible with the Wake-Sleep algorithm.

Lastly, our modeling focus has been exclusively on cortical plasticity and hallucination effects: it should be noted that our model has little bearing on other important features of the psychedelic experience of potential therapeutic relevance, because we have not included the effects of psychedelics on subcortical structures including the serotonergic system [119], which plays an important role in regulating mood and may be where psychedelics exert some of their antidepressant effects. Many studies of the effects of psychedelics on fear extinction focus on the hippocampus or the amygdala [120, 121, 122, 123]. These areas receive extensive innervation directly from serotonergic synapses originating from the dorsal raphe nucleus, which have been

shown to play an important role in emotional learning [124]; because classical psychedelics may play a more direct role in modulating this serotonergic innervation, it is possible that fear conditioning results (in addition to the anxiolytic effects of psychedelics) cannot be attributed to a shift in balance between apical and basal synapses induced by psychedelic administration.

Conclusions

Here we have proposed a hypothesis for the mechanism of action of psychedelic drugs in terms of its excitatory effects on the apical dendrites of pyramidal neurons, which we propose pushes network dynamics into a state normally reserved for offline replay and learning; we have also proposed a number of testable predictions which could be used to validate or invalidate our hypothesis. If validated, our model would describe a mechanism by which psychedelic drug administration causes ordinary sensory perception to become literally more dream-like; it further suggests that the plasticity increases observed during both sleep and psychedelic experience could occur via a common mechanism dedicated to sensory representation learning in the brain. Beyond classical psychedelics, further studying the balance between apical and basal dendritic inputs to pyramidal neurons in connection to replay during sleep may be relevant for explaining the hallucinatory effects of other drugs (such as ketamine) or mental disorders like schizophrenia [125].

Methods

Model architecture and training

To model the effects of psychedelics on neural network dynamics and plasticity, we first constructed a simple model of the early visual system by training neural networks on two different image datasets (MNIST [62] and CIFAR10 [63]). Networks were trained with the Wake-Sleep algorithm [33], which requires, for each layer, two modes of stochastic network activity: a 'generative mode', and an 'inference mode'. For the 'inference' mode we must specify a probability distribution $b(\mathbf{r}^{(l)} \mid \mathbf{r}^{(l-1)})$, while for the 'generative' mode we must specify a separate distribution $p(\mathbf{r}^{(l)} \mid \mathbf{r}^{(l+1)})$ (As a notational convention, we will use letters when referring to mathematical objects from the generative, top-down distribution, and their vertical reflection when referring to the inference, bottom-up distribution (e.g. p and b)). Notice here that activity in 'inference' mode is conditioned on 'bottom-up' network states ($\mathbf{r}^{(l-1)}$), while activity in generative mode is conditioned on 'top-down' network states ($\mathbf{r}^{(l+1)}$) (Figure 1a).

The 'inference mode' specifies a probability distribution over neural activity, conditioned on the nextlower layer (where the lowest layer is the stimulus layer, i.e. $\mathbf{r}^{(0)} = \mathbf{s}$)—mechanistically it corresponds to activity generated by feedforward projections. To increase the expressive power of our neural units, we use multicompartmental neuron models similar to [126] with N_d dendritic compartments, whose voltages are summed nonlinearly to form the full input to the basal dendrites. For $l > 0$, layer activity is sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(h(\mathbf{r}^{(l-1)}), \sigma_b^2 \mathbf{I})$, where for neuron i in layer l, $h_i(\mathbf{r}^{(l-1)})$ is given by:

$$h_i(\mathbf{r}^{(l-1)}) = \phi\left(\sum_{n=0}^{N_d} w_{in}^{(l)} \phi_d\left(\mathbf{W}_{in}^{(l)} \mathbf{r}^{(l-1)} + c_{in}^{(l)}\right) + b_i^{(l)}\right), \qquad (2)$$

where $\mathbf{W}_{in}^{(l)}$ is a $1 \times N^{(l-1)}$ matrix of synaptic weights onto dendrite n, c_{in} is the corresponding bias for the nth dendritic compartment, $w_{in}^{(l)}$ is the strictly positive weight given to the nth dendritic branch (roughly corresponding to a conductance), and $b_i^{(l)}$ is the bias for the entire basal compartment. $\Phi_d(\cdot)$ and $\Phi(\cdot)$ are nonlinearities for the dendritic branches and the total basal compartment, respectively: both are the sequential composition of the tanh nonlinearity, followed by batch normalization [127]. For the dendritic branch nonlinearities, we allow for learnable affine parameters (scale and bias), but for the entire basal dendritic compartment we constrain activity to be zero-mean and unit variance across batches in order to prevent indeterminacy between apical and basal scale parameters. For the final inference layer $\mathbf{r}^{(L)}$, as in the variational autoencoder [34], we parameterize both the mean *and* a diagonal covariance matrix of the

inference distribution: $\mathbf{r}^{(L)} \sim \mathcal{N}h(\mathbf{r}^{(L-1)})$, diag($h_2(\mathbf{r}^{(L-1)})$), where $h_2(\cdot)$ is also a multicompartmental model, in this case replacing the final batch normalization with an exponential nonlinearity to ensure positivity.

The 'generative' mode specifies a probability distribution over neural activity, conditioned on the next-higher layer—it corresponds mechanistically to activity generated by feedback projections. The highest layer, $\mathbf{r}^{(L)}$ is sampled from an $N^{(L)}$-dimensional independent standard normal distribution, $\mathbf{r}^{(L)} \sim \mathcal{N}(0, \mathbf{I})$, and all subsequent layers are sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(\mu(\mathbf{r}^{(l+1)}), \sigma_p^2 \mathbf{I})$, where for the ith neuron, $\mu_i(\mathbf{r}^{(l+1)})$ is given by:

$$\mu_i(\mathbf{r}^{(l+1)}) = \phi\left(\sum_{n=0}^{N_d} m_{in}^{(l)} \phi_d\left(\mathbf{M}_{in}^{(l)} \mathbf{r}^{(l+1)} + d_{in}^{(l)}\right) + a_i^{(l)}\right), \tag{3}$$

where $\mathbf{M}_{in}^{(l)}$ is a $1 \times N^{(l+1)}$ matrix of synaptic weights onto apical dendritic branch n, $d_{in}^{(l)}$ is the corresponding bias for the nth dendritic compartment, $m_{in}^{(l)}$ is the strictly positive weight given to the nth dendritic branch, and $a_i^{(l)}$ is the bias for the entire apical compartment. Again, $\Phi_d(\cdot)$ and $\Phi(\cdot)$ are nonlinearities, identical to the inference (basal) pathway.

While the neuron model used here is more complicated than is normally used for single-unit neuron models, functions of this kind could feasibly be implemented by nonlinear dendritic computations [126]; we further found that using this nonlinearity qualitatively improved generative performance (Supplemental Figure S2). Given these parameterized probability distributions, we then determined the neural activity for each layer l according to Eq. 1. Our network trained on MNIST was composed of 3 layers, with widths [32, 16, 6], listed in ascending order. A full list of network hyperparameters for both our MNIST and CIFAR10-trained networks can be found in the Supplemental Methods.

All synaptic weights and parameters in our networks were trained via the Wake-Sleep algorithm [33], which is known to produce 'local' parameter updates for a wide range of neuron models (and rate or spike-based output distributions), though the specific functional form of the update may vary depending on the neuron model chosen [128]. These updates, for reasonable choices of neural network architecture, can be interpreted as predictions for how synaptic plasticity should look in the brain, if learning were really occurring via the Wake-Sleep algorithm or some approximation thereof.

Consider a generic inference (basal dendrite) parameter for neuron i, $\theta_b \in \{w_{in}^{(l)}, \mathbf{W}_{in}^{(l)}, b_i^{(l)}, c_{in}^{(l)} : n = 0, ..., N_d\}$. The Wake-Sleep algorithm gives the following update, for a single stimulus presentation:

$$\Delta\theta_b = (\alpha)\eta \frac{\left(\mathbf{r}_i^{(l)} - h_i(\mathbf{r}^{(l-1)}, \theta_b)\right)}{\sigma_b^2} \frac{\partial h_i(\mathbf{r}^{(l-1)}, \theta_b)}{\partial \theta_b}, \tag{4}$$

where η is a learning rate, and the gate α ensures that learning only occurs during sleep mode. Further, for reasons of computational efficiency, we average weight updates across a batch of 512 stimulus presentations; similar results could in principal be obtained with purely online updates [129], but we opted to present stimuli in batches here in order to parallelize computations. $\frac{\partial h_i(\mathbf{r}^{(l-1)}, \theta_b)}{\partial \theta_b}$ changes depending on the parameter θ, reflecting that particular parameter's contribution to basal dendritic activity. For a dendritic branch weight $w_{in}^{(l)}$, we have:

$$\frac{\partial h_i(\mathbf{r}^{(l-1)}, w_{in}^{(l)})}{\partial w_{in}^{(l)}} = \phi'(\mathbf{v}_i^{total}) \phi_d(\mathbf{v}_{in}), \tag{5}$$

where $\mathbf{v}_i^{total}$ is the total input to the basal dendritic compartment, and $\mathbf{v}_{in} = \mathbf{W}_{in}^{(l)} \mathbf{r}^{(l-1)} + c_{in}^{(l)}$ is the total input to the nth dendritic branch. This update has the functional form of a classical 'delta' learning rule [88], where a compartmental prediction error between local dendritic activity and neuronal firing rate is multiplicatively combined with branch-specific input to provide changes in the conductance for the nth branch. Similarly, for the jth synapse on the nth dendritic branch, $\mathbf{W}_{inj}^{(l)}$, we have:

$$\frac{\partial h_i\left(\mathbf{r}^{(l-1)}, \mathbf{W}_{inj}^{(l)}\right)}{\partial \mathbf{W}_{inj}^{(l)}} = \phi'(\mathbf{v}_i^{total}) w_{in}^{(l)} \phi'(\mathbf{v}_{in}) \mathbf{r}_j^{(l-1)}. \quad (6)$$

Unlike for simple one-compartment neuron models, computation of parameter updates for dendritic synapses $\mathbf{W}_{inj}^{(l)}$ requires weighting the 'delta' error by the conductance of the corresponding dendritic branch (w_{in}), which could be approximated by the passive diffusion of signaling molecules from the principal basal dendritic compartment back along dendritic branches to individual synapses.

For generative parameters $(\theta_p \in \{m_{in}^{(l)}, \mathbf{M}_{in}^{(l)}, a_i^{(l)}, d_{in}^{(l)} : n = 0, ..., N_d\})$, we have a nearly identical update for a single stimulus presentation:

$$\Delta\theta_p = (1-\alpha)\eta \frac{\left(\mathbf{r}_i^{(l)} - \mu_i(\mathbf{r}^{(l+1)}, \theta_p)\right)}{\sigma_p^2} \frac{\partial \mu_i(\mathbf{r}^{(l+1)}, \theta_p)}{\partial \theta_p}, \quad (7)$$

where now input in the apical dendritic compartment, $\mu_i(\mathbf{r}^{(l+1)})$, is being compared to the activity of the neuron as a whole to determine the magnitude and sign of plasticity. The $(1 - \alpha)$ gate in this case ensures that plasticity only occurs during the Wake mode. We provide pseudocode (Algorithm 1) for our Wake-Sleep implementation, as well as a full list of algorithm and optimizer hyperparameters (Tables S1 and S2) in the Supplemental Materials (Code for reproducing all results from Wake-Sleep-trained models this study is available here: *https://github.com/colinbredenberg/oneirogen-hypothesis*).

Modeling Hallucinations

During training, neural network activity is either dominated entirely by bottom-up inputs (Wake, $\alpha = 0$) or by top-down inputs (Sleep, $\alpha = 1$). As a consequence, sampling neural activity is computationally low-cost, and can be performed in a single time step. During Wake, one can take a sampled stimulus variable **s**, determine the activity at layer 1, then 2, and so on until layer L, while during Sleep, one can sample a latent network state in layer L and traverse the layers in reverse order, down to the stimulus layer. However, this is not possible if $\alpha \notin \{0, 1\}$, because activity in each layer l should depend simultaneously on layer $l + 1$ and layer $l-1$. For this reason, we chose to model hallucinatory neural activity *dynamically*, as follows:

$$\mathbf{r}_t^{(l)} = (1 - \frac{1}{\tau})\mathbf{r}_{t-1}^{(l)} + \frac{1}{\tau} f\left(h(\mathbf{r}_{t-1}), \mu(\mathbf{r}_{t-1}), \alpha\right) + \frac{f(\sigma_b, \sigma_p, \alpha)}{\sqrt{\tau}} \boldsymbol{\eta}_{t-1}, \quad (8)$$

where τ is a time constant that determines how much of the previous network state is retained, and $\boldsymbol{\eta} t-1 \sim \mathcal{N}(0, \mathbf{I})$. Critically, if we take $\tau = 1$ these dynamics reduce to the sampling procedure used during training (Eq. 1). *A priori*, the choice of interpolation function $f(a, b, \alpha)$ is arbitrary. We selected the following function:

$$f(a, b, \alpha) = \kappa \log\left[(1-\alpha)\exp\frac{a}{\kappa} + \alpha \exp\frac{b}{\kappa}\right], \quad (9)$$

where $\kappa = 0.35$ is a free parameter. This function is equivalent to linear interpolation as $\kappa \to \infty$, and is equivalent to the maximum function between arguments a and b as $\kappa \to 0$ if $\alpha = 0.5$. By selecting $\kappa = 0.35$, we are biasing the system towards registering positive inputs from apical *or* basal sources (in the inclusive sense). We found that this produced 'hallucinatory' percepts in stimulus space that did not reduce the intensity of input stimuli as α increased; rather, inputs maintained their intensity and hallucinations were added on top if they were of greater intensity than the ground-truth image. All simulations were run for 800 timesteps, with $\tau = 0.1$. As a control, we compared our results to network dynamics produced purely by increases in noise, without increases in apical dendritic influence (which we refer to as our noise-based hallucination protocol). For these control simulations, we produced network activity time series with the following equation:

$$\mathbf{r}_t^{(l)} = (1 - \frac{1}{\tau})\mathbf{r}_{t-1}^{(l)} + \frac{1}{\tau} h(\mathbf{r}_{t-1}) + \frac{\sigma_b + \alpha}{\sqrt{\tau}} \boldsymbol{\eta}_{t-1}, \quad (10)$$

so that the standard deviation of the injected noise increased linearly with α.

Apical and Basal Alignment

To measure the alignment between inputs in the apical and basal dendritic compartments of our model neurons, we computed the 'Wake' neural responses to the full test dataset and measured the activity in both the basal and apical compartments of our neurons ($h(\mathbf{r}^{(l-1)})$ and $\mu(\mathbf{r}^{(l+1)})$, respectively). We then calculated the correlation coefficient between apical and basal compartments for the same neuron, compared to the correlation between compartments for two randomly selected neurons.

Quantifying plasticity

To quantify the total amount of plasticity induced in our model system by the administration of psychedelic drugs, we measured the change in *relative* parameter strength (averaging across all synapses in the network and an ensemble of 512 test images). For each test image, we simulated network dynamics according to Eq. 8. Subsequently, for each parameter θ, we calculated the net amount of plasticity induced by viewing all test images, $\Delta\theta$. We subsequently reported the relative change:

$$\Delta\theta_{rel} = \frac{|\Delta\theta|}{|\theta| + \epsilon}, \tag{11}$$

under conditions in which α values gate plasticity (as in ordinary Wake-Sleep) and under conditions in which psychedelic drug administration does not also affect plasticity gating. Here, we took $\epsilon = 10^{-2}$ to avoid numerical instabilities.

Classifier training

As we trained our neural network using the Wake-Sleep algorithm, we simultaneously trained a *separate* classifier network based on Wake-phase neural activity in the second network layer on a cross-entropy loss, to identify the stimulus class of the input to the system. For our classifier, we used a multilayer perceptron neural network with a single 256-unit hidden layer and tanh(·) nonlinearities.

We then quantified the accuracy of the classifier on the test set, based on neural activity drawn from the final time step T of hallucination simulations with various values of α. We further measured the average variance of the 10-dimensional output logits of the neural network.

Quantifying correlation matrix similarity before and after psychedelics

To quantify how similar the pairwise correlations between neurons in our model networks were before and after the administration of psychedelics, we recorded hallucinatory network dynamics for an ensemble of 512 test images, and measured pairwise correlations between neurons in the first network layer. To compare these matrices, we then report the correlation coefficient between the flattened $N \times N$ matrices. For this metric, a value of 1 indicates that the correlation matrices are perfectly aligned, while a value of -1 indicates that pairwise correlations are fully inverted.

Quantifying interareal causality through inactivations

To quantify changes in interareal functional connectivity induced by psychedelics, we performed two different types of inactivation. In the first, we inactivated the apical dendritic compartments of all neurons in the stimulus layer, and measured how this inactivation affected across-stimulus variability of neurons relative to the fully active state. In the second method, we inactivated all neurons in the deepest layer, and measured the same effect in across-stimulus variability in the stimulus layer. For both inactivation schemes, we report the mean and standard error of the variance ratio:

$$VR = \frac{\mathrm{Var}_{inact}(\mathbf{r}^{(0)}) + \epsilon_v}{\mathrm{Var}(\mathbf{r}^{(0)}) + \epsilon_v}, \tag{12}$$

where we added $\epsilon_v = 10^{-3}$ to the denominator to prevent numerical instability and to the numerator ensure that the ratio evaluates to 1 if the two variances are equivalent.

Generating hallucinations in hierarchical variational autoencoders

To model more complex hallucination phenomena than could be observed in our simpler Wake-Sleep-trained networks, we used pretrained VDVAE [79] models trained on Tiny ImageNet [80], a 64x64 pixel variant of ImageNet, and FFHQ-256 [81], a dataset of 256x256 pixel human faces. VDVAE models are very similar to our Wake-Sleep-trained models: they are trained on the same unsupervised representation learning objective function (the ELBO), and every layer of the multilayer network models are parameterized by a bottom-up inference distribution b and a top-down generative distribution p. VDVAE models are top-down VAEs [77], which means that the inference distribution is conditioned on bottom-up stimuli *and* latent network activity at higher layers, i.e. the distribution is written $b(\mathbf{r}^{(l)} \mid h(\mathbf{s}, \mathbf{r}^{(l+1)}))$, where $h(\cdot)$ is a parameterized neural network. By contrast the generative distribution is conditioned only on top-down inputs, and is written $p(\mathbf{r}^{(l)} \mid \mu(\mathbf{r}^{(l+1)}))$, where $\mu(\mathbf{r}^{(l+1)})$ is also a parameterized neural network.

For our Wake-Sleep-trained networks, we modeled hallucinations by simulating a stochastic time series at each layer (Eq. 8), but for the VDVAE models we found this to be computationally infeasible. Instead, we modeled hallucinations with a single bottom-up and top-down pass through the network, as follows:

$$\mathbf{r}^{(l)} = (1-\alpha)\mathbf{r}_b^{(l)} + (\alpha)\mathbf{r}_p^{(l)}, \tag{13}$$

where $\mathbf{r}_b^l \sim b(\mathbf{r}^{(l)}|h(\mathbf{s}, \mathbf{r}^{(l+1)}))$ is a sample from the inference distribution, and $\mathbf{r}_p^l \sim p(\mathbf{r}^{(l)}|\mu(\mathbf{r}^{(l+1)}))$ is a sample from the generative distribution. This generation scheme is simpler and less computationally expensive than our previous method, while still producing purely Wake-stage sampling when $\alpha = 0$ and Sleep-stage sampling when $\alpha = 1$; intermediate values of α correspond to modeled hallucinatory network states (Code for reproducing results obtained with pretrained VDVAE models is available here: *https://github.com/colinbredenberg/vdvae*). Our Laplacian pyramid analysis of generated images was performed using the Pyrtools package [130].

Supplementary materials

Recurrent network model

To explore the extent to which our results hold for different neuron models, and to give our generative model more expressive power than the traditional Helmholtz machine [137], we constructed a network model with a single timestep of within-layer recurrent denoising in each layer, which gives our model some similarities to denoising diffusion approaches [138]. For both our 'inference' mode and our 'generative' mode we specify both a denoised network state $\bar{\mathbf{r}}^{(l)}$ and a noise-corrupted network state $\mathbf{r}^{(l)}$ for layer l; specifying a neural network model is then equivalent to specifying, for each layer, a joint probability distribution over denoised and noise-corrupted network states for both the inference and generative modes, i.e. for the 'inference' mode we must specify a probability distribution $b(\bar{\mathbf{r}}^{(l)}, \mathbf{r}^{(l)} \mid \mathbf{r}^{(l-1)})$, while for the 'generative' mode we must specify a separate distribution $p(\bar{\mathbf{r}}^{(l)}, \mathbf{r}^{(l)} | \bar{\mathbf{r}}^{(l+1)})$ (As a notational convention, we will use letters when referring to mathematical objects from the generative, top-down distribution, and their vertical reflection when referring to the inference, bottom-up distribution (e.g. p and b)). Notice here that activity in 'inference' mode is conditioned on 'bottomup' network states ($\mathbf{r}^{(l-1)}$), while activity in generative mode is conditioned on 'top-down' network states ($\bar{\mathbf{r}}^{(l+1)}$) (Figure 1a).

The 'inference mode' specifies a probability distribution over neural activity, conditioned on the nextlower layer (where the lowest layer is the stimulus layer, i.e. $\mathbf{r}^{(0)} = \mathbf{s}$)—mechanistically it corresponds to activity generated by feedforward projections. For $l > 0$, layer activity is sampled from the distribution $\bar{\mathbf{r}}^{(l)} \sim \mathcal{N}(h(\mathbf{r}^{(l-1)}), \sigma_b^2 \mathbf{I})$, where $h(\mathbf{r}^{(l-1)})$ is given by:

$$h(\mathbf{r}^{(l-1)}) = \tanh(\mathbf{W}^{(l)}\mathbf{r}^{(l-1)} + b^{(l)}). \quad (14)$$

Subsequently, we add additional noise to get a noise-corrupted network state $\mathbf{r}^{(l)} \sim \mathcal{N}((1-\bar{\sigma}_b^2\mathbf{I})\bar{\mathbf{r}}^{(l)}, \bar{\sigma}_b^2)$; while noise corruption is a natural feature of network dynamics in the brain [139], we include it here in our model because it has been shown that *denoising* is a critical aspect of many powerful generative modeling approaches [140, 141, 142], and we have likewise found that it improves the quality of generated images in our learned networks (Supplemental Figure S2).

The 'generative' mode specifies a probability distribution over neural activity, conditioned on the next-higher layer—it corresponds mechanistically to activity generated by feedback projections. The highest layer, $\mathbf{r}^{(L)}$ is sampled from an $N^{(L)}$-dimensional independent standard normal distribution, $\mathbf{r}^{(L)} \sim \mathcal{N}(0, \mathbf{I})$, and all subsequent layers are sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(\mu(\bar{\mathbf{r}}^{(l+1)}), \sigma_p^2\mathbf{I})$, where $\mu(\bar{\mathbf{r}}^{(l+1)})$ is given by:

$$\mu(\bar{\mathbf{r}}^{(l+1)}) = \tanh(\mathbf{M}^{(l)}\bar{\mathbf{r}}^{(l+1)} + a^{(l)}), \quad (15)$$

where $m^{(l)}$ is a $N^{l} \times N^{(l+1)}$ weight matrix, and a is a bias term. Subsequently, the network goes through a single timestep of recurrent denoising, so that $\bar{\mathbf{r}}^{(l)} \sim \mathcal{N}(\bar{\mu}(\mathbf{r}^{(l)}), \bar{\sigma}_p^2\mathbf{I})$, where $\bar{\mu}(\mathbf{r}^{(l)})$ is given by:

$$\bar{\mu}(\mathbf{r}^{(l)}) = \mathbf{r}^{(l)} + \sigma(\mathbf{C}_1\mathbf{r}^{(l)} + \mathbf{c}_1)\tanh(\mathbf{C}_2\mathbf{r}^{(l)} + \mathbf{c}_2), \quad (16)$$

where $\sigma(\cdot)$ is a sigmoid nonlinearity that acts as a gating function similar those used in the LSTM [143] and GRU [144], $\mathbf{C}_1$ and $\mathbf{C}_2$ are $N^{(l)} \times N^{(l)}$ recurrent weight matrices, and $\mathbf{c}_1$ and $\mathbf{c}_2$ are biases. While this is a more complicated nonlinearity than is normally used for single-unit neuron models, functions of this kind could feasibly be implemented by nonlinear dendritic computations [126]; we further found that using this nonlinearity qualitatively improved generative performance. Given these parameterized probability distributions, we then determined the neural activity for each layer l according to Eq. (1). As with our multicompartmental neuron model, inference and generative parameters were updated according to Eqs. (4) and (7), respectively. Recurrent network hyperparameters are available in Table S3.

Simplified neuron model

As a control, we also tested our results using a simplified multilayer perceptron neuron model, which used neither batch normalization nor multiple dendritic branches. For the 'inference' mode within the simplified model, for $l > 0$, layer activity is sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(h(\mathbf{r}^{(l-1)}), \sigma_b^2\mathbf{I})$, where for neuron i in layer l, $h_i(\mathbf{r}^{(l-1)})$ is given by:

$$h_i(\mathbf{r}^{(l-1)}) = \tanh\left(\mathbf{W}_{i:}^{(l)}\mathbf{r}^{(l-1)} + b_i^{(l)}\right), \quad (17)$$

where $\mathbf{W}_{i:}^{(l)}$ is a $1 \times N^{(l-1)}$ matrix of basal synaptic weights onto neuron i, and b_i is the corresponding bias.

The simplified 'generative' mode likewise replaces the branched neuron model used in the main text with a multilayer perceptron model. The highest layer, $\mathbf{r}^{(L)}$ is sampled from an $N^{(L)}$-dimensional independent standard normal distribution, $\mathbf{r}^{(L)} \sim \mathcal{N}(0, \mathbf{I})$, and all subsequent layers are sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(\mu(\mathbf{r}^{(l+1)}), \sigma_p^2\mathbf{I})$, where for the ith neuron, $\mu_i(\mathbf{r}^{(l+1)})$ is given by:

$$\mu_i(\mathbf{r}^{(l+1)}) = \tanh\left(\mathbf{M}_{i:}^{(l)}\mathbf{r}^{(l+1)} + a_i^{(l)}\right), \quad (18)$$

where $\mathbf{M}_{i:}^{(l)}$ is a $1 \times N^{(l+1)}$ matrix of apical synaptic weights onto neuron i, and $a_i^{(l)}$ is the corresponding bias. As with the branched neuron model, inference and generative parameters were updated according to Equations (4) and (7), respectively. For optimization, we used the identical hyperparameters to the multicompartment neuron model (Table S1).

Algorithm 1. ***Wake-Sleep Pseudocode***

$e \leftarrow 0$
$\mathcal{L}_{prev} \leftarrow \infty$
while $e < L$ **do** ▷ Repeat for L data epochs
  for t in range$(0, T_w)$ **do** ▷ Wake Phase
    for k in range$(0, K)$ **do** ▷ Sample network states
      $\mathbf{r}^{(k)} \sim b(\mathbf{r}, \theta_b)$
      if epoch is complete **then** $e \leftarrow e + 1$
      end if
    end for
    $\mathcal{L}(\theta_p) \leftarrow \frac{1}{K} \sum_{k=0}^{K} -\log p(\mathbf{r}^{(k)}, \theta_p)$ ▷ Update generative parameters θ_p
    $\Delta\theta_p \leftarrow \eta_p \nabla_{\theta_p} \mathcal{L}(\theta_p)$
    $\theta_p \leftarrow \text{Adam}(\theta_p, \Delta\theta_p)$
  end for
  if $\mathcal{L}(\theta_p) < \mathcal{L}_{prev}$ **then**
    $\mathcal{L}_{prev} \leftarrow \mathcal{L}(\theta_p)$
    for t in range$(0, T_s)$ **do** ▷ Sleep Phase
      for k in range$(0, K)$ **do** ▷ Sample network states
        $\mathbf{r}^{(k)} \sim p(\mathbf{r}, \theta_p)$
      end for
      $\mathcal{L}(\theta_b) \leftarrow \frac{1}{K} \sum_{k=0}^{K} -\log b(\mathbf{r}^{(k)}, \theta_b)$ ▷ Update inference parameters θ_b
      $\Delta\theta_b \leftarrow \eta_b \nabla_{\theta_b} \mathcal{L}(\theta_b)$
      $\theta_b \leftarrow \text{Adam}(\theta_b, \Delta\theta_b)$
    end for
  end if
end while

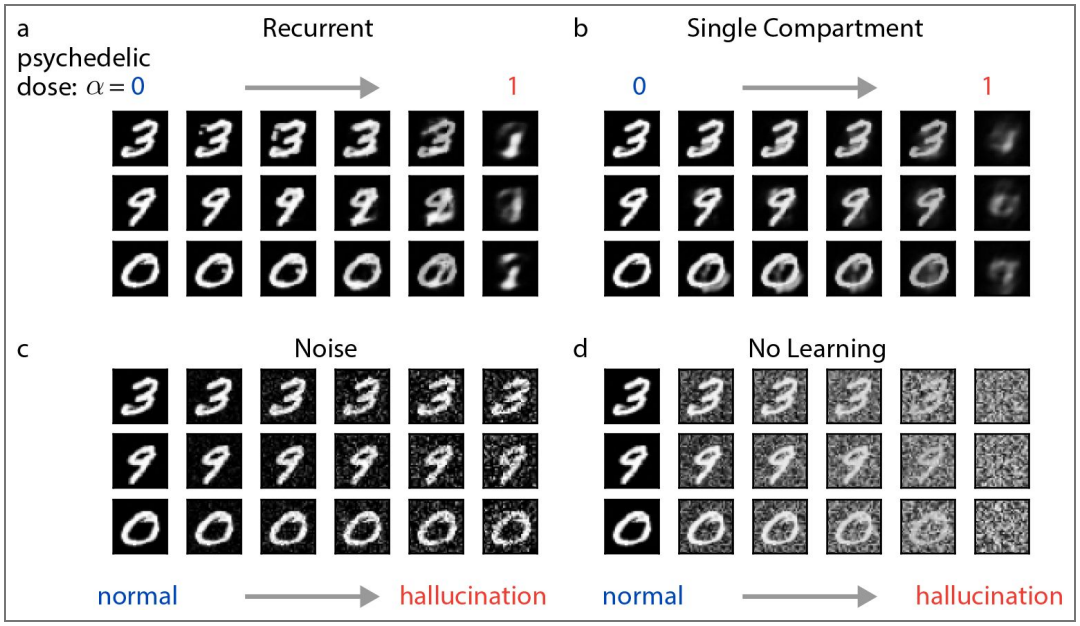


Figure S1. **Visualizing the effects of psychedelics for alternative model architectures.** We model the effects of classical psychedelics by progressively increasing α from 0 to 1 in alternative model architectures. We visualize the effects of psychedelics on the network representation by inspecting the stimulus layer **s. a)** Example stimulus-layer activity (rows) in response to an MNIST digit presentation as psychedelic dose increases (columns, left to right) in the recurrent network model. **b)** Same as (a) but for our single compartment neuron model. **c)** Same as (a) using the multicompartment neuron model used for our main results, but for our noise-based hallucination hallucination protocol. **d)** Same as (c), but in a network in which neither the generative nor inference pathways have been trained beyond initialization.

Figure S2. Example generated images for different model architectures and datasets.

Generated images sampled from Eq. (1) with $\alpha = 1$ for: **a)** Our primary multicompartment neuron model trained on MNIST, **b)** A multicompartment neuron model trained on CIFAR10, **c)** The recurrent network model, **d)** The single compartment neuron model.

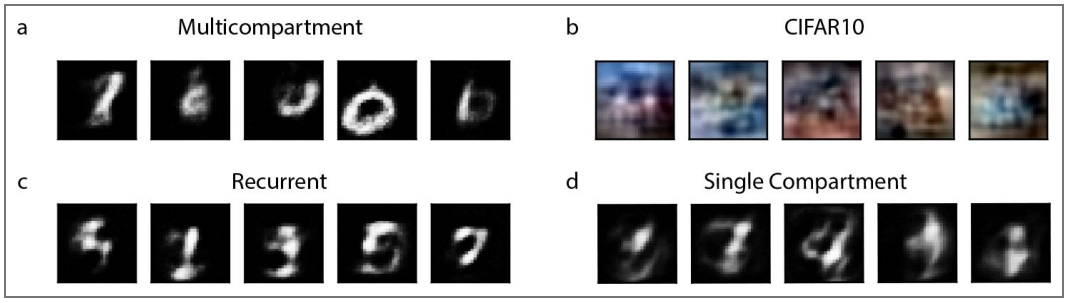


Figure S3. Alignment between apical and basal dendritic compartments for different model architectures and datasets.

Apical-basal alignment for: **a)** An untrained multicompartment neuron model trained on MNIST, **b)** A single compartment neuron model, **c)** A recurrent network model, **d)** A multicompartment neuron model trained on CIFAR10.

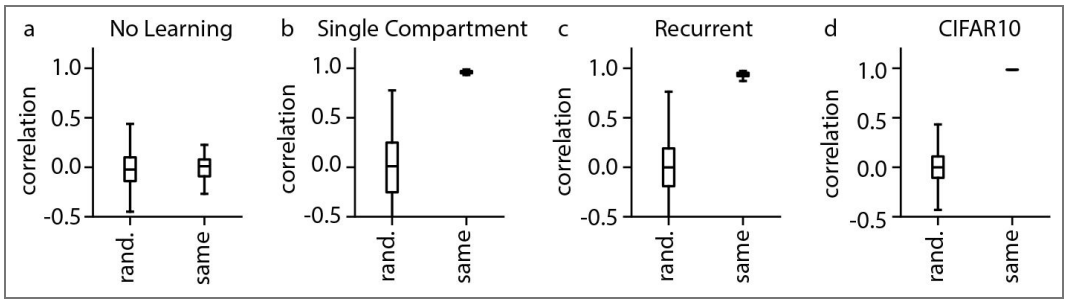
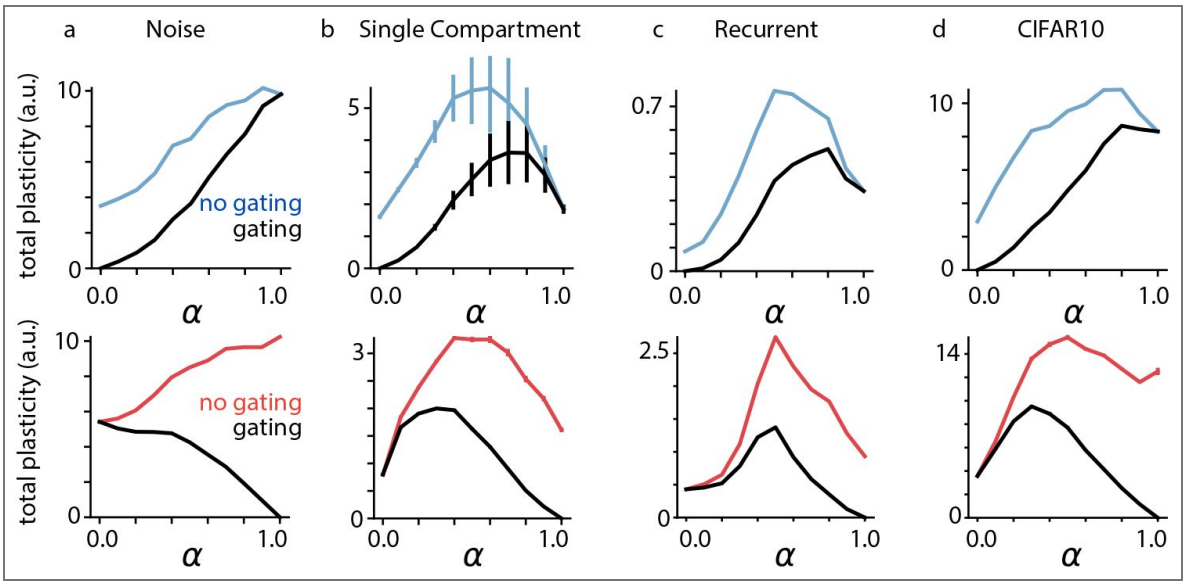


Figure S4. Hallucination-induced synaptic plasticity for different neuron models.

a) Basal (top) and apical (bottom) plasticity as a function of α for a multicompartment neuron model trained on MNIST, using our noise-based hallucination protocol as a control. **b)** Same as (a) for a single compartment neuron model, using our primary hallucination protocol. **c)** Same as (b) for a recurrent network model, **d)** Same as (b) for a multicompartment neuron model trained on CIFAR10. Error bars indicate +/-1 s.e.m.



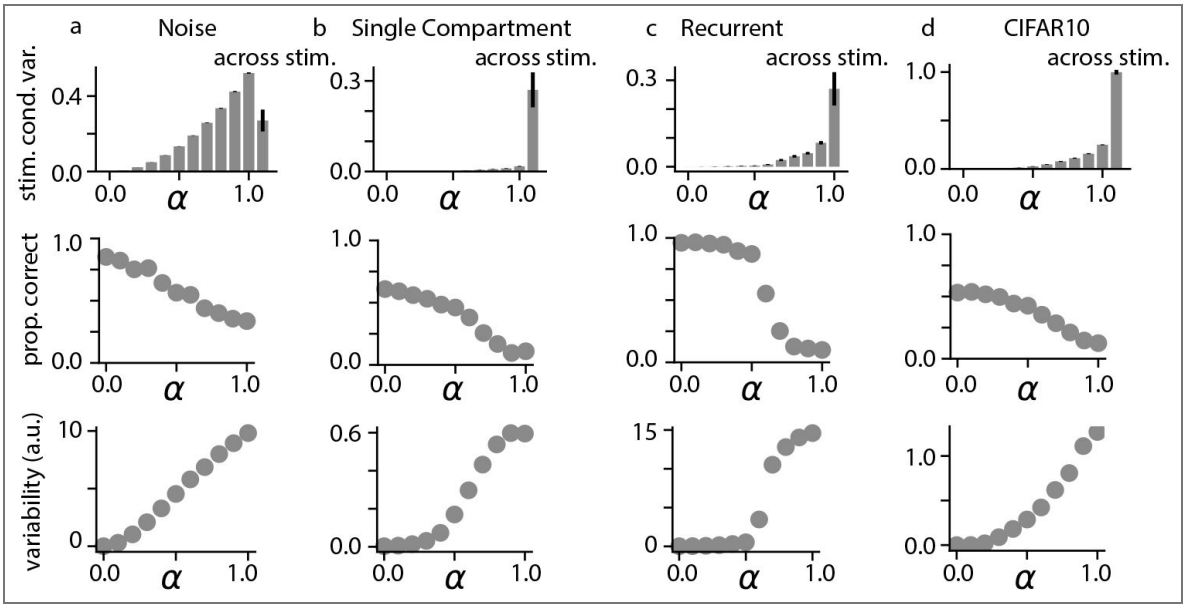


Figure S5. Neural variability changes for different neuron models.

a) Stimulus-conditioned variability (top), classifier accuracy (middle), and classifier output variability (bottom) as a function of *α* for a multicompartment neuron model trained on MNIST, using our noise-based hallucination protocol as a control. **b)** Same as (b) for a single compartment neuron model, using our primary hallucination protocol. **c)** Same as (b) for a recurrent network model, **d)** Same as (b) for a multicompartment neuron model trained on CIFAR10. Error bars indicate +/-1 s.e.m.

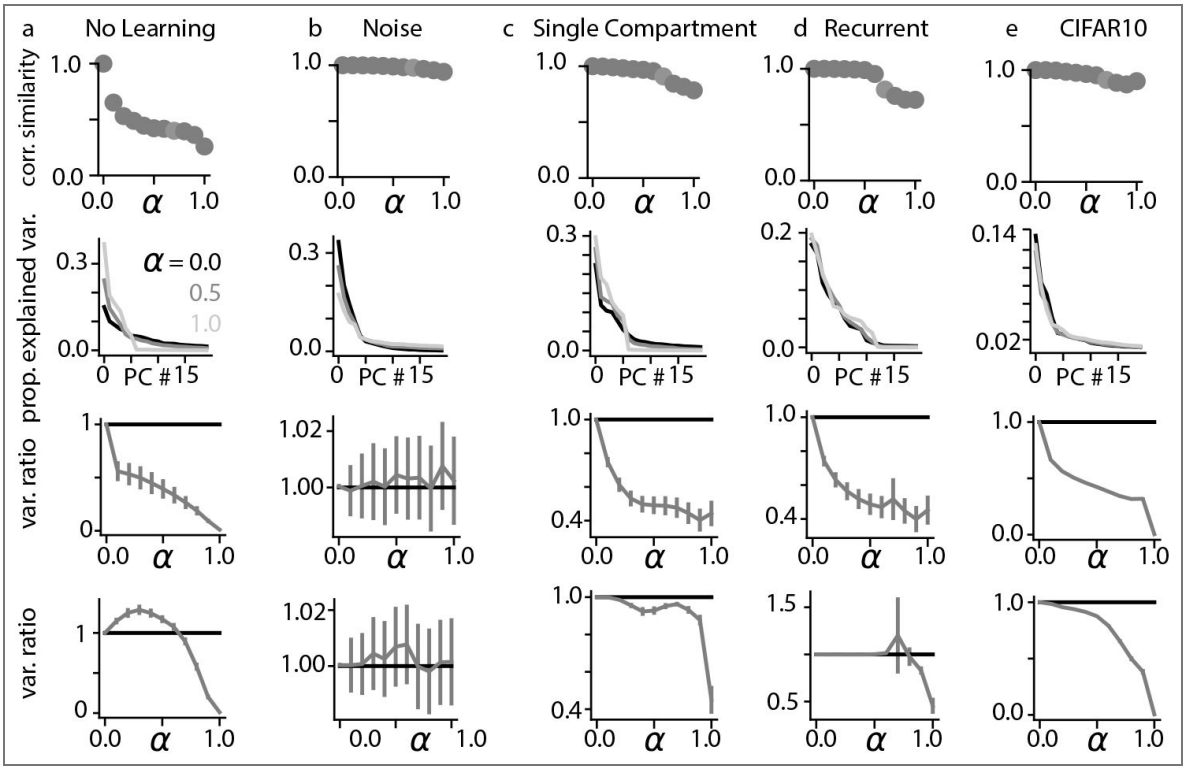


Figure S6. Network-level effects of psychedelics for different network architectures and training datasets.

For each network architecture, we examine: correlation similarity as a function of α (top row), the proportion explained variance across stimuli as a function of principal component number (second row), the ratio of across-stimulus variance in stimulus layer neurons when apical dendrites have been inactivated compared to baseline conditions across different α values (third row), and the ratio of across-stimulus variance in stimulus layer neurons when the deepest network layer has been inactivated across different α values (fourth row). **a)** Results for an untrained multicompartment neuron. **b)** Results for a multicompartment neuron model trained on MNIST, using our noise-based hallucination protocol. **c)** Results for a single compartment neuron model. **d)** Results for a recurrent network model. **e)** Results for a multicompartment neuron model trained on CIFAR10. Error bars indicate +/-1 s.e.m.

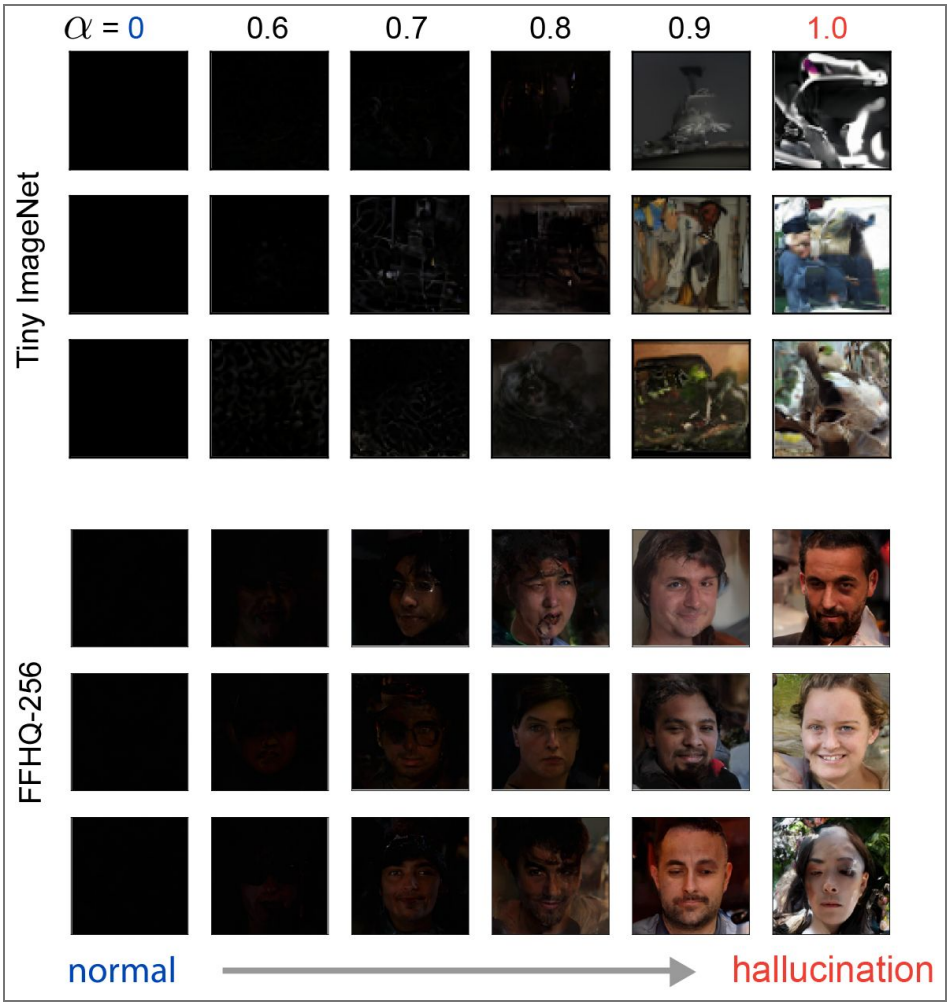


Figure S7. Visualizing the eyes-closed effects of psychedelics in pretrained VDVAE models.

Decoded outputs of a pretrained VDVAE model trained on Tiny ImageNet (Top) and FFHQ-256 (Bottom) based on hallucinations generated in the top 35 layers of the model. Black input images were used for samples in all rows. Hallucination intensity, parameterized by *α*, increases along columns.

Figure S8. Analyzing the image- and network-level effects of psychedelics in a Tiny ImageNet-pretrained VDVAE model.

a) Laplacian pyramid features for a grayscale example input image from the Tiny ImageNet dataset (top left). Pyramid levels increase along columns, corresponding to decreasing resolution, and hallucination intensity increases along rows. **b)** Correlation similarity across pyramid levels between the base image and a hallucinated image across different α values, averaged over 100 image samples. **c)** Stimulus-conditioned variance of units in layer 30 (descending from the top of the network) across different α values, averaged over 100 sample images and 32 distinct trials. **d)** Correlation similarity calculated between correlation matrices for units in layer 30 across different α values, averaged over spatial positions and 100 sample images. **e)** Ratio of across-stimulus variance in individual units of layer 30 when the highest 20 layers of the network have been inactivated, versus baseline conditions across different α values. Error bars indicate +/-1 s.e.m.

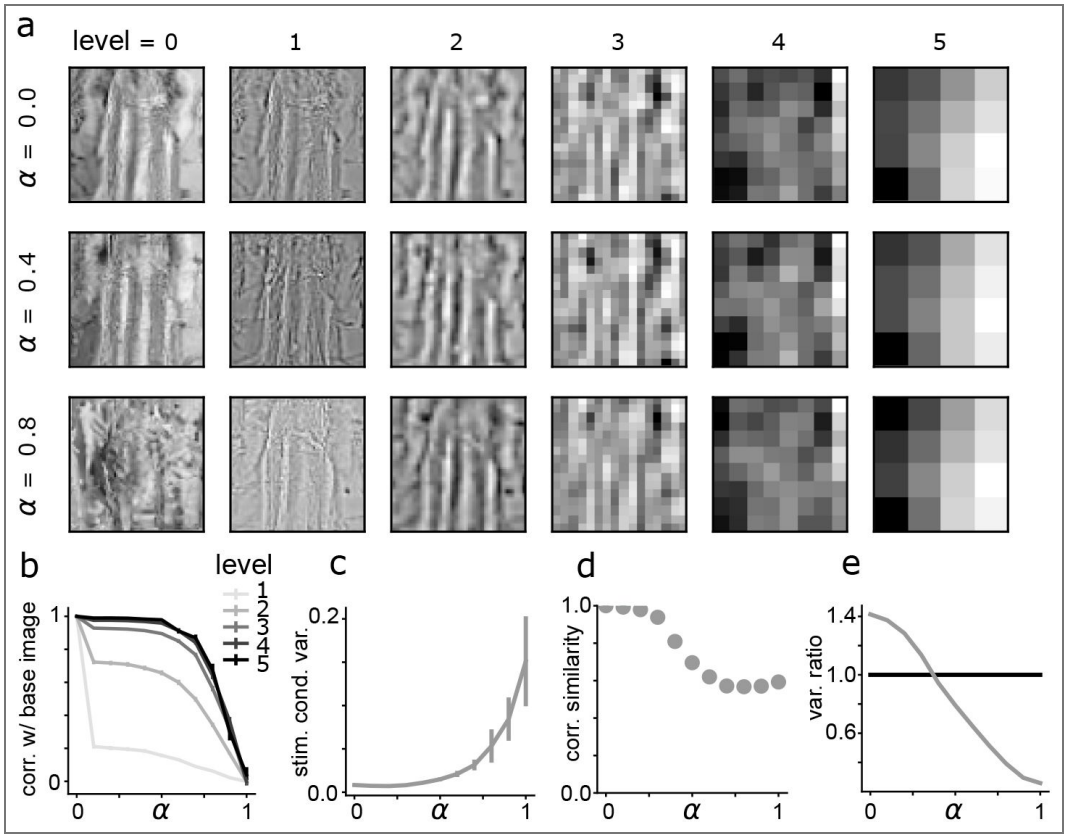


Table S1. MNIST multicompartment network hyperparameters.

Parameter	Function	Value
$\{N^{(l)}\}$	Layer Widths	$[32, 16, 6]$
σ_b	Inference standard dev.	0.01
σ_p	Generative standard dev.	0.01
N_d	Number of dendritic branches	16
L	Epoch number	50
T_w	Length of Wake phase	200
T_s	Length of Sleep phase	200
K	Batch number	512
η_p	Generative parameter learning rate	0.001
η_b	Inference parameter learning rate	0.001
$[\beta_1, \beta_2]$	Adam optimizer parameters	[0.9, 0.999]

Table S2. **CIFAR10 multicompartment network hyperparameters.**

Parameter	Function	Value
$\{N^{(l)}\}$	Layer Widths	$[128, 128, 96]$
σ_b	Inference standard dev.	0.01
σ_p	Generative standard dev.	0.01
N_d	Number of dendritic branches	16
L	Epoch number	50
T_w	Length of Wake phase	200
T_s	Length of Sleep phase	200
K	Batch number	256
η_p	Generative parameter learning rate	0.001
η_b	Inference parameter learning rate	0.001
$[\beta_1, \beta_2]$	Adam optimizer parameters	[0.9, 0.999]

Table S3. **Recurrent network hyperparameters.**

Parameter	Function	Value
$\{N^{(l)}\}$	Layer Widths	$[512, 256, 256, 128, 64, 32, 16]$
σ_b	Denoised inference standard dev.	0.01
$\bar{\sigma}_b$	Noisy inference standard dev.	0.01
σ_p	Denoised generative standard dev.	0.01
$\bar{\sigma}_p$	Noisy generative standard dev.	0.01
L	Epoch number	400
T_w	Length of Wake phase	200
T_s	Length of Sleep phase	50
K	Batch number	512
η_p	Generative parameter learning rate	0.001
η_b	Inference parameter learning rate	0.0001
$[\beta_1, \beta_2]$	Adam optimizer parameters	[0.9, 0.999]

Data availability

Code for reproducing all results from Wake-Sleep-trained models this study is available here: *https://github.com/* colinbredenberg/oneirogen-hypothesis. Code for reproducing results obtained with pretrained VDVAE models is available here: *https://github.com/* colinbredenberg/vdvae.

Acknowledgements

We would like to thank members of both G.L. and B.R.'s labs, as well as James M. Shine, Brandon Munn, Christopher Whyte, Veronica Chelu, Jiameng Wu, Matthew Larkum, Santiago Jaramillo, Michael Wehr, Neil Savalia, Alexandra Klein, Sarah Cook, Conor Lane, Anousheh Bakhti-Suroosh, Runchong Wang, Michael Okun, and Jordan O'Byrne for insightful discussions and feedback. This work was supported by: [GL] NSERC Discovery Grant (RGPIN-2018-04821), Canada CIFAR AI Chair Program, Canada Research Chair in Neural Computations and Interfacing (CIHR, tier 2). [BR] NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031; Arthur B. McDonald Fellowship: 566355-2022) and CIFAR (Canada AI Chair; Learning in Machine and Brains Fellowship). [CB] is supported in part by the FRQNT Strategic Clusters Program (Centre UNIQUE - Quebec Neuro-AI Research Center). The authors acknowledge the material support of NVIDIA in the form of computational resources.

Additional information

Ethics Declarations

Psychedelic drug research has a long history fraught with many instances of unethical research practice [131]. Further, psychedelic drug use itself has long been stigmatized and punished through legal measures [132], often at the expense of indigenous peoples, who have long incorporated psychoactive substances into their cultural and spiritual practices [133]. In the interest of avoiding a repetition of past mistakes, we feel compelled to provide explicit guidance on how our work should be interpreted and used. To do so, we will take inspiration from two principal ethical frameworks: the Montreal Declaration on Responsible AI [134], and the EQUIP framework for equity-oriented healthcare [135, 136]. We strongly encourage anyone considering extending our research or using our work in any form of clinical setting to ensure that subsequent research adheres to these frameworks.

Below, drawing from these ethical frameworks, we will provide a set of guidelines for how our work should be interpreted and used. Though these guidelines are by no means exhaustive, our hope is that adherence to them will help promote the potential positive outcomes of our work while limiting potential negative consequences.

Guidelines for the ethical use of this study

Do

1. Ensure that the elements of our hypothesis have been adequately tested, as outlined in our discussion, *before* using our framework in any form of clinical or therapeutic setting.

2. Use our ideas to inform further basic neuroscience research on perception, learning, sleep, and replay phenomena.

3. Explore our ideas as an opportunity to inform your own understanding of cognition, learning, and perception, with the understanding that these ideas have not yet been fully validated experimentally.

4. Feel free to ask us if you are worried that your proposed use of our work may have negative impacts.

Do not

1. Report our results as scientific fact. We have outlined a *hypothesis*, which is designed to be tested by the experimental neuroscience community.

2. Cite or interpret our results without an adequate understanding of the evidence supporting the various claims made in this study. Feel free to ask us if you are worried that you may be misinterpreting our results.

3. Use our results to extract undue or inequitable profit. The ideas developed in this paper are the product of decades of research and public funding, built upon *centuries* of exploration of psychedelics. Any knowledge or value contained within this paper is the common heritage of all humanity, with particular recognition due to the indigenous and marginalized communities that have historically suffered and are currently suffering from oppressive government and industry policies.

4. Use our results for any application that could violate human rights or harm human beings in any way.

Funding

Funder	Grant reference number	Author
Natural Sciences and Engineering Research Council of Canada (NSERC)	RGPIN-2018-04821	Guillaume Lajoie
CIFAR AI Chair Program		Guillaume Lajoie Blake Richards
Canada Research Chair in Neural Computations and Interfacing (CIHR, tier 2)		Guillaume Lajoie
Natural Sciences and Engineering Research Council of Canada (NSERC)	RGPIN-2020-05105	Blake Richards
Natural Sciences and Engineering Research Council of Canada (NSERC)	RGPAS-2020-00031	Blake Richards
Arthur B. McDonald Fellowship	566355-2022	Blake Richards
CIFAR Learning in Machine and Brains Fellowship		Blake Richards
FRQNT Strategic Clusters Program (Centre UNIQUE - Quebec Neuro-AI Research Center)		Colin Bredenberg

Author ORCID iDs

Colin Bredenberg: https://orcid.org/0000-0002-9749-9228

Additional files

Figure 2a multicompartmental mnist hallucination video

Figure 2c multicompartmental cifar10 hallucination video

References

[1] **Jakab Robert L,** Goldman-Rakic Patricia S (1998) 5-hydroxytryptamine2a serotonin receptors in the primate cerebral cortex: possible site of action of hallucinogenic and antipsychotic drugs in pyramidal cell apical dendrites. *Proceedings of the National Academy of Sciences* **95**:735-740

[2] **Almeida J De,** Mengod G (2007) Quantitative analysis of glutamatergic and GABAergic neurons expressing 5-HT2A receptors in human and monkey prefrontal cortex. *Journal of neurochemistry* **103**:475-486

[3] **Preller Katrin H,** Vollenweider Franz X (2018) Phenomenology, structure, and dynamic of psychedelic states. *Behavioral neurobiology of psychedelic drugs* 221-256

[4] **Shao Ling-Xiao,** Liao Clara, Gregg Ian, Davoudian Pasha A, Savalia Neil K, Delagarza Kristina, Kwan Alex C (2021) Psilocybin induces rapid and persistent growth of dendritic spines in frontal cortex in vivo. *Neuron* **109**:2535-2544

[5] **Grieco Steven F,** Castrén Eero, Knudsen Gitte M, Kwan Alex C, Olson David E, Zuo Yi, Holmes Todd C, Xu Xiangmin (2022) Psychedelics and neural plasticity: therapeutic implications. *Journal of Neuroscience* **42**:8439-8449

[6] **Muttoni Silvia,** Ardissino Maddalena, John Christopher (2019) Classical psychedelics for the treatment of depression and anxiety: A systematic review. *Journal of affective disorders* **258**:11-24

[7] **Krediet Erwin,** Bostoen Tijmen, Breeksema Joost, van Schagen Annette, Passie Torsten, Vermetten Eric (2020) Reviewing the potential of psychedelics for the treatment of PTSD. *International Journal of Neuropsychopharmacology* **23**:385-400

[8] **Vargas Maxemiliano V,** Dunlap Lee E, Dong Chunyang, Carter Samuel J, Tombari Robert J, Jami Shekib A, Cameron Lindsay P, Patel Seona D, Hennessey Joseph J, Saeger Hannah N, *et al.* (2023) Psychedelics promote neuroplasticity through the activation of intracellular 5-HT2A receptors. *Science* **379**:700-706

[9] **Kraehenmann Rainer,** Pokorny Dan, Vollenweider Leonie, Preller Katrin H, Pokorny Thomas, Seifritz Erich, Vollenweider Franz X (2017) Dreamlike effects of LSD on waking imagery in humans depend on serotonin 2a receptor activation. *Psychopharmacology* **234**:2031-2046

[10] **Green A Richard,** Mechan Annis O, Elliott J Martin, O'Shea Esther, Colado M Isabel (2003) The pharmacology and clinical pharmacology of 3, 4-methylenedioxymethamphetamine (MDMA,"ecstasy"). *Pharmacological reviews* **55**:463-508

[11] **Marona-Lewicka Danuta,** Kurrasch-Orbaugh Deborah M, Selken Jennifer R, Cumbay Medhane G, Lisnicchia Joshua G, Nichols David E (2002) Re-evaluation of lisuride pharmacology: 5hydroxytryptamine1a receptor-mediated behavioral effects overlap its other properties in rats. *Psychopharmacology* **164**:93-107

[12] **Carhart-Harris Robin** (2007) Waves of the unconscious: the neurophysiology of dreamlike phenomena and its implications for the psychodynamic model of the mind. *Neuropsychoanalysis* **9**:183-211

[13] **Thomas Christopher W,** Blanco-Duque Cristina, Bréant Benjamin J, Goodwin Guy M, Sharp Trevor, Bannerman David M, Vyazovskiy Vladyslav V (2022) Psilocin acutely alters sleep-wake architecture and cortical brain activity in laboratory mice. *Translational psychiatry* **12**:77

[14] **Dudysová Daniela,** Janků Karolina, Šmotek Michal, Saifutdinova Elizaveta, Kopřivová Jana, Bušková Jitka, Mander Bryce Anthony, Brunovský Martin, Zach Peter, Korčák Jakub, *et al.* (2020) The effects of daytime psilocybin administration on sleep: implications for antidepressant action. *Frontiers in pharmacology* **11**:602590

[15] **Barbanoj Manel J,** Riba Jordi, Clos S, Giménez S, Grasa E, Romero S (2008) Daytime ayahuasca administration modulates REM and slow-wave sleep in healthy volunteers. *Psychopharmacology* **196**:315-326

[16] **Girardeau Gabrielle,** Benchenane Karim, Wiener Sidney I, Buzsáki György, Zugaro Michaël B (2009) Selective suppression of hippocampal ripples impairs spatial memory. *Nature neuroscience* **12**:1222-1223

[17] **Deuker Lorena,** Olligs Jan, Fell Juergen, Kranz Thorsten A., Mormann Florian, Montag Christian, Reuter Martin, Elger Christian E., Axmacher Nikolai (2013) Memory consolidation by replay of stimulus-specific neural activity. *Journal of Neuroscience* **33**:19373-19383

[18] **Lavilléon Gaetan D.,** Lacroix Marie Masako, Rondi-Reig Laure, Benchenane Karim (2015) Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nature neuroscience* **18**:493-495

[19] **Maingret Nicolas,** Girardeau Gabrielle, Todorova Ralitsa, Goutierre Marie, Zugaro Michaël (2016) Hippocampo-cortical coupling mediates memory consolidation during sleep. *Nature neuroscience* **19**:959-964

[20] **Fernández-Ruiz Antonio,** Oliva Azahara, de Oliveira Eliezyer Fermino, Rocha-Almeida Florbela, Tingley David, Buzsáki György (2019) Long-duration hippocampal sharp wave ripples improve memory. *Science* **364**:1082-1086

[21] **Walker Matthew P,** Stickgold Robert (2004) Sleep-dependent learning and memory consolidation. *Neuron* **44**:121-133

[22] **Nádasdy Zoltán,** Hirase Hajime, Czurkó András, Csicsvari Jozsef, Buzsáki György (1999) Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience* **19**:9497-9507

[23] **Lee Albert K,** Wilson Matthew A (2002) Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* **36**:1183-1194

[24] **Foster David J** (2017) Replay comes of age. *Annual review of neuroscience* **40**:581-602

[25] **Ji Daoyun,** Wilson Matthew A (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature neuroscience* **10**:100-107

[26] **Euston David R,** Tatsuno Masami, McNaughton Bruce L (2007) Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *science* **318**:1147-1150

[27] **Peyrache Adrien,** Khamassi Mehdi, Benchenane Karim, Wiener Sidney I, Battaglia Francesco P (2009) Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature neuroscience* **12**:919-926

[28] **Kenet Tal,** Bibitchkov Dmitri, Tsodyks Misha, Grinvald Amiram, Arieli Amos (2003) Spontaneously emerging cortical representations of visual attributes. *Nature* **425**:954-956

[29] **Xu Shengjin,** Jiang Wanchen, Poo Mu-ming, Dan Yang (2012) Activity recall in a visual cortical ensemble. *Nature neuroscience* **15**:449-455

[30] **Hoffman Kari L,** McNaughton Bruce L (2002) Coordinated reactivation of distributed memory traces in primate neocortex. *Science* **297**:2070-2073

[31] **Louie Kenway,** Wilson Matthew A (2001) Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**:145-156

[32] **Andrillon Thomas,** Nir Yuval, Cirelli Chiara, Tononi Giulio, Fried Itzhak (2015) Single-neuron activity and eye movements during human rem sleep and awake vision. *Nature communications* **6**:7884

[33] **Hinton Geoffrey E,** Dayan Peter, Frey Brendan J, Neal Radford M (1995) The “wake-sleep” algorithm for unsupervised neural networks. *Science* **268**:1158-1161

[34] **Rezende Danilo Jimenez,** Mohamed Shakir, Wierstra Daan (2014) Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning. PMLR. pp. 1278-1286

[35] **Kingma Diederik P,** Welling Max (2013) Auto-encoding variational bayes. *arXiv*

[36] **Simoncelli Eero P** (2003) Vision and the statistics of the visual environment. *Current opinion in neurobiology* **13**:144-149

[37] **Simoncelli Eero P,** Olshausen Bruno A (2001) Natural image statistics and neural representation. *Annual review of neuroscience* **24**:1193-1216

[38] **Ballé Johannes,** Laparra Valero, Simoncelli Eero P (2016) End-to-end optimized image compression. *arXiv*

[39] **DiCarlo James J,** Zoccolan Davide, Rust Nicole C (2012) How does the brain solve visual object recognition?. *Neuron* **73**:415-434

[40] **Higgins Irina,** Matthey Loic, Pal Arka, Burgess Christopher P, Glorot Xavier, Botvinick Matthew M, Mohamed Shakir, Lerchner Alexander (2017) beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)* **3**

[41] **Ikeda Shiro,** Amari Shun-ichi, Nakahara Hiroyuki (1998) Convergence of the wake-sleep algorithm. *Advances in neural information processing systems* **11**

[42] **Kirby Kevin G** (2006) A tutorial on Helmholtz Machines. Department of Computer Science, Northern Kentucky University.

[43] **Larkum Matthew** (2013) A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in neurosciences* **36**:141-151

[44] **Aru Jaan,** Siclari Francesca, Phillips William A, Storm Johan F (2020) Apical drive—a cellular mechanism of dreaming?. *Neuroscience & Biobehavioral Reviews* **119**:440-455

[45] **Körding Konrad P,** König Peter (2001) Supervised and unsupervised learning with two sites of synaptic integration. *Journal of computational neuroscience* **11**:207-215

[46] **Urbanczik Robert,** Senn Walter (2014) Learning by the dendritic prediction of somatic spiking. *Neuron* **81**:521-528

[47] **Guerguiev Jordan,** Lillicrap Timothy P, Richards Blake A (2017) Towards deep learning with segregated dendrites. *eLife* **6**:e22901 https://doi.org/10.7554/eLife.22901

[48] **Sacramento João,** Costa Rui Ponte, Bengio Yoshua, Senn Walter (2018) Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems* **31**

[49] **Richards Blake A,** Lillicrap Timothy P (2019) Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology* **54**:28-36

[50] **Payeur Alexandre,** Guerguiev Jordan, Zenke Friedemann, Richards Blake A, Naud Richard (2021) Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience* **24**:1010-1019

[51] **Bredenberg Colin,** Lyo Benjamin, Simoncelli Eero, Savin Cristina (2021) Impression learning: Online representation learning with synaptic plasticity. *Advances in Neural Information Processing Systems* **34**:11717-11729

[52] **George Tom M,** Stachenfeld Kimberly L, Barry Caswell, Clopath Claudia, Fukai Tomoki (2024) A generative model of the hippocampal formation trained with theta driven local learning rules. *Advances in Neural Information Processing Systems* **36**

[53] **Seibt Julie,** Richard Clément J, Sigl-Glöckner Johanna, Takahashi Naoya, Kaplan David I, Doron Guy, De Limoges Denis, Bocklisch Christina, Larkum Matthew E (2017) Cortical dendritic activity correlates with spindle-rich oscillations during sleep in rodents. *Nature communications* **8**:684

[54] **Miyamoto Daisuke,** Hirai Daichi, Fung CCA, Inutsuka Ayumu, Odagawa Maya, Suzuki Takayuki, Boehringer Roman, Adaikkan Chinnakkaruppan, Matsubara Chie, Matsuki Norio, *et al.* (2016) Top-down cortical input during nrem sleep consolidates perceptual memory. *Science* **352**:1315-1318

[55] **Li Wei,** Ma Lei, Yang Guang, Gan Wen-Biao (2017) Rem sleep selectively prunes and maintains new synapses in development and learning. *Nature neuroscience* **20**:427-437

[56] **Zhou Yanmei,** Lai Cora Sau Wan, Bai Yang, Li Wei, Zhao Ruohe, Yang Guang, Frank Marcos G, Gan Wen-Biao (2020) Rem sleep promotes experience-dependent dendritic spine elimination in the mouse cortex. *Nature communications* **11**:4819

[57] **Aime Mattia,** Niccoló Calcini, Borsa Micaela, Campelo Tiago, Rusterholz Thomas, Sattin Andrea, Fellin Tommaso, Adamantidis Antoine (2022) Paradoxical somatodendritic decoupling supports cortical plasticity during rem sleep. *Science* **376**:724-730

[58] **Aghajanian GK,** Marek GJ (1997) Serotonin induces excitatory postsynaptic potentials in apical dendrites of neocortical pyramidal cells. *Neuropharmacology* **36**:589-599

[59] **Aghajanian George K,** Marek Gerard J (1999) Serotonin and hallucinogens. *Neuropsychopharmacology* **21**:16S-23S

[60] **Arvanov Viktor L,** Liang Xiaofu, Russo Angelo, Wang Rex Y (1999) LSD and DOB: interaction with 5-HT2A receptors to inhibit NMDA receptor-mediated transmission in the rat prefrontal cortex. *European Journal of Neuroscience* **11**:3064-3072

[61] **Hidalgo Javier,** Kaup Kristjan, Aru Jaan (2025) Electrophysiological mechanisms of psychedelic drugs: A systematic review. *bioRxiv*

[62] **Deng Li** (2012) The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine* **29**:141-142

[63] **Krizhevsky Alex,** Hinton Geoffrey, et al. (2009) Learning multiple layers of features from tiny images. University of Toronto.

[64] **Beaulieu-Laroche Lou,** Toloza Enrique HS, Brown Norma J, Harnett Mark T (2019) Widespread and highly correlated somato-dendritic activity in cortical layer 5 neurons. *Neuron* **103**:235-241

[65] **O'Hare Justin K,** Wang Jamie, Shala Margjele D, Polleux Franck, Losonczy Attila (2024) Distal tuft dendrites shape and maintain new place fields. *bioRxiv*

[66] **Nardou Romain,** Sawyer Edward, Song Young Jun, Wilkinson Makenzie, Padovan-Hernandez Yasmin, de Deus Júnia Lara, Wright Noelle, Lama Carine, Faltin Sehr, Goff Loyal A, *et al.* (2023) Psychedelics reopen the social reward learning critical period. *Nature* **618**:790-798

[67] **la Mario de,** Revenga Fuente, Zhu Bohan, Guevara Christopher A, Naler Lynette B, Saunders Justin M, Zhou Zirui, Toneatti Rudy, Sierra Salvador, Wolstenholme Jennifer T, *et al.* (2021) Prolonged epigenomic and synaptic plasticity alterations following single exposure to a psychedelic in mice. *Cell reports* **37**

[68] **Juliani Arthur,** Chelu Veronica, Graesser Laura, Safron Adam (2024) A dual-receptor model of serotonergic psychedelics: therapeutic insights from simulated cortical dynamics. *bioRxiv*

[69] **Carhart-Harris Robin L,** Friston Karl J (2019) Rebus and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews* **71**:316-344

[70] **Lebedev Alexander V,** Kaelen Mendel, Lövdén Martin, Nilsson Jonna, Feilding Amanda, Nutt David J, Carhart-Harris Robin L (2016) LSD-induced entropic brain activity predicts subsequent personality change. *Human brain mapping* **37**:3203-3213

[71] **Carhart-Harris Robin Lester,** Leech Robert, Hellyer Peter John, Shanahan Murray, Feilding Amanda, Tagliazucchi Enzo, Chialvo Dante R, Nutt David (2014) The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience* **8**:20

[72] **Siegel Joshua S,** Subramanian Subha, Perry Demetrius, Kay Benjamin P, Gordon Evan M, Laumann Timothy O, Reneau Rick, Metcalf Nicholas V, Chacko Ravi V, Gratton Caterina, *et al.* (2024) Psilocybin desynchronizes the human brain. *Nature*

[73] **Horrocks Max,** Mohn Jennifer L, Jaramillo Santiago (2024) The serotonergic psychedelic DOI impairs deviance detection in the auditory cortex. *bioRxiv*

[74] **Shanon Benny** (2002) Ayahuasca visualizations a structural typology. *Journal of Consciousness Studies* **9**:3-30

[75] **Díaz José Luis** (2010) Sacred plants and visionary consciousness. *Phenomenology and the Cognitive Sciences* **9**:159-170

[76] **Goodfellow Ian,** Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, Courville Aaron, Bengio Yoshua (2020) Generative adversarial networks. *Communications of the ACM* **63**:139-144

[77] **Casper Kaae Sønderby,** Raiko Tapani, Lars Maaløe, Kaae Sønderby Søren, Winther Ole (2016) Ladder variational autoencoders. *Advances in neural information processing systems* **29**

[78] **Vahdat Arash,** Kautz Jan (2020) Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems* **33**:19667-19679

[79] **Child Rewon** (2020) Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv*

[80] **Wu Jiayu,** Zhang Qixiang, Xu Guoxi (2017) Tiny imagenet challenge. Stanford University.

[81] **Karras Tero,** Laine Samuli, Aila Timo (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401-4410

[82] **Lillicrap Timothy P,** Santoro Adam, Marris Luke, Akerman Colin J, Hinton Geoffrey (2020) Backpropagation and the brain. *Nature Reviews Neuroscience* **21**:335-346

[83] **Pogodin Roman,** Mehta Yash, Lillicrap Timothy, Latham Peter E (2021) Towards biologically plausible convolutional networks. *Advances in neural information processing systems* **34**:13924-13936

[84] **Csikor Ferenc,** Meszéna Balázs, Szabó Bence, Orbán Gergó (2022) Top-down inference in an early visual cortex inspired hierarchical variational autoencoder. *arXiv*

[85] **Evarts Edward V,** Landau William, jr Walter Freygang, Marshall Wade H (1955) Some effects of lysergic acid diethylamide and bufotenine on electrical activity in the cat's visual system. *American Journal of Physiology-Legacy Content* **182**:594-598

[86] **Azimi Zohre,** Barzan Ruxandra, Spoida Katharina, Surdin Tatjana, Wollenweber Patric, Mark Melanie D, Herlitze Stefan, Jancke Dirk (2020) Separable gain control of ongoing and evoked activity in the visual cortex by serotonergic input. *eLife* **9**:e53552 https://doi.org/10.7554/eLife.53552

[87] **Michaiel Angie M,** Parker Philip RL, Niell Cristopher M (2019) A hallucinogenic serotonin-2a receptor agonist reduces visual response gain and alters temporal dynamics in mouse v1. *Cell reports* **26**:3475-3483

[88] **Widrow Bernard,** Lehr Michael A (1990) 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE* **78**:1415-1442

[89] **Jesper Sjöström Per,** Häusser Michael (2006) A cooperative switch determines the sign of synaptic plasticity in distal dendrites of neocortical pyramidal neurons. *Neuron* **51**:227-238

[90] **Froemke Robert C,** Letzkus Johannes J, Kampa Björn M, Hang Giao B, Stuart Greg J (2010) Dendritic synapse location and neocortical spike-timing-dependent plasticity. *Frontiers in synaptic neuroscience* **2**:29

[91] **Mediano Pedro AM,** Rosas Fernando E, Timmermann Christopher, Roseman Leor, Nutt David J, Feilding Amanda, Kaelen Mendel, Kringelbach Morten L, Barrett Adam B, Seth Anil K, *et al.* (2024) Effects of external stimulation on psychedelic state neurodynamics. *ACS Chemical Neuroscience* **15**:462-471

[92] **Bredenberg Colin,** Savin Cristina (2024) Desiderata for normative models of synaptic plasticity. *Neural Computation* **36**:1245-1285

[93] **Levenstein Daniel,** Alvarez Veronica A, Amarasingham Asohan, Azab Habiba, Chen Zhe S, Gerkin Richard C, Hasenstaub Andrea, Iyer Ramakrishnan, Jolivet Renaud B, Marzen Sarah, *et al.* (2023) On the role of theory and modeling in neuroscience. *Journal of Neuroscience* **43**:1074-1088

[94] **Ermentrout G Bard,** Cowan Jack D (1979) A mathematical theory of visual hallucination patterns. *Biological cybernetics* **34**:137-150

[95] **Suzuki Keisuke,** Roseboom Warrick, Schwartzman David J, Seth Anil K (2017) A deep-dream virtual reality platform for studying altered perceptual phenomenology. *Scientific reports* **7**:15982

[96] **Stickgold Robert,** Hobson Allen, Fosse Roar, Fosse Magdalena (2001) Sleep, learning, and dreams: off-line memory reprocessing. *Science* **294**:1052-1057

[97] **Domenico Carli,** Haggerty Daniel, Mou Xiang, Ji Daoyun (2021) Lsd degrades hippocampal spatial representations and suppresses hippocampal-visual cortical interactions. *Cell reports* **36**

[98] **Whyte Christopher J,** Redinbaugh Michelle J, Shine James M, Saalmann Yuri B (2024) Thalamic contributions to the state and contents of consciousness. *Neuron* **112**:1611-1625

[99] **Gaddis Andrew,** Lidstone Daniel E, Nebel Mary Beth, Griffiths Roland R, Mostofsky Stewart H, Mejia Amanda F, Barrett Frederick S (2022) Psilocybin induces spatially constrained alterations in thalamic functional organizaton and connectivity. *Neuroimage* **260**:119434

[100] **Pizzi Stefano Delli,** Chiacchiaretta Piero, Sestieri Carlo, Ferretti Antonio, Tullo Maria Giulia, Penna Stefania Della, Martinotti Giovanni, Onofrj Marco, Roseman Leor, Timmermann Christopher, *et al.* (2023) Lsd-induced changes in the functional connectivity of distinct thalamic nuclei. *NeuroImage* **283**:120414

[101] **Singh Dhairyya,** Norman Kenneth A, Schapiro Anna C (2022) A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *Proceedings of the national academy of sciences* **119**:e2123432119

[102] **Bressloff Paul C,** Cowan Jack D, Golubitsky Martin, Thomas Peter J, Wiener Matthew C (2001) Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **356**:299-330

[103] **Keller T Anderson,** Welling Max (2023) Neural wave machines: learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks. In: International Conference on Machine Learning. PMLR. pp. 16168-16189

[104] **Keller T Anderson,** Muller Lyle, Sejnowski Terrence, Welling Max (2023) Traveling waves encode the recent past and enhance sequence learning. *arXiv*

[105] **Rao Rajesh PN,** Ballard Dana H (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* **2**:79-87

[106] **Rajpal Hardik,** Mediano Pedro AM, Rosas Fernando E, Timmermann Christopher B, Brugger Stefan, Muthukumaraswamy Suresh, Seth Anil K, Bor Daniel, Carhart-Harris Robin L, Jensen Henrik J (2022) Psychedelics and schizophrenia: Distinct alterations to Bayesian inference. *NeuroImage* **263**:119624

[107] **Alamia Andrea,** Timmermann Christopher, Nutt David J, VanRullen Rufin, CarhartHarris Robin L (2020) Dmt alters cortical travelling waves. *eLife* **9**:e59784 https://doi.org/10.7554/eLife.59784

[108] **Ackley David H,** Hinton Geoffrey E, Sejnowski Terrence J (1985) A learning algorithm for Boltzmann machines. *Cognitive science* **9**:147-169

[109] **Reichert David P,** Series Peggy, Storkey Amos J (2013) Charles bonnet syndrome: evidence for a generative model in the cortex?. *PLoS computational biology* **9**:e1003134

[110] **O'Donohue Thomas L,** Millington William R, Handelmann Gail E, Contreras Patricia C, Chronwall Bibie M (1985) On the 50th anniversary of Dale's law: multiple neurotransmitter neurons. *Trends in Pharmacological Sciences* **6**:305-308

[111] **Cornford Jonathan,** Kalajdzievski Damjan, Leite Marco, Lamarquette Amélie, Kullmann Dimitri M, Richards Blake (2020) Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. *bioRxiv*

[112] **Munn Brandon R,** Medel Vicente, Naismith Sharon L, Lizier Joseph T, Sanders Robert D, Shine James M (2023) Neuronal connected burst cascades bridge macroscale adaptive signatures across arousal states. *Nature Communications* **14**:6846

[113] **Podlaski Bill,** Machens Christian K (2020) Biological credit assignment through dynamic inversion of feedforward networks. *Advances in Neural Information Processing Systems* **33**:10065-10076

[114] **Lindsay Grace W** (2020) Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience* **14**:29

[115] **Islah Nizar,** Etter Guillaume, Tugsbayar Mashbayar, Gurbuz Busra Tugce, Richards Blake, Muller Eilif B (2025) Learning to combine top-down context and feed-forward representations under ambiguity with apical and basal dendrites. *Cerebral Cortex* **35**:bhaf134

[116] **Gilra Aditya,** Gerstner Wulfram (2017) Predicting non-linear dynamics by stable local learning in a recurrent spiking neural network. *eLife* **6**:e28295 https://doi.org/10.7554/eLife.28295

[117] **Meulemans Alexander,** Farinha Matilde Tristany, Javier García Ordóñez, Aceituno Pau Vilimelis, Sacramento João, Grewe Benjamin F (2021) Credit assignment in neural networks through deep feedback control. *Advances in Neural Information Processing Systems* **34**:4674-4687

[118] **Roelfsema Pieter R,** van Ooyen Arjen (2005) Attention-gated reinforcement learning of internal representations for classification. *Neural computation* **17**:2176-2214

[119] **Carhart-Harris Robin L,** Nutt DJ (2017) Serotonin and brain function: a tale of two receptors. *Journal of psychopharmacology* **31**:1091-1120

[120] **Bombardi Cristiano,** Di Giovanni Giuseppe (2013) Functional anatomy of 5-ht2a receptors in the amygdala and hippocampal complex: relevance to memory functions. *Experimental brain research* **230**:427-439

[121] **Jiang Xiaolong,** Xing Guoqiang, Yang Chunhui, Verma Ajay, Zhang Lei, Li He (2009) Stress impairs 5-ht2a receptor-mediated serotonergic facilitation of gaba release in juvenile rat basolateral amygdala. *Neuropsychopharmacology* **34**:410-423

[122] **Kelly Thomas J,** Bonniwell Emma M, Mu Lianwei, Liu Xiaojie, Hu Ying, Friedman Vladislav, Yu Hao, Su Wantang, McCorvy John D, Liu Qing-song (2024) Psilocybin analog 4-oh-dipt enhances fear extinction and gabaergic inhibition of principal neurons in the basolateral amygdala. *Neuropsychopharmacology* **49**:854-863

[123] **Tiwari Praachi,** Davoudian Pasha A, Kapri Darshana, Vuruputuri Ratna Mahathi, Karaba Lindsay A, Sharma Mukund, Zanni Giulia, Balakrishnan Angarika, Chaudhari Pratik R, Pradhan Amartya, *et al.* (2024) Ventral hippocampal parvalbumin interneurons gate the acute anxiolytic action of the serotonergic psychedelic doi. *Neuron* **112**:3697-3714

[124] **Lesch Klaus-Peter,** Waider Jonas (2012) Serotonin in the modulation of neural plasticity and networks: implications for neurodevelopmental disorders. *Neuron* **76**:175-191

[125] **Corlett Philip R,** Frith Chris D, Fletcher Paul C (2009) From drugs to deprivation: a bayesian framework for understanding models of psychosis. *Psychopharmacology* **206**:515-530

[126] **Poirazi Panayiota,** Brannon Terrence, Mel Bartlett W (2003) Pyramidal neuron as two-layer neural network. *Neuron* **37**:989-999

[127] **Ioffe Sergey** (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*

[128] **Bredenberg Colin,** Williams Ezekiel, Savin Cristina, Richards Blake, Lajoie Guillaume (2024) Formalizing locality for normative synaptic plasticity models. In: Advances in Neural Information Processing Systems. **36**

[129] **Williams Ezekiel,** Bredenberg Colin, Lajoie Guillaume (2023) Flexible phase dynamics for bioplausible contrastive learning. In: International Conference on Machine Learning. PMLR. pp. 37042-37065

[130] **Simoncelli Eero,** Young Rob, Broderick William, Fiquet tienne, Wang Zhuo, Kadkhodaie Zahra, Parthasarathy Nikhil, Ward Brian (2025) Pyrtools: tools for multi-scale image processing.

[131] **Strauss Dana,** Salle Sara de la, Sloshower Jordan, Williams Monnica T (2022) Research abuses against people of colour and other vulnerable groups in early psychedelic research. *Journal of Medical Ethics* **48**:728-737

[132] **Bauml James A,** Schaefer Stacy B (2016) *Peyote: History, tradition, politics, and conservation* Bloomsbury Publishing USA.

[133] **Samorini Giorgio** (2019) The oldest archeological data evidencing the relationship of homo sapiens with psychoactive plants: A worldwide overview. *Journal of Psychedelic Studies* **3**:63-80

[134] **Dilhac Marc-Antoine,** Abrassart Christophe, Voarino Nathalie (2018) Report of the Montréal declaration for a responsible development of artificial intelligence.

[135] **Rea Kerri,** Wallace Bruce (2021) Enhancing equity-oriented care in psychedelic medicine: Utilizing the EQUIP framework. *International Journal of Drug Policy* **98**:103429

[136] **Browne Annette J,** Varcoe Colleen, Ford-Gilboe Marilyn, Wathen C Nadine, EQUIP Research Team (2015) EQUIP healthcare: An overview of a multi-component intervention to enhance equityoriented care in primary health care settings. *International journal for equity in health* **14**:1-11

[137] **Dayan Peter,** Hinton Geoffrey E, Neal Radford M, Zemel Richard S (1995) The Helmholtz machine. *Neural computation* **7**:889-904

[138] **Toosi Tahereh,** Issa Elias (2024) Brain-like flexible visual inference by harnessing feedback feedforward alignment. *Advances in Neural Information Processing Systems* **36**

[139] **Faisal A Aldo,** Selen Luc PJ, Wolpert Daniel M (2008) Noise in the nervous system. *Nature reviews neuroscience* **9**:292-303

[140] **Vincent Pascal** (2011) A connection between score matching and denoising autoencoders. *Neural computation* **23**:1661-1674

[141] **Kadkhodaie Zahra,** Simoncelli Eero (2021) Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems* **34**:13242-13254

[142] **Rombach Robin,** Blattmann Andreas, Lorenz Dominik, Esser Patrick, Ommer Björn (2022) Highresolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684-10695

[143] **Hochreiter Sepp,** Schmidhuber Jürgen (1997) Long short-term memory. *Neural computation* **9**:1735-1780

[144] **Chung Junyoung,** Gulcehre Caglar, Cho KyungHyun, Bengio Yoshua (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv*

Peer reviews

Reviewer #1 (Public review):

Bredenberg et al. aim to model some of the visual and neural effects of psychedelics via the Wake-Sleep algorithm. This is an interesting study with findings that challenge certain mainstream ideas in psychedelic neuroscience.

While some of my concerns have been addressed in revision, I am still not convinced that this model applies to 5-HT2A hallucinogens, as opposed to a pharmacologically distinct hallucinogen. I think it is important to justify which class of hallucinogens this model applies to and distinguish it from other hallucinogens. While some researchers tend to group several hallucinogens together (e.g., 5-HT2A agonists, NMDA antagonists, kappa-opioids agonists), I'm not convinced this is warranted, when they have distinct subjective and cognitive effects (including quite different visual distortions, and again I point out that the kappa-opioid agonist salvinorin A, which is referred to as an "oneirogen," has been described as particularly dream-like, perhaps more so than 5-HT2A hallucinogens), as well as some differences in therapeutic outcomes (ketamine seems to not have as persisting of therapeutic effects, and kappa-opioid agonist have yet to be shown to be therapeutic). Their use patterns highlight this (e.g., 5-HT2A drugs are used less in non-festival/rave social settings compared to NMDA drugs like ketamine, which can be used frequently enough to the point of abuse; kappa-opioid agonists have quite mixed effects in terms of pleasurable outcomes, thereby rarely being used/abused and almost never to my knowledge being used recreationally).

In sum, more is needed to justify the claim that this work applies to 5-HT2A drugs in particular.

https://doi.org/10.7554/eLife.105968.2.sa2

Reviewer #2 (Public review):

This work is a nice contribution to the literature in articulating a specific, testable theory of how psychedelics act to generate hallucinations and plasticity.

I believe my concerns from the first round of review have been addressed in this version.

https://doi.org/10.7554/eLife.105968.2.sa1

Author response:

The following is the authors' response to the original reviews.

First, we thank the reviewers for the valuable and constructive reviews. Thanks to these, we believe the article has been considerably improved. We have organized our response to address points that are relevant to both reviewers first, after which we address the unique concerns of each individual reviewer separately. We briefly paraphrase each concern and provide comments for clarification, outlining the precise changes that we have made to the text.

> ***Common Concerns (R1 & R2):***
>
> *Can you clarify how NREM and REM sleep relate to the oneirogen hypothesis?*

Within the submission draft we tried to stay agnostic as to whether mechanistically similar replay events occur during NREM or REM sleep; however, upon a more thorough literature review, we think that there is moderately greater evidence in favor of Wake-Sleep-type replay occurring during REM sleep which is related to classical psychedelic drug mechanisms of action.

First, we should clarify that replay has been observed during both REM and NREM sleep, and dreams have been documented during both sleep stages, though the characteristics of dreams differ across stages, with NREM dreams being more closely tied to recent episodic experience and REM dreams being more bizarre/hallucinatory (see Stickgold et al., 2001 for a review). Replay during sleep has been studied most thoroughly during NREM sharp-wave ripple events, in which significant cortical-hippocampal coupling has been observed (Ji & Wilson, 2007). However, it is critical to note that the quantification methods used to identify replay events in the hippocampal literature usually focus on identifying what we term 'episodic replay,' which involves a near-identical recapitulation of neural trajectories that were recently experienced during waking experimental recordings (Tingley & Peyrach, 2020). In contrast, our model focuses on 'generative replay,' where one expects only a statistically similar reproduction of neural activity, without any particular bias towards recent or experimentally controlled experience. This latter form of replay may look closer to the 'reactivation' observed in cortex by many studies (e.g. Nguyen et al., 2024), where correlation structures of neural activity similar to those observed during stimulus-driven experience are recapitulated. Under experimental conditions in which an animal is experiencing highly stereotyped activity repeatedly, over extended periods of time, these two forms of replay may be difficult to dissociate.

Interestingly, though NREM replay has been shown to couple hippocampal and cortical activity, a similar study in waking animals administered psychedelics found hippocampal replay without any obvious coupling to cortical activity (Domenico et al., 2021). This could be because the coupling was not strong enough to produce full trajectories in the cortex (psychedelic administration did not increase 'alpha' enough), and that a causal manipulation of apical/basal influence in the cortex may be necessary to observe the increased coupling. Alternatively, as Reviewer 1 noted, it may be that psychedelics induce a form of

hippocampus-decoupled replay, as one would expect from the REM stage of a recently proposed complementary learning systems model (Singh et al., 2022).

Evidence in favor of a similarity between the mechanism of action of classical psychedelics and the mechanism of action of memory consolidation/learning during REM sleep is actually quite strong. In particular, studies have shown that REM sleep increases the activity of soma-targeting parvalbumin (PV) interneurons and decreases the activity of apical dendrite-targeting somatostatin (SOM) interneurons (Niethard et al., 2021), that this shift in balance is controlled by higher-order thalamic nuclei, and that this shift in balance is critical for synaptic consolidation of both monocular deprivation effects in early visual cortex (Zhou et al., 2020) and for the consolidation of auditory fear conditioning in the dorsal prefrontal cortex (Aime et al., 2022). These last studies were not discussed in our previous text–we have added them, in addition to a more nuanced description of the evidence connecting our model to NREM and REM replay.

Relevant modifications: Page 4, 1st paragraph; Page 11, 1st paragraph.

> *Can you explain how synaptic plasticity induced by psychedelics within your model relates to learning at a behavioral level?*

While the Wake-Sleep algorithm is a useful model for unsupervised statistical learning, it is not a model of reward or fear-based conditioning, which likely occur via different mechanisms in the brain (e.g. dopamine-dependent reinforcement learning or serotonin-dependent emotional learning). The Wake-Sleep algorithm is a 'normative plasticity algorithm,' that connects synaptic plasticity to the formation of structured neural representations, but it is not the case that all synaptic plasticity induced by psychedelic administration within our model should induce beneficial learning effects. According to the Wake-Sleep algorithm, plasticity at apical synapses is enhanced during the Wake phase, and plasticity at basal synapses is enhanced during the Sleep phase; under the oneirogen hypothesis, hallucinatory conditions (increased 'alpha') cause an increase in plasticity at both apical and basal sites. Because neural activity is in a fundamentally aberrant state when 'alpha' is increased, there are no theoretical guarantees that plasticity will improve performance on any objective: psychedelic-induced plasticity within our model could perhaps better be thought of as 'noise' that may have a positive or negative effect depending on the context.

In particular, such 'noise' may be beneficial for individuals or networks whose synapses have become locked in a suboptimal local minimum. The addition of large amounts of random plasticity could allow a system to extricate itself from such local minima over subsequent learning (or with careful selection of stimuli during psychedelic experience), similar to simulated annealing optimization approaches. If our model were fully validated, this view of psychedelic-induced plasticity as 'noise' could have relevance for efforts to alleviate the adverse effects of PTSD, early life trauma, or sensory deprivation; it may also provide a cautionary note against repeated use of psychedelic drugs within a short time frame, as the plasticity changes induced by psychedelic administration under our model are not guaranteed to be good or useful in-and-of themselves without subsequent re-learning and compensation.

We should also note that we have deliberately avoided connecting the oneirogen hypothesis model to fear extinction experimental results that have been observed through recordings of the hippocampus or the amygdala (Bombardi & Giovanni, 2013; Jiang et al., 2009; Kelly et al., 2024; Tiwari et al., 2024). Both regions receive extensive innervation directly from serotonergic synapses originating in the dorsal raphe nucleus, which have been shown to play an important role in emotional learning (Lesch & Waider, 2012); because classical psychedelics may play a more direct role in modulating this serotonergic innervation, it is possible that fear conditioning results (in addition to the anxiolytic effects of psychedelics)

cannot be attributed to a shift in balance between apical and basal synapses induced by psychedelic administration. We have provided a more detailed review of these results in the text, as well as more clarity regarding their relation to our model.

Relevant modifications: Page 9, final paragraph; Page 12, final paragraph.

> ***Reviewer 1 Concerns:***
>
> *Is it reasonable to assign a scalar parameter 'alpha' to the effects of classical psychedelics? And is your proposed mechanism of action unique to classical psychedelics? E.g. Could this idea also apply to kappa opioid agonists, ketamine, or the neural mechanisms of hallucination disorders?*

We have clarified that within our model 'alpha' is a parameter that reflects the balance between apical and basal synapses in determining the activity of neurons in the network. For the sake of simplicity we used a single 'alpha' parameter, but realistically, each neuron would have its own 'alpha' parameter, and different layers or individual neurons could be affected differentially by the administration of any particular drug; therefore, our scalar 'alpha' value can be thought of as a mean parameter for all neurons, disregarding heterogeneity across individual neurons.

There are many different mechanisms that could theoretically affect this 'alpha' parameter, including: 5-HT2a receptor agonism, kappa opioid receptor binding, ketamine administration, or possibly the effects of genetic mutations underlying the pathophysiology of complex developmental hallucination disorders. We focused exclusively on 5-HT2a receptor agonism for this study because the mechanism is comparatively simple and extensively characterized, but similar mechanisms may well be responsible for the hallucinatory symptoms of a variety of drugs and disorders.

Relevant modifications: Page 4, first paragraph; Page 13, first paragraph.

> *Can you clarify the role of 5-HT2a receptor expression on interneurons within your model?*

While we mostly focused on the effects of 5-HT2a receptors on the apical dendrites of pyramidal neurons, these receptors are also expressed on soma-targeting parvalbumin (PV) interneurons. This expression on PV interneurons is consistent with our proposed psychedelic mechanism of action, because it could lead to a coordinated decrease in the influence of somatic and proximal dendritic inputs while increasing the influence of apical dendritic inputs. We have elaborated on this point, and moved the discussion earlier in the text.

Relevant modifications: Page 1, 1st paragraph; Page 4, 2nd paragraph.

> *Discussions of indigenous use of psychedelics over millenia may amount to over-romanticization.*

We ultimately decided to remove these discussions from the main text, as they had little bearing on the content of our work. Within the Ethics Declarations section we softened our claims from "millenia" to "centuries," as indigenous psychedelic use over this latter period of time is well-substantiated.

Relevant modifications: removed from introduction; modified Ethics Declarations

> *You isolate the 5-HT2a agonism as the mechanism of action underlying 'alpha' in your model, but there exist 5-HT2a agonists that do not have hallucinatory effects (e.g. lisuride). How do you explain this?*

Lisuride has much-reduced hallucinatory effects compared to other psychedelic drugs at clinical doses (though it does indeed induce hallucinations at high doses; Marona-Lewicka et al., 2002), and we should note that serotonin (5-HT) itself is pervasive in the cortex without inducing hallucinatory effects during natural function. Similarly, MDMA is a partial agonist for 5-HT2a receptors, but it has much-reduced perceptual hallucination effects relative to classical psychedelics (Green et al., 2003) in addition to many other effects not induced by classical psychedelics.

Therefore, while we argue that 5-HT2a agonism induces an increase in influence of apical dendritic compartments and a decrease in influence of basal/somatic compartments, and that this change induces hallucinations, we also note that there are many other factors that control whether or not hallucinations are ultimately produced, so that not all 5-HT2a agonists are hallucinogenic. There are two possible additional factors that could contribute to this phenomenon: 5-HT receptor binding affinity and cellular membrane permeability.

Importantly, many 5-HT2a receptor agonists are also 5-HT1a receptor agonists (e.g. serotonin itself and lisuride), while MDMA has also been shown to increase serotonin, norepinephrine, and dopamine release (Green et al., 2003). While 5-HT2a receptor agonism has been shown to reduce sensory stimulus responses (Michaiel et al., 2019), 5-HT1a receptor agonism inhibits spontaneous cortical activity (Azimi et al., 2020); thus one might expect the net effect of administering serotonin or a nonselective 5-HT receptor agonist to be widespread inhibition of a circuit, as has been observed in visual cortex (Azimi et al., 2020). Therefore, selective 5-HT2a agonism is critical for the induction of hallucinations according to our model, though any intervention that jointly excites pyramidal neurons' apical dendrites and inhibits their basal/somatic compartments across a broad enough area of cortex would be predicted to have a similar effect. Lisuride has a much higher binding affinity for 5-HT1a receptors than, for instance, LSD (Marona-Lewicka et al., 2002).

Secondly, it has recently been shown that both the head-twitch effect (a coarse behavioral readout of hallucinations in animals) and the plasticity effects of psychedelics are abolished when administering 5-HT2a agonists that are impermeable to the cellular membrane because of high polarity, and that these effects can be rescued by temporarily rendering the cellular membrane permeable (Vargas et al., 2023). This suggests that the critical hallucinatory effects of psychedelics (apical excitation according to our model) may be mediated by intracellular 5-HT2a receptors. Notably, serotonin itself is not membrane permeable in the cortex.

Therefore, either of these two properties could play a role in whether a given 5-HT2a agonist induces hallucinatory effects. We have provided an extended discussion of these nuances in our revision.

Relevant modifications: Page 1, paragraph 2.

> *Your model proposes that an increase in top-down influence on neural activity underlies the hallucinatory effects of psychedelics. How do you explain experimental results that show increases in bottom-up functional connectivity (either from early sensory areas or the thalamus)?*

Firstly, we should note that our proposed increase in top-down influence is a causal, biophysical property, not necessarily a statistical/correlative one. As such, we will stress that the best way to test our model is via direct intervention in cortical microcircuitry, as opposed to correlative approaches taken by most fMRI studies, which have shown mixed results with regard to this particular question. Correlative approaches can be misleading due to dense recurrent coupling in the system, and due to the coarse temporal and spatial resolution provided by noninvasive recording technologies (changes in statistical/functional connectivity do not necessarily correspond to changes in causal/mechanistic connectivity, i.e. correlation does not imply causation).

There are two experimental results that appear to contradict our hypothesis that deserve special consideration. The first shows an increase in directional thalamic influence on the distributed cortical networks after psychedelic administration (Preller et al., 2018). To explain this, we note that this study does not distinguish between lower-order sensory thalamic nuclei (e.g. the lateral and medial geniculate nuclei receiving visual and auditory stimuli respectively) and the higher-order thalamic nuclei that participate in thalamocortical connectivity loops (Whyte et al., 2024). Subsequent more fine-grained studies have noted an increase in influence of higher order thalamic nuclei on the cortex (Pizzi et al., 2023; Gaddis et al., 2022), and in fact extensive causal intervention research has shown that classical psychedelics (and 5-HT2a agonism) decrease the influence of incoming sensory stimuli on the activity of early sensory cortical areas, indicating decoupling from the sensory thalamus (Evarts et al., 1955; Azimi et al., 2020; Michaiel et al. 2019). The increased influence of higher-order thalamic nuclei is consistent with both the cortico-striatal-thalamo-cortical (CTSC) model of psychedelic action as well as the oneirogen hypothesis, since higher-order thalamic inputs modulate the apical dendrites of pyramidal neurons in cortex (Whyte et al., 2024).

The second experimental result notes that DMT induces traveling waves during resting state activity that propagate from early visual cortex to deeper cortical layers (Alamia et al., 2020). There are several possibilities that could explain this phenomenon: 1) it could be due to the aforementioned difficulties associated with directed functional connectivity analyses, 2) it could be due to a possible high binding affinity for DMT in the visual cortex relative to other brain areas, or 3) it could be due to increases in apical influence on activity caused by local recurrent connectivity within the visual cortex which, in the absence of sensory input, could lead to propagation of neural activity from the visual cortex to the rest of the brain. This last possibility is closest to the model proposed by (Ermentrout & Cowan, 1979), and which we believe would be best explained within our framework by a topographically connected recurrent network architecture trained on video data; a potentially fruitful direction for future research.

Relevant modifications: Page 9, paragraph 1; Page 10, final paragraph; Page 11, final paragraph.

> *Shouldn't the hallucinations generated by your model look more 'psychedelic,' like those produced by the DeepDream algorithm?*

We believe that the differences in hallucination visualization quality between our Wake-Sleep-trained models and DeepDream are mostly due to differences in the scale and power of the models used across these two studies. We are confident that with more resources (and potentially theoretical innovations to improve the Wake-Sleep algorithm's performance) the produced hallucination visualizations could become more realistic.

We note that more powerful generative models trained with backpropagation are able to produce surreal images of comparable quality (Rezende et al., 2014; Goodfellow et al., 2020; Vahdat & Kautz, 2020), though these have not yet been used as a model of psychedelic hallucinations. However, the DeepDream model operates on top of large pretrained image processing models, and does not provide an biologically mechanistic/testable interpretation of its hallucination effects. When training smaller models with a local synaptic plasticity rule (as opposed to backpropagation), the hallucination effects are less visually striking due to the reduced quality of our trained generative model, though they are still strongly tied to the statistics of sensory inputs, as quantified by our correlation similarity metric (Fig. 5b).

To demonstrate that our proposed hallucination mechanism is capable of producing more complex hallucinations in larger, more powerful models, we employed our same hallucination generation mechanism in a pretrained Very Deep Variational Autoencoder (VDVAE) (Child et al., 2021), which is a hierarchical variational autoencoder with a nearly

identical structure compared to our Wake-Sleep-trained networks, with both a bottom-up inference pathway and a top-down generative pathway that maps cleanly onto our multicompartmental neuron model. VDVAEs are trained on the same objective function as our Wake-Sleep-trained networks, but using the backpropagation algorithm. The VDVAE models were able to generate much more complex hallucinations (emergence of complex geometric patterns, smooth deformations of objects and faces), whose complexity arguably exceeds those produced by the DeepDream algorithm. Therefore while the VDVAEs are less biologically realistic (they do not learn via local synaptic plasticity), they function as a valuable high-level model of hallucination generation that complements our Wake-Sleep-trained approach. As further validation, we were also able to replicate our key results and testable predictions with these models.

Relevant modifications: Results section "Modeling hallucinations in large-scale pretrained networks"; Figure 6, S7, S8; Page 12, paragraph 3; Methods section "Generating hallucinations in hierarchical variational autoencoders."

> *Your model assumes domination by entirely bottom-up activity during the 'wake' phase, and domination entirely by top-down activity during 'sleep,' despite experimental evidence indicating that a mixture of top-down and bottom-up inputs influence neural activity during both stages in the brain. How do you explain this?*

Our use of the Wake-Sleep algorithm, in which top-down inputs (Sleep) or bottom-up inputs (Wake) dominate network activity is an over-simplification made within our model for computational and theoretical reasons. Models that receive a mixture of top-down and bottom-up inputs during 'Wake' activity do exist (in particular the closely related Boltzmann machine (Ackley et al., 1985)), but these models are considerably more computationally costly to train due to a need to run extensive recurrent network relaxation dynamics for each input stimulus. Further, these models do not generalize as cleanly to processing temporal inputs. For this reason, we focused on the Wake-Sleep algorithm, at the cost of some biological realism, though we note that our model should certainly be extended to support mixed apical-basal waking regimes. We have added a discussion of this in our 'Model Limitations' section.

Relevant modifications: Page 12, paragraph 4.

> *Your model proposes that 5-HT2a agonism enhances glutamatergic transmission, but this is not true in the hippocampus, which shows decreases in glutamate after psychedelic administration.*

We should note that our model suggests only compartment specific increases in glutamatergic transmission; as such, our model does not predict any particular directionality for measures of glutamatergic transmission that includes signaling at both apical and basal compartments in aggregate, as was measured in the provided study (Mason et al., 2020).

> *You claim that your model is consistent with the Entropic Brain theory, but you report increases in variance, not entropy. In fact, it has been shown that variance decreases while entropy increases under psychedelic administration. How do you explain this discrepancy?*

Unfortunately, 'entropy' and 'variance' are heavily overloaded terms in the noninvasive imaging literature, and the particularities of the method employed can exert a strong influence on the reported effects. The reduction in variance reported by (Carhart-Harris et al., 2016) is a very particular measure: they are reporting the variance of resting state synchronous activity, averaged across a functional subnetwork that spans many voxels; as such, the reduction in variance in this case is a reduction in broad, synchronous activity. We do not have any resting state synchronous activity in our network due to the simplified

nature of our model (particularly an absence of recurrent temporal dynamics), so we see no reduction in variance in our model due to these effects.

Other studies estimate 'entropy' or network state disorder via three different methods that we have been able to identify. 1) (Carhart-Harris et al., 2014) uses a different measure of variance: in this case, they subtract out synchronous activity within functional subnetworks, and calculate variability across units in the network. This measure reports increases in variance (Fig. 6), and is the closest measure to the one we employ in this study. 2) (Lebedev et al., 2016) uses sample entropy, which is a measure of temporal sequence predictability. It is specifically designed to disregard highly predictable signals, and so one might imagine that it is a measure that is robust to shared synchronous activity (e.g. resting state oscillations). 3) (Mediano et al., 2024) uses Lempel-Ziv complexity, which is, similar to sample entropy, a measure of sequence diversity; in this case the signal is binarized before calculation, which makes this method considerably different from ours. All three of the preceding methods report increases in sequence diversity, in agreement with our quantification method. Our strongest explanation for why the variance calculation in (Carhart-Harris et al., 2016) produces a variance reduction is therefore due to a reduction in low-rank synchronous activity in subnetworks during resting state.

As for whether the entropy increase is meaningful: we share Reviewer 1's concern that increases in entropy could simply be due to a higher degree of cognitive engagement during resting state recordings, due to the presence of sensory hallucinations or due to an inability to fall asleep. This could explain why entropy increases are much more minimal relative to non-hallucinating conditions during audiovisual task performance (Siegel et al., 2024; Mediano et al., 2024). However, we can say that our model is consistent with the Entropic Brain Theory without including any form of 'cognitive processing': we observe increases in variability during resting state in our model, but we observe highly similar distributions of activity when averaging over a wide variety of sensory stimulus presentations (Fig. 5b-c). This is because variability in our model is not due to unstructured noise: it corresponds to an exploration of network states that would ordinarily be visited by some stimulus. Therefore, when averaging across a wide variety of stimuli, the distribution of network states under hallucinating or non-hallucinating conditions should be highly similar.

One final point of clarification: here we are distinguishing Entropic Brain Theory from the REBUS model–the oneirogen hypothesis is consistent with the increase in entropy observed experimentally, but in our model this entropy increase is not due to increased influence of bottom-up inputs (it is due instead to an increase in top-down influence). Therefore, one could view the oneirogen hypothesis as consistent with EBT, but inconsistent with REBUS.

Relevant modifications: Page 10, paragraph 1.

> *You relate your plasticity rule to behavioral-timescale plasticity (BTSP) in the hippocampus, but plasticity has been shown to be reduced in the hippocampus after psychedelic administration. Could you elaborate on this connection?*

When we were establishing a connection between our 'Wake-Sleep' plasticity rule and BTSP learning, the intended connection was exclusively to the mathematical form of the plasticity rule, in which activity in the apical dendrites of pyramidal neurons functions as an instructive signal for plasticity in basal synapses (and vice versa): we will clarify this in the text. Similarly, we point out that such a plasticity rule tends to result in correlated tuning between apical and basal dendritic compartments, which has been observed in hippocampus and cortex: this is intended as a sanity check of our mapping of the Wake-Sleep algorithm to cortical microcircuitry, and has limited further bearing on the effects of psychedelics specifically.

Reduction in plasticity in the hippocampus after psychedelic administration could be due to a complementary learning systems-type model, in which the hippocampus becomes partly decoupled from the cortex during REM sleep (Singh et al., 2022); were this to be the case, it would not be incompatible with our model, which is mostly focused on the cortex. Notably, potentiating 5HT-2a receptors in the ventral hippocampus does not induce the head-twitch response, though it does produce anxiolytic effects (Tiwari et al., 2024), indicating that the hallucinatory and anxiolytic effects of classical psychedelics may be partly decoupled.

> ***Reviewer 2 Concerns:***
>
> *Could you provide visualizations of the 'ripple' phenomenon that you're referring to?*

In our revised submission, 'ripple' phenomena are now visible in two places: Fig 2c-d, and Fig 6 (rows 2 and 3). Because the VDVAE models used to generate Figure 6 produce higher quality generated images, the ripples appearing in these plots are likely more prototypical, but it is not easy to evaluate the quality of these visualizations relative to subjective hallucination phenomena.

> *Could you provide a more nuanced description of alternative roles for top-down feedback, beyond being used exclusively for learning as depicted in your model?*

For the sake of simplicity, we only treat top-down inputs in our model as a source of an instructive teaching signal, the originator of generative replay events during the Sleep phase, and as the mechanism of hallucination generation. However, as discussed in a response to a previous question, in the cortex pyramidal neurons receive and respond to a mixture of top-down and bottom-up processing.

There are a variety of theories for what role top-down inputs could play in determining network activity. To name several, top-down input could function as: 1) a denoising/pattern completion signal (Kadkhodaie & Simoncelli, 2021), 2) a feedback control signal (Podlaski & Machens, 2020), 3) an attention signal (Lindsay, 2020), 4) ordinary inputs for dynamic recurrent processing that play no specialized role distinct from bottom-up or lateral inputs except to provide inputs from higher-order association areas or other sensory modalities (Kar et al., 2019; Tugsbayar et al., 2025). Though our model does not include these features, they are perfectly consistent with our approach.

In particular, denoising/pattern completion signals in the predictive coding framework (closely related to the Wake-Sleep algorithm) also play a role as an instructive learning signal (Salvatori et al., 2021); and top-down control signals can play a similar role in some models (Gilra & Gerstner, 2017; Meulemans et al., 2021). Thus, options 1 and 2 are heavily overlapping with our approach, and are a natural consequence of many biologically plausible learning algorithms that minimize a variational free energy loss (Rao & Ballard, 1997; Ackley et al., 1985). Similarly, top-down attentional signals can exist alongside top-down learning signals, and some models have argued that such signals can be heavily overlapping or mutually interchangeable (Roelfsema & van Ooyen, 2005). Lastly, generic recurrent connectivity (from any source) can be incorporated into the Wake-Sleep algorithm (Dayan & Hinton, 1996), though we avoided doing this in the present study due to an absence of empirical architecture exploration in the literature and the computational complexity associated with training on time series data.

To conclude, there are a variety of alternative functions proposed for top-down inputs onto pyramidal neurons in the cortex, and we view these additional features as mutually compatible with our approach; for simplicity we did not include them in our Wake-Sleep-trained model, but we believe that these features are unlikely to interfere with our testable predictions or empirical results. In fact, the pretrained VDVAE models that we worked with

do include top-down influence during the Wake-stage inference process, and these models recapitulated our key results and testable predictions (Fig. S8).

Relevant modifications: Fig. S8; Page 12, paragraph 4.

https://doi.org/10.7554/eLife.105968.2.sa0