TYPE Original Research
PUBLISHED 03 January 2024
DOI 10.3389/fnhum.2023.1159821

Modelling phenomenological differences in aetiologically distinct visual hallucinations using deep neural networks

Keisuke Suzuki[1,2,3]*, Anil K. Seth[1,2,4] and David J. Schwartzman[1,2]

[1]Sussex Centre for Consciousness Science, University of Sussex, Brighton, United Kingdom, [2]Department of Informatics, University of Sussex, Brighton, United Kingdom, [3]Center for Human Nature, Artificial Intelligence and Neuroscience (CHAIN), Hokkaido University, Sapporo, Japan, [4]Program on Brain, Mind, and Consciousness, Canadian Institute for Advanced Research, Toronto, ON, Canada

Visual hallucinations (VHs) are perceptions of objects or events in the absence of the sensory stimulation that would normally support such perceptions. Although all VHs share this core characteristic, there are substantial phenomenological differences between VHs that have different aetiologies, such as those arising from Neurodegenerative conditions, visual loss, or psychedelic compounds. Here, we examine the potential mechanistic basis of these differences by leveraging recent advances in visualising the learned representations of a coupled classifier and generative deep neural network—an approach we call 'computational (neuro)phenomenology'. Examining three aetiologically distinct populations in which VHs occur—Neurodegenerative conditions (Parkinson's Disease and Lewy Body Dementia), visual loss (Charles Bonnet Syndrome, CBS), and psychedelics—we identified three dimensions relevant to distinguishing these classes of VHs: realism (veridicality), dependence on sensory input (spontaneity), and complexity. By selectively tuning the parameters of the visualisation algorithm to reflect influence along each of these phenomenological dimensions we were able to generate 'synthetic VHs' that were characteristic of the VHs experienced by each aetiology. We verified the validity of this approach experimentally in two studies that examined the phenomenology of VHs in Neurodegenerative and CBS patients, and in people with recent psychedelic experience. These studies confirmed the existence of phenomenological differences across these three dimensions between groups, and crucially, found that the appropriate synthetic VHs were rated as being representative of each group's hallucinatory phenomenology. Together, our findings highlight the phenomenological diversity of VHs associated with distinct causal factors and demonstrate how a neural network model of visual phenomenology can successfully capture the distinctive visual characteristics of hallucinatory experience.

KEYWORDS

visual hallucinations, machine learning, Parkinson's disease, Lewy Body Dementia, Charles Bonnet Syndrome, psychedelics, phenomenology, computational neurophenomenology

Check for updates

OPEN ACCESS

EDITED BY
Daniel Collerton,
Newcastle University, United Kingdom

REVIEWED BY
Shigetoshi Nara,
Okayama University, Japan
Thomas Parr,
University of Oxford, United Kingdom

*CORRESPONDENCE
Keisuke Suzuki
✉ ksk@chain.hokudai.ac.jp

RECEIVED 06 February 2023
ACCEPTED 11 September 2023
PUBLISHED 03 January 2024

CITATION
Suzuki K, Seth AK and Schwartzman DJ (2024) Modelling phenomenological differences in aetiologically distinct visual hallucinations using deep neural networks.
Front. Hum. Neurosci. 17:1159821.
doi: 10.3389/fnhum.2023.1159821

1 Introduction

Visual hallucinations (VHs) are perceptual experiences that occur in the absence of the sensory stimulation that would normally accompany such experiences. The decoupling of perceptual experience from sensation, together with the varied nature of VHs across different aetiologies, provides an opportunity to investigate the computational processes and neural mechanisms that may underlie them, with implications for those that may underlie visual experience in general.

VHs are commonly experienced as a result of Neurodegenerative disorders that share a common pathological accumulation of insoluble α-synuclein protein in neurons, nerve fibres or glial cells, which include Parkinson's disease (PD; Fénelon et al., 2000; Barnes and David, 2001; Mosimann et al., 2006; Boubert and Barnes, 2015) and Lewy Body Dementia (LBD; Harding et al., 2002). We use the term Neurodegenerative to specifically refer to PD and LBD populations in which VHs occur. They may also occur as the result of marked visual impairment, as in Charles Bonnet Syndrome (CBS; Schultz and Melzack, 1991; Teunisse et al., 1996; Ffytche, 2005; Abbott et al., 2007; Yacoub and Ferrucci, 2011; Nair et al., 2015). VHs also occur in psychiatric conditions such as schizophrenia, although they are less frequent than auditory hallucinations (Bauer et al., 2011; Waters et al., 2014). VHs are also common features of psychedelic states, such as those produced by the ingestion of classical hallucinogens such as LSD, psilocybin and DMT (Nichols, 2016). The VHs associated with each of these aetiologies have distinctive phenomenological (experiential) characteristics, raising the question of what aspects of the underlying neurophysiology and dynamics are responsible for these differences.

A popular approach to understanding the basis of VHs is through the lens of 'predictive processing' (PP) theories of perception and brain function (Rao and Ballard, 1999; Friston, 2005; Hohwy and Seth, 2020). These theories view perception as an iterative process in which the brain is always trying to optimise an evolving 'best guess' (Bayesian posterior belief) about the most likely causes of the sensory inputs it encounters (Lee and Mumford, 2003; Knill and Pouget, 2004; Clark, 2013). In standard PP, this is achieved by the reciprocal exchange of top-down perceptual predictions and bottom-up (sensory) prediction error signals in a continual process of prediction error minimisation, which approximates Bayesian inference on the causes of sensory signals. Importantly, the outcome of this (approximate) inference process depends on the relative influence of (top-down) predictions and (bottom-up) prediction errors, a balance which is mediated by the estimated precision (informally, the 'reliability') of sensory signals relative to perceptual predictions (Yuille and Kersten, 2006; Fletcher and Frith, 2009; Friston and Kiebel, 2009; Zarkali et al., 2019). Within this framework, VHs can be broadly understood as resulting from aberrant inference in which this balance is disrupted in some way (Friston, 2005; Powers et al., 2016; Corlett et al., 2019; Zarkali et al., 2019).

Such aberrant inference can take many different forms. Some researchers have interpreted Neurodegenerative VHs as resulting from overly strong perceptual predictions which overwhelm sensory prediction error signals (Powers et al., 2016; O'Callaghan et al., 2017). Others have focused on psychedelic VHs and adopted a hierarchical perspective, suggesting that psychedelic hallucinations occur due to a relaxation of high-level perceptual predictions (or Bayesian 'priors'), which has the effect of increasing the influence of bottom-up signalling on perceptual inference (Alonso et al., 2015; Swanson, 2018; Timmermann et al., 2018; Carhart-Harris and Friston, 2019). However, despite accumulating empirical evidence linking perceptual experience to inference and prediction error minimization in both normal (Hardstone et al., 2021) and hallucinatory (Powers et al., 2016; Carhart-Harris and Friston, 2019; Schartner and Timmermann, 2020) perception, the computational basis of phenomenological differences between different kinds of VH has remained unclear.

To shed light on this question, we take a computational neurophenomenology approach (Suzuki et al., 2017; Ramstead et al., 2022). Neurophenomenology emphasises the importance of capturing first-person descriptions of phenomena of interest (such as VHs) that are amenable to neurocognitive methodologies (Lutz, 2002; Lutz et al., 2002; Gould et al., 2014). This approach becomes computational neurophenomenology when computational modelling is used to simulate specific properties of perceptual experience (rather than, for example, the functions associated with perception, such as classification or discrimination), and where the computational models used have useful interpretations with respect to theories of perception or neurophysiological mechanisms [see also phenomenological approaches in robotics, e.g., Tani (2016)]. We highlight that our implementation of this approach is not designed to map directly between the pathological mechanisms and variations in hallucinatory experience observed in different aetiologies.

1.1 Phenomenological variation in hallucinatory experience

Whilst there are substantial phenomenological differences in VHs both across and within different aetiological categories, we identified three dimensions which broadly characterise variations in VHs arising from Neurodegenerative, CBS and psychedelic origins. These dimensions are complexity, veridicality, and spontaneity.

1.1.1 Complexity

VHs can generally be categorised as being either simple (e.g., shapes, flashes or grid-like lattice patterns) or complex (e.g., well-defined recognisable forms, such as objects or people). Neurodegenerative VHs are typically complex, featuring false perceptions of family, other people, or animals (Frucht and Bernsohn, 2002; Mosimann et al., 2006; Papapetropoulos et al., 2008). In contrast, the most commonly reported class of VH in CBS are simple (Ffytche and Howard, 1999; Santhouse et al., 2000; Abbott et al., 2007), with complex VHs being reported less frequently (Schultz and Melzack, 1991; Teunisse et al., 1996). Psychedelic VHs also vary in their complexity. VHs arising from low doses are usually associated with visual distortion and/or simple VHs, such as brightly coloured geometric 'form constants' including lattices, cobwebs, tunnels and spirals (Kluver, 1926; Díaz, 2010; Studerus et al., 2011; Nichols, 2016), with complex VHs not as frequently reported (Studerus et al., 2011; Kometer et al., 2013; Schmid et al., 2015). At higher doses, complex VHs are more likely to occur, including fully formed VHs comprising visual scenes with elaborate structural content such as landscapes, cities, and galaxies, as well as specific forms including human figures and animals (Strassman et al., 1994; Shanon, 2002; Studerus et al., 2011; Kometer et al., 2013; Liechti, 2017). The rarity of complex psychedelic VHs of this sort may reflect either their relative rarity and/or the difficulties inherent in providing subjective reports in high dose situations (Studerus et al., 2011; Kometer et al., 2013; Kometer and Vollenweider, 2018).

1.1.2 Veridicality

Neurodegenerative and CBS complex VHs typically display a high degree of perceptual veridicality. That is, they are reported as being similar in visual quality to normal perceptual experience (Schultz and Melzack, 1991; Teunisse et al., 1996; Barnes and David, 2001; Mosimann et al., 2006; Boubert and Barnes, 2015). There are some exceptions: for example, complex VHs in CBS are sometimes described as distorted or cartoon-like, however, the majority are described as vivid and life-like (Schultz and Melzack, 1991; Ffytche and Howard, 1999; Santhouse et al., 2000; Ffytche, 2005; Abbott et al., 2007). In contrast, psychedelic complex VHs are typically reported to have a lower degree of veridicality, which may not be surprising thanks to their dream-like qualities (for a review, see Sanz et al., 2018) and the prominence of visual distortion, unrealistic colours, patterns and kaleidoscopic imagery (Studerus et al., 2011; Carhart-Harris et al., 2016; Preller and Vollenweider, 2018; Timmermann et al., 2018). Indeed, some have suggested that the alterations in visual perception induced by psychedelics, such as psilocybin, rarely represent true hallucinations because, at least at moderate doses, they can be readily distinguished from real perceptions (Preller and Vollenweider, 2018).

1.1.3 Spontaneity

Neurodegenerative and CBS complex VHs typically occur spontaneously, that is, they are not experienced as being transformations of existing content within a perceived scene (Schultz and Melzack, 1991; Teunisse et al., 1996; Frucht and Bernsohn, 2002; Mosimann et al., 2006; Papapetropoulos et al., 2008). Spontaneous VHs, therefore, correspond closely to the folk-psychological idea of hallucination as perception in the absence of sensory stimulation. In contrast, anecdotal reports of psychedelic complex VHs describe them as frequently being driven by visual 'seeds' from within a perceived scene, bearing similarity to the everyday phenomenon of 'pareidolia'—or 'seeing patterns in noise' (Kometer and Vollenweider, 2018; Preller and Vollenweider, 2018; Swanson, 2018).

From this short review, it is possible to broadly characterise each aetiologically distinct category of VH in terms of complexity, veridicality, and spontaneity. Neurodegenerative VHs typically display high veridicality, are spontaneous, and are mainly complex in nature. CBS VHs also typically display high veridicality, and are also spontaneous, but can occur in both simple and complex forms. Psychedelic VHs generally display lower veridicality compared to Neurodegenerative and CBS VHs, they tend to not be spontaneous and can occur in both simple and complex forms. We used these phenomenological profiles to tune the parameters used to generate synthetic VHs representative of each category.

1.2 Simulating dimensions of hallucinatory phenomenology

To simulate these specific aspects of hallucinatory phenomenology, we used the coupled neural network architecture of Nguyen et al. (2016), which combines pre-trained classifier (DCNN) and generative (DGN) networks, to generate synthetic snapshots of aetiologically distinct VHs (see Materials and methods). Intuitively, within this architecture, the image fed into the model can be viewed as analogous to visual input and the synthetic image produced by the model, the resulting perceptual experience.

1.2.1 Veridicality

DCNNs trained for natural image recognition are highly complex, with many parameters and nodes, such that their analysis requires innovative visualisation methods. A popular solution to this challenge has been to use algorithms that visualise the information processed by a target neuron (or group of neurons) within a DCNN. Classical Activation Maximisation (ClassicalAM) is one such visualisation technique (Erhan et al., 2009; Mahendran and Vedaldi, 2014; Simonyan et al., 2014). The core idea behind ClassicalAM is simple: generate an input image that maximises the activation of a target neuron(s) of interest. Instead of updating the weights between the neurons in the network as occurs during training, ClassicalAM modifies the image in a way that maximises the activity of the target neuron (see Supplementary material 1.1.2). This is done by defining an error function to return larger errors when the activation of the neuron is low and smaller errors when the activation is high. Repeating this process (iterating) gradually changes the image rather than altering the network to match the features of the image with what is represented by the target neuron. A slight modification of this approach is the well-known Deep-Dream visualisation algorithm, which applies ClassicalAM across a user-defined layer of the DCNN instead of a single target neuron (Mordvintsev et al., 2015).

Previous studies using ClassicalAM have shown that it typically produces unrealistic, uninterpretable images (Erhan et al., 2009; Mahendran and Vedaldi, 2014; Simonyan et al., 2014). This is due to the vast set of possible images that may excite a target neuron, making it likely that the image produced will not resemble the natural images that the neuron has learned to respond to.

To produce more 'human-interpretable' images, Nguyen et al. (2016) developed a new type of AM, called GenerativeAM, which combined a pre-trained Deep Generator Network (DGN) capable of generating realistic synthetic images with the same DCNN as used in ClassicalAM (Figure 1). For each iteration, GenerativeAM uses the learned natural image prior of the DGN to generate a new output image that maximises the activity of a target neuron within the DCNN, instead of updating the input image directly as in ClassicalAM (see Supplementary material 1.1.2). Using this approach Nguyen et al. (2016) found that GenerativeAM produced output images that were much more realistic than those produced by ClassicalAM.

To simulate differences in the veridicality of aetiologically distinct VHs, we generated synthetic VHs both with (GenerativeAM; high veridicality), and without (ClassicalAM; low veridicality), the natural image prior provided by the DGN. We reasoned that because GenerativeAM can produce realistic-looking, interpretable outputs it could be used to simulate the reported high veridicality of (complex and simple) Neurodegenerative and CBS VHs (Nguyen et al., 2016). In contrast, our previous work using ClassicalAM demonstrated its suitability to simulate the reduced veridicality of both simple and complex psychedelic VHs (Figure 1, top; Suzuki et al., 2017).

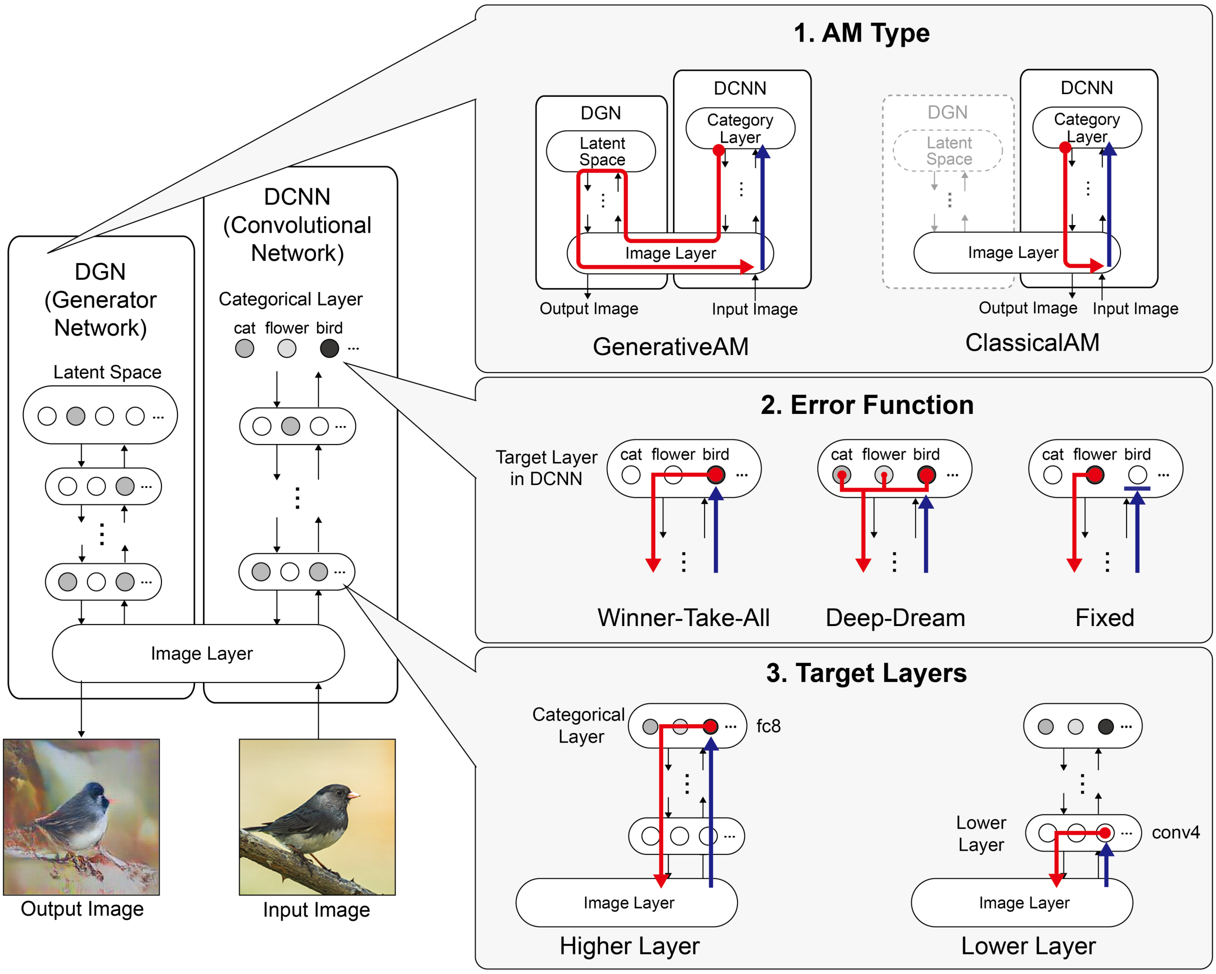

FIGURE 1
Schematic of the model architecture adapted from Nguyen et al. (2016). Left: The model consists of a DGN (left) that constrains the network towards producing realistic images that maximally activate a target neuron within the DCNN (right). Right: the three independent manipulations that we applied to simulate specific forms of hallucinatory phenomenology. AM Type (top), we used two forms of AM, that either iterated through just the DCNN (ClassicalAM) or through both the DCNN and DGN (GenerativeAM). Error Function (middle) returns errors at the terminated layer for either a single predetermined neuron (Fixed), all neurons activated by the input image (Deep-Dream) or the single neuron that is maximally activated by the input image (Winner-Take-All). Target Layers (bottom) allows the selection of the specific layer that the AM process terminates at within the DCNN and returns errors from this layer. Note that these three manipulations can be applied orthogonally.

For both types of AM, in order for the input image to be altered, the processes of the network must be repeated (iterated) many times. With each iteration the image is altered further. Therefore, generating synthetic VHs using increasing numbers of iterations provides an additional inbuilt method of manipulating veridicality, as the veridicality of synthetic VHs decreases with increasing iteration number. We leveraged this inherent feature of the networks to better capture natural variations in the veridicality of hallucinatory content within a specific hallucinatory population.

1.2.2 Spontaneity

A key challenge when designing network visualisation techniques centres on the parameters of the error function used to select target neuron(s). ClassicalAM and GenerativeAM both use a Fixed error function that returns an error for a pre-defined (randomly selected) target neuron, whilst all other neurons return zero errors (Figure 1, middle). Consequently, the output image is optimised to activate the selected target neuron only (usually representing a particular object category thanks to the DCNN/DGN training regime) regardless of the content of the input image. In contrast, the Deep-Dream error function, used in our previous simulations of hallucinatory phenomenology (Mordvintsev et al., 2015; Suzuki et al., 2017), returns errors for all the neurons within a selected layer of the DCNN that were activated by the visual features of the input image. Therefore, in this case the output image is dependent on the visual features contained within the input image and is modified by a dynamic interplay between these features and the neurons that are activated within the user-defined layer of the DCNN.

To simulate the reported differences in the spontaneity of VHs we therefore selectively altered the error function used to select the target neuron(s) within the DCNN. To simulate spontaneous VHs we used an error function that is not dependent on the input image (Fixed), and to simulate VHs that are driven by visual 'seeds' from within a perceived scene we used an error function that is dependent on the input image (Deep-Dream) to represent the apparent dependency between visual input and hallucinatory content (i.e., the lower spontaneity) reported in psychedelic VHs.

Another reason we selected this approach is that it relates to the theorised disruption between (top–down) predictions and (bottom-up) prediction errors, which has been proposed to underlie VHs within PP accounts of perception (Alonso et al., 2015; Swanson, 2018; Timmermann et al., 2018; Carhart-Harris and Friston, 2019). Within this framework, the use of a predetermined target neuron by the Fixed error function can be viewed as simulating the overly strong perceptual predictions thought to be the cause of Neurodegenerative/CBS VHs (Powers et al., 2016; O'Callaghan et al., 2017). The Deep-Dream error function can be viewed as simulating the increased influence of bottom-up signalling on perceptual inference thought to underlie psychedelic VHs, because in this case the selected target neurons are dependent on, and driven by, the properties of the input image.

1.2.3 Complexity

Based on our previous research (Suzuki et al., 2017), we reasoned that simple hallucinatory phenomenology could be simulated by restricting the layer that AM terminates to a lower layer of the DCNN, resulting in an overemphasis of low-level visual features extracted by these layers during training (Figure 1, bottom). Together with literature describing the characteristics of psychedelic VHs (Kluver, 1926; Díaz, 2010; Studerus et al., 2011; Nichols, 2016), and experimentation visualising activity within different lower layers, we found that lower layer conv4 produced synthetic VHs that were most representative of simple psychedelic VHs—for examples of how altering the lower layer AM terminates at affects image generation see Mahendran and Vedaldi (2016). In contrast, to simulate complex hallucinatory phenomenology, we restricted the layer that AM terminates to the highest (categorical) layer of the DCNN.

Finally, to provide a benchmark of the Nguyen et al. (2016) model's performance against which the above simulations of hallucinatory phenomenology could be compared, we produced simulations of non-hallucinatory (veridical) perceptual phenomenology. Normal perceptual experience is naturally characterised by high veridicality and a close dependence between visual input and perceptual experience. To mimic these features, we used GenerativeAM and developed a new class of error function, which we call Winner-Take-All (Figure 1, middle). In Winner-Take-All, as with Deep-Dream, the user selects which layer of the DCNN the error function is to operate so that the target neuron is selected by the visual features of the input image, however, only a single neuron with the highest activation to the input is selected as the target neuron, whilst the errors of all the other neurons are set to zero, as with Fixed. In this way Winner-Take-All mimics the balance between sensory signals and perceptual predictions associated with normal perceptual experience. The idea of this benchmark is to create an image set using the same network architecture when it is not tuned to model a VH of any kind.

Figure 1 summarises the manipulations to AM: type of AM: (GenerativeAM or ClassicalAM), error functions (Winner-Take-All, Deep-Dream, or Fixed), and user-selected target layer for AM to terminate within the DCNN [Higher (fc8) or Lower (conv4)].

Figure 2 provides illustrative examples of the effects of applying these three orthogonal manipulations on the output of the model. First, generating images with (GenerativeAM) or without (ClassicalAM) the natural image prior of the DGN produces output images with high and low veridicality, respectively. Second, selectively applying different Error Functions (Deep-Dream, Winner-Take-All, Fixed) alters the dependency between the input and output image, modulating spontaneity. Third, modifying the target layer at which AM terminates between a Higher Layer (fc8) and Lower Layer (conv4) alters the complexity of the output image.

Using the above approach, we set out to test if the coupled DCNN-DGN neural network architecture of Nguyen et al. (2016) could be used to simulate three aspects of Neurodegenerative, CBS and psychedelic hallucinatory phenomenology: veridicality, spontaneity, and complexity. Summarising, we simulated these three phenomenological dimensions by the inclusion or omission of the natural image prior of the DGN (veridicality), altering the error function used to select the target neuron(s) within the DCNN (spontaneity), and restricting the level within the DCNN that AM terminates to either the lower (conv4) or higher layer (fc8; complexity).

1.3 Do synthetic VHs provide an accurate representation of hallucinatory experience?

To draw conclusions about the validity of the synthetic VHs produced by this approach, it is important to establish the extent to which they match the subjective experience associated with each group's hallucinatory experience. To investigate this question, we performed two additional experimental studies. In the first in-person study we developed a semi-structured interview that was used to enquire about the visual phenomenology associated with CBS and Neurodegenerative VHs and how closely these reports matched the appropriate synthetic VHs. The second online study recruited participants with recent psychedelic experience to assess how closely a recalled psychedelic experience matched examples of the corresponding synthetic VHs. Both studies enquired about the general phenomenology, including specific questions regarding the complexity, veridicality, and spontaneity of their VHs and critically, also asked participants to directly rate the visual similarity between the appropriate synthetic VHs and their hallucinatory experiences.

2 Materials and methods

2.1 DCNN-DGN model

Here we summarise the details of the coupled DCNN-DGN model architecture that we use without modifications, taken from Nguyen et al. (2016) (Figure 3). Modelling distinct classes of VH was instead achieved by applying the following modifications to the visualisation algorithm: (1) we added a function that allowed us to switch between the GenerativeAM and ClassicalAM visualisation algorithms. (2) we added a function that allowed us to use different error functions when optimising images for both forms of AM. (3) We modified both forms of AM so that they could be applied in a layer-specific manner to the DCNN.

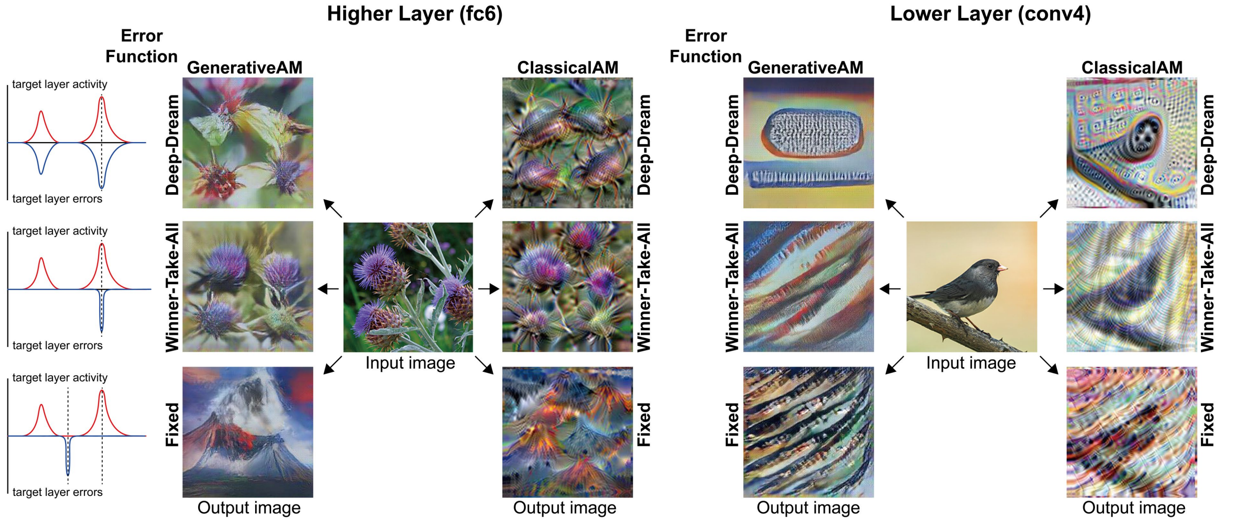

FIGURE 2
Examples of input image transformations based on the three orthogonal manipulations in our model. (i) the inclusion or omission of the natural image prior of the DGN: GenerativeAM vs. ClassicalAM (left and right image columns in each panel, corresponding to veridicality). (ii) altering the error function used to select the target neuron(s) within the DCNN: Fixed, Winner-Take-All or Deep-Dream (rows, corresponding to spontaneity). (iii) restricting the level within the DCNN that AM terminates: Higher Layer (fc8) vs. Lower Layer (conv4; left panel vs. right panel. Corresponding to complexity and to the level at which AM terminates). Fixing higher layers (left panel) tends to produce output images similar to more complex hallucinations, whilst fixing lower layers (right panel) tends to create output images better resembling simpler geometric hallucinations. Including the natural image prior (GenerativeAM) leads to output images with higher veridicality than without this prior (ClassicalAM). Turning to the error functions, the far-left column illustrates how each error function operates based on the activation of the target layer and the errors generated. In the Deep-Dream Error Function, errors are generated proportional to the activation of the target layer neurons. This may lead to the output image containing multiple categorical features. In the Winner-Take-All Error Function, only the neuron that is maximally activated by the input image remains activated. In the Fixed Error Function, the target neuron is pre-selected by the researcher and therefore only a single neuron returns the error. See Supplementary Figure S2 for more sample output images for all the combinations of model manipulations.

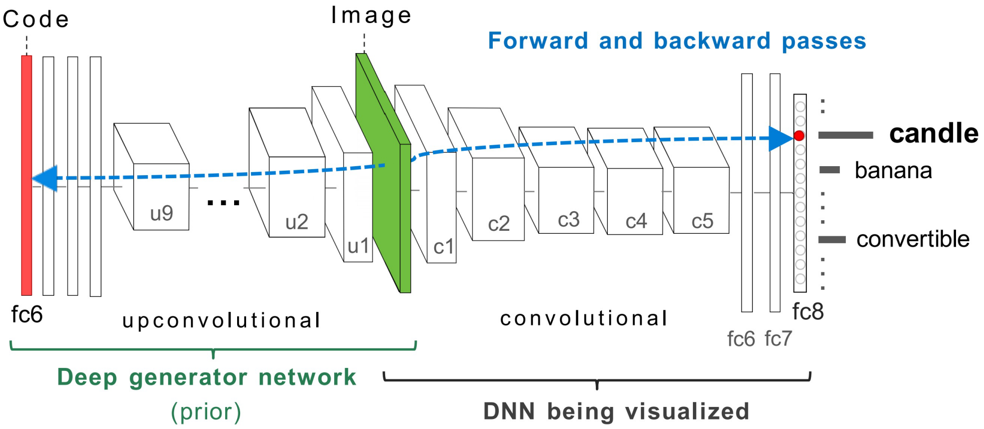

FIGURE 3
Coupled classifier (DCNN) and generative (DGN) neural network architecture used in this study, taken from Nguyen et al. (2016). The two networks, DCNN (labelled DNN) on the right, and DGN on the left are combined via the Image layer at the bottom of both networks (green). To synthesise a preferred input for a fixed target neuron (representing a 'candle') located in layer fc8 of the DCNN, the latent vector layer (red bar) of the DGN is optimised to produce an image that strongly activates the target neuron. The gradient information (blue-dashed line) flows from the layer containing the target neuron in the DCNN via backpropagation through the image to the latent vector layer of the DGN. This process generates a new image that maximally activates the target neuron within the DCNN. In GenerativeAM, the gradient information does not terminate at the bottom layer of the DCNN but is passed through the image layer to optimise the latent vector of the DGN, which is then used to generate a preferred image for the selected target neuron. Image adapted with permission from Nguyen et al. (2016).

2.1.1 Model architecture

The DCNN used by Nguyen et al. (2016) is the CaffeNet architecture (Jin, 2014), a minor variant of the AlexNet architecture (Krizhevsky et al., 2017). The DCNN was pre-trained for image classification through supervised learning on a dataset of natural photographs, ImageNet (Russakovsky et al., 2015; Schmid et al., 2015;

Szegedy et al., 2015; Krizhevsky et al., 2017), based on the thousand categories of ImageNet. The CaffeNet architecture consists of five convolutional layers (c1–c5) and three fully connected layers (fc6, fc7, and fc8; Figure 3). The layer fc8 is the highest layer of the DCNN (pre-softmax) and has 1,000 outputs, each corresponding to one of the ImageNet categorical labels. During training, the visual characteristics of ImageNet photographs are extracted across all layers of the DCNN. The network learns via backpropagation to associate these features with distinct categorical labels. Consequently, the trained network implements a mapping from the pixel level of the input image to the respective categorical labels, represented as activation within specific neurons within the highest layer of the network (fc8).

The DGN used by Nguyen et al. (2016) was pre-trained independently from the DCNN [DGN taken from Dosovitskiy and Brox (2016a,b)], using a Generative Adversarial Network (GAN) approach, combined with the representation learning method (see Supplementary material 1.1.1 for details of the GAN training). This DGN consists of nine up-convolutional (up-sampling and a subsequent convolutional) layers (u1–u9) one 'reshape' layer and three fully connected (fc) layers (Figure 3) which are designed to invert a convolutional network.

2.2 Parameters used to simulate aetiologically distinct VHs

Based on the phenomenological profile of each aetiologically distinct category of VH (in terms of their complexity, veridicality, and spontaneity; see Table 1), we applied specific manipulations to AM to simulate the hallucinatory phenomenology of each group: AM type (ClassicalAM or GenerativeAM), error functions (Fixed, Deep-Dream, or Winner-Take-All), and user-selected target layer for AM to terminate within the DCNN [Lower (conv4) or Higher (fc8)].

For all simulations we ran the model for a total of 1,000 iterations. This value was derived from experimentation with the optimal number of iterations to allow GenerativeAM to converge on a stable output. After extensive piloting, we found that 1,000 iterations were sufficient to ensure the output image reached saturation, that is, it was not altered any further by subsequent iterations. For all further simulations we generated synthetic images following 10, 50, 100, and 1,000 iterations. See Supplementary material 1.1.4 for details of the input images and minor optimisations used for all simulations and Supplementary Table S1 for the pseudo code used to generate all images. The code used to generate all synthetic images is freely available at https://github.com/ksk-S/ModellingHallucinations.

2.2.1 Benchmark simulations

To create a benchmark against which simulations of hallucinatory phenomenology could be compared, we used the model to simulate non-hallucinatory (veridical) perceptual phenomenology by applying GenerativeAM with the Winner-Takes-All error function and terminating the upward pass of AM at the top categorical layer of the DCNN (fc8).

2.2.2 Neurodegenerative VHs

To simulate complex Neurodegenerative VHs, we used GenerativeAM (high veridicality), the Fixed error function (spontaneous) and terminated the upward pass of AM at the top categorical layer (fc8) of the DCNN. Five target neurons were randomly selected from layer fc8, resulting in the Fixed error function maximising the activity of the DCNN target neurons representing flower, bird, mushroom, lamp and volcano. Producing synthetic VHs using an increasing number of iterations provided an additional means of manipulating the veridicality of the image. We found that using GenerativeAM with increasing iteration number resulted in a decrease in the veridicality of the synthetic VH, an effect that became saturated following approximately 200 iterations. We used the same parameters to simulate simple Neurodegenerative VHs except we terminated the upward pass of AM at the lowest layer (conv4) of the DCNN. Again, five target neurons were randomly selected from the conv4 layer of the DCNN. We found that unlike the critical role that a target neuron has in shaping the content of the synthetic complex VHs, the target neuron used to generate simple VHs did not have as much of an influence in shaping the resulting synthetic VHs. This is due to neurons within this layer representing similar low-level visual features of natural images.

2.2.3 CBS VHs

To simulate simple and complex CBS VHs we used the same parameters as for Neurodegenerative VHs, except that we introduced representative features of the visual deficits associated with CBS (central blur) into the input image.

TABLE 1 Summary of phenomenological characteristics associated with Neurodegenerative, CBS and psychedelic VHs (clear) and the corresponding model manipulations (shaded) used to simulate these characteristics.

Type of visual phenomenology	Veridicality	AM type	Complexity	Layer of DCNN	Spontaneity	Type of error function
Benchmark	High	GenerativeAM	Complex	Highest layer ('fc8')	Dependent	Winner-Take-All
Neurodegenerative Complex VH	High	GenerativeAM	Complex	Highest layer ('fc8')	Independent	Fixed
CBS Complex VH	High	GenerativeAM	Complex	Highest layer ('fc8')	Independent	Fixed
CBS Simple VH	High	GenerativeAM	Simple	Lower layer ('conv4')	Independent	Fixed
Psychedelic Complex VH	Low	ClassicalAM	Simple	Lower layer ('conv4')	Dependent	Deep-Dream
Psychedelic Simple VH	Low	ClassicalAM	Complex	Highest layer ('fc8')	Dependent	Deep-Dream

Veridicality refers to the similarity in the visual quality of a given experience compared to normal visual experience. *AM Type* refers to the use of either GenerativeAM or ClassicalAM. *Complexity* refers to the occurrence of simple or complex visual experience. *Layer of DCNN* describes the level within the DCNN that AM terminates. *Spontaneity* refers to whether the visual experience is dependent or independent of sensory input. *Type of Error* refers to whether Fixed, Deep-Dream or WTA error function was used. To simulate CBS VHs, we manually blurred the centre of each input image to mimic the central visual deficit associated with CBS.

2.2.4 Psychedelic VHs

To simulate complex Psychedelic VHs, we used ClassicalAM (low veridicality), the Deep Dream error function (dependent) and terminated the upward pass of AM at the top categorical layer (fc8) of the DCNN. Unlike GenerativeAM, we found that the synthetic VHs generated by ClassicalAM did not ever reach saturation and increasing the number of iterations resulted in a graded decrease in veridicality of the resulting synthetic VHs. We used the same parameters to simulate simple Psychedelic VHs, except we terminated the upward pass of AM at a lower layer (conv4) of the DCNN.

2.3 Image similarity analysis

To provide an objective method of measuring the realism of our simulations of hallucinatory phenomenology we used a commonly applied method for assessing the realism of synthetic images produced by generative models: the Inception Score (IS; Salimans et al., 2016). To calculate the IS an image is entered into an 'inception network', a class of DCNN (e.g., GoogleNet), trained for image classification (Szegedy et al., 2015). An image produces the highest IS when it activates only a single neuron in the categorical layer of the DCNN, meaning that the features of the image converge into a single categorical label. We compared the IS between input images, benchmark, and all simulations of synthetic VHs.

To calculate the IS, we utilized the same five images for all simulations. Additionally, we selected 27 random image categories from CaffeNet and performed an internet search to find a single exemplar image of each category, using the specific CaffeNet label of each category (see Supplementary material section 1.1.4 for more detail and Figure S3 for the full set of images used). This led to a total of 32 input images. For each image we then generated synthetic examples of non-hallucinatory and hallucinatory experience using the model parameters described in section 2.2, resulting in six main categories: (1) Non-hallucinatory (veridical) perceptual phenomenology (2) Simple CBS VHs, (3) Simple psychedelic VHs, (4) Complex Neurodegenerative VHs, (5) Complex Neurodegenerative VHs, (6) Complex psychedelic VHs (see Supplementary material 1.1.6 for all synthetic non-hallucinatory and hallucinatory images generated).

To provide a baseline IS for our model we first calculated the IS for the 32 input (unaltered) images. We then calculated the IS for each of the six categories. This resulted in a single IS score between 0 and 32 for the baseline and six categories, with higher numbers denoting greater realism of the images.

2.4 Experimental studies: verifying the validity of synthetic VHs

To assess the ability of our model to accurately simulate etiologically distinct VHs, we conducted two separate studies: (i) an online survey that targeted participants with recent experience of classical hallucinogens, and (ii) a semi-structured series of phenomenological interviews in which patients with Lewy Body Dementia, Parkinson's Disease or Charles Bonnet Syndrome described their VHs in detail.

2.4.1 Online psychedelic survey

This experiment was designed to assess how closely participants' chosen psychedelic experiences matched the synthetic VHs produced by our model. We reasoned that if our simulations were successful in producing representative examples of psychedelic experience, participants would be more likely to select synthetic psychedelic VHs, compared to Neurodegenerative VHs, as most closely matching their experience. The experiment was carried out in accordance with approved guidelines provided by the University of Sussex, Research Ethics Committee, and was pre-registered using OSF.[1] Here, we report abbreviated methods, for a full description of the experiment see Supplementary material 1.2.

2.4.1.1 Participants

Eighty-one participants completed the online survey who reported having taken a classical psychedelic hallucinogen (LSD, psilocybin, or DMT) within the previous 12 months.

2.4.1.2 Procedure

The experiment consisted of collection of background information, an image selection task, and assessment of hallucinatory phenomenology. It began by asking eligible participants to select a particular psychedelic experience from within the last 12 months, that they would use to answer all the questions in the survey, the type of classical hallucinogen this experience related to and the subjective potency of this experience using a scale from 0 (not potent at all) to 5 (as potent as my most intense psychedelic experience).

Participants then completed 5 practice trials of the image selection task. Each trial of the image selection task presented 6 randomly selected synthetic VHs images (3 Neurodegenerative and 3 psychedelic). Participants were asked to select which image (if any) that most closely resembled their chosen psychedelic experience (see Supplementary material 1.2 for more detail).[2]

The main experiment used the same trial structure as the practice session but consisted of a total of four blocks, each with 32 trials. Two blocks contained images simulating simple VHs (3 Neurodegenerative and 3 psychedelic), and two blocks contained images simulating complex VHs (3 Neurodegenerative and 3 psychedelic). Within each block, 2 catch trials were placed in random positions that contained task irrelevant instruction (e.g., click the top-right image), to maintain participants' attention.

Following the main experiment, participants were asked to indicate if their chosen psychedelic experience contained simple, complex or both types of hallucinatory content. They were then asked separately about the veridicality [from 1 (identical) to 10 (completely different)], and the spontaneity [scale from 1 (completely dependent) to 5 (completely independent)] of each type of hallucinatory content in relation to their existing visual experience (see Supplementary Tables S3–S6 for actual questions).

2.4.1.3 Pre-registered analysis

We followed the analysis plan and exclusion criteria described in the pre-registration document (see text footnote 1). We excluded participants who skipped more than 75% of the trials, those who had an average reaction time of less than 1 s, and those who did not meet the 1-min time limit for more than 50% of all trials. In addition to the

[1] https://osf.io/qu8xe

[2] https://osf.io/nr4ke/files/osfstorage

pre-registered exclusion criteria, we set post-hoc exclusion criteria in which participants who answered four or more of the eight catch trials incorrectly were excluded.

To investigate if synthetic psychedelic VHs were phenomenologically similar to psychedelic experiences, we used a one-tailed, one-sample *t* test using the test variable of the sample mean and a chance level of 0.5 (50%) to compare the ratio of responses between psychedelic and Neurodegenerative synthetic VHs for both simple and complex blocks separately. A resulting value of p equal to or less than 0.05 ($p \leq 0.05$) was interpreted as evidence that participants selected psychedelic synthetic VHs at a higher-than-chance frequency. This standard *t* test was accompanied with a Bayesian *t* test. A Bayes Factor (BF) greater than 3 was taken to indicate sensitive evidence for the hypothesis that participants selected simulations of psychedelic VHs at a higher-than-chance frequency, whilst a BF < 1/3 was taken to indicate sensitive evidence for the null hypothesis. BFs between these two thresholds (>1/3 but <3) were taken as evidence that the data was insensitive (Jeffreys, 1939; Dienes, 2014).

2.4.1.4 Exploratory analysis

We performed two exploratory analyses which were not included in the pre-registration. First, to explore if any of the other image generation parameters used to produce the synthetic VHs affected how representative of psychedelic experience they were, we conducted two separate ANOVAs (simple VHs and complex VHs) for the most frequently selected hallucination type (either psychedelic or Neurodegenerative VHs) based on the pre-registered results, using image selection frequency as the dependent variable. For complex VHs we conducted a 3 × 2 ANOVA of image selection frequency, with factors iteration number (10, 100, or 1,000) and error function (WTA or Fixed). For simple VHs we conducted another separate 3 × 2 ANOVA of image selection frequency, with factors iteration number (10, 100, or 1,000) and target layers (conv3 or conv4).

Second, we examined the correlation across participants between the reported potency of the chosen psychedelic experience and the number of iterations used to generate the synthetic VHs again for the most selected hallucination type (either psychedelic or Neurodegenerative), reasoning that those participants who had reported a higher potency psychedelic experience would be more likely to select synthetic VHs produced by a higher number of iterations. First, for each participant, we ran a linear regression across all stimuli, plotting the iteration level of each stimulus against how frequently that stimulus was selected. The slope value (coefficient) of each regression reflects the iteration level preference for a participant, with a positive slope indicating a tendency to select a higher iteration value. We then ran a one-tailed bivariate correlation (Pearson's coefficient) across participants using the variables of 'iteration preference', taken from the slope of the regression, and potency rating (0–5).

2.4.2 Clinical semi-structured phenomenological interview

Twenty-two participants (13 female) took part in a semi-structured interview designed to enquire about the visual phenomenology associated with CBS and Neurodegenerative VHs (PD and LBD, see S5; mean age = 67 years, SD 15.9). Nine participants had received a formal diagnosis of Parkinson's Disease (PD), three a formal diagnosis of Lewy Body Dementia (LBD), and 10 reported visual loss resulting in a formal diagnosis of Charles Bonnet Syndrome (CBS). This experiment was carried out in accordance with approved guidelines provided by the University of Sussex, Research Ethics Committee.

All participants were interviewed by author D.J.S. by video conference call (zoom etc.) or telephone, and each interview lasted no longer than 45 min. After an introduction, and verbal consent was provided, the semi-structured interview began by collecting background information about general features of participants' VHs, such as how long they had experienced visual hallucinations, how frequently they occurred, how long on average each VH lasted, the differing types of content of their VHs, the complexity of their VHs (simple/complex) and if they had ever confused their VHs as being real (reality monitoring; see Supplementary material 1.4 for the full interview material).

Each participant was then asked to describe in as much detail as possible their most recent (target) experience of a VH. We used open-ended questioning to allow participants to answer in detail. For each target VH, the spontaneity of the VH was assessed by the context of the description. A VH was interpreted as being spontaneous if it was described as being a transformation of pre-existing aspects of a visual scene, as compared to occurring in the absence of any corresponding pre-existing content. If there was any ambiguity in participants' descriptions regarding spontaneity, the interviewer guided the participant towards reporting on this aspect of their VH, by asking them specifically if their VHs occurred in the absence of pre-existing content within the visual scene.

Participants were then asked to rate the veridicality of the VH: 'On a scale of 1–10, 10 being identical to normal visual experiences, 1 being not at all like normal visual experience, how would you rate the visual quality of this VH?'. For CBS participants in which no vision was preserved, they were asked to compare the veridicality of the target VH to memories of their visual experience before the onset of eye disease. This procedure was repeated with other instances of VHs that the participant reported, up to a maximum of 5 descriptions.

Following this portion of the interview, each participant was shown 5 × 10 grids of images displaying examples of Neurodegenerative (GenerativeAM) and psychedelic (ClassicalAM) synthetic VHs. The first column displayed the 5 unaltered input images used in all simulations. Each successive column displayed synthetic complex GenerativeAM or ClassicalAM VHs following 5, 10, 50, 100, 200, 400, 600, 800, and 1,000 iterations. The participants were instructed: 'please select the column, if any, that displays the closest visual similarity to your experience of visual hallucinations'. If a participant reported simple VHs, they were also shown similar grids of Neurodegenerative (GenerativeAM) and psychedelic (ClassicalAM) simple synthetic VHs. Three CBS participants were too visually impaired to perform this task. See Supplementary Figure S12 for all images used in the interview.

3 Results

We simulated three distinct aspects of clinical and psychedelic visual hallucinatory phenomenology—their complexity, veridicality and spontaneity—by manipulating a pre-trained coupled DGN-DCNN model. These distinct aspects were selected to reflect the reported phenomenology of VHs experienced by people with certain Neurodegenerative conditions, by people with visual loss, and by neurotypical people following the ingestion of classical psychedelics (see Table 1). We assessed the output of the model objectively by comparing the Inception Scores from our benchmark simulation of non-hallucinatory experience to all other simulations of hallucinatory phenomenology. The Inception Scores of all the simulation results are presented in Table 2. We also conducted both questionnaire and semi-structured interview surveys with people from each of the above three groups to assess the subjective match between model output and reported experience.

TABLE 2 Comparison of Inception Scores for 32 arbitrary input images and synthetic outputs of the model for each simulation.

Image type	Inception score (IS)
Input images	28.54
Non-hallucinatory perception (benchmark)	22.59
Neurodegenerative	19.92
CBS complex	22.40
CBS simple	6.68
Psychedelic complex	5.52
Psychedelic simple	3.42

A higher number denotes a greater degree of visual realism of the image set. Note that the maximum Inception Score is derived from the total number of images used in the calculation, i.e., 32.

3.1 Simulating non-hallucinatory perceptual phenomenology

To provide a benchmark of model performance, we first simulated non-hallucinatory (veridical) perceptual phenomenology. Note that if the performance of the model was perfect in simulating non-hallucinatory perceptual phenomenology, then the input and output should be identical. Figure 4 shows the procedure and representative results from the benchmark simulation.

Using the Inception Score to measure the realism of the benchmark simulation we found that the synthetic outputs of this simulation displayed the highest Inception Score (22.59) of all the reported simulations when compared to the (unaltered) input images (28.54). We used these visualisations and Inception Score as a benchmark with which to assess the veridicality of further simulations.

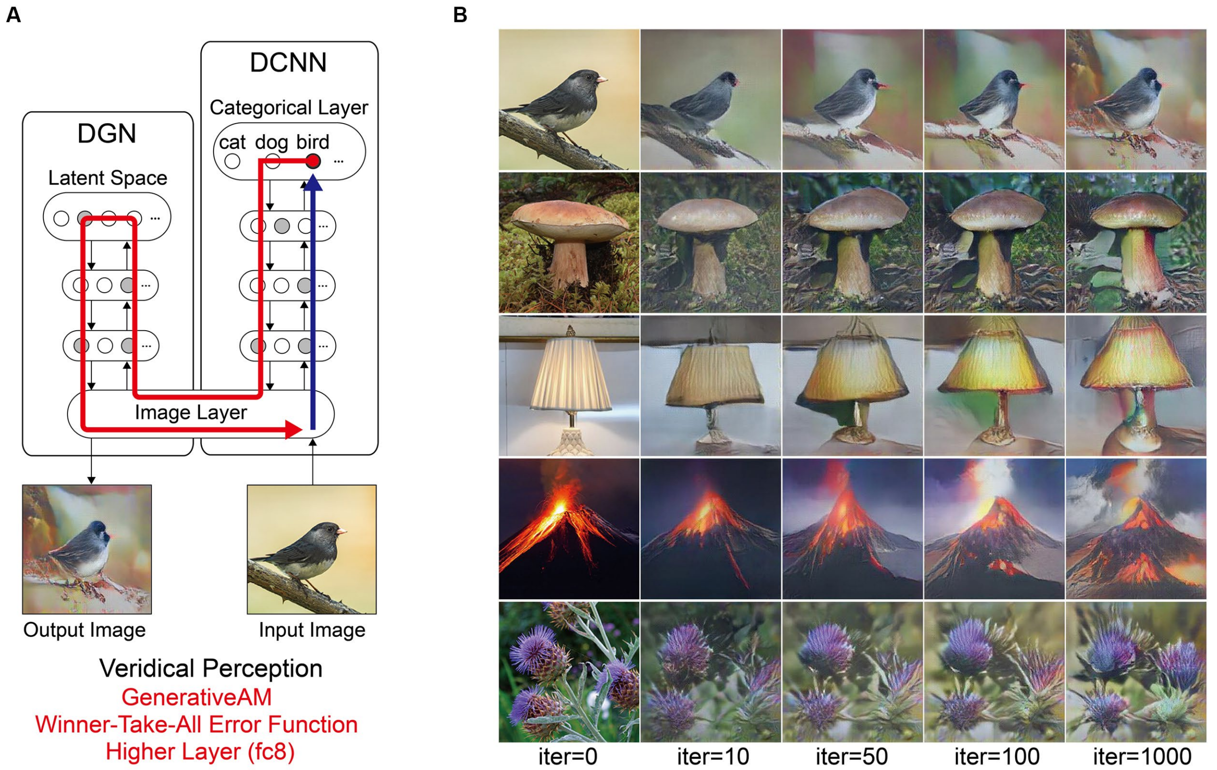

FIGURE 4
Schematic of the model architecture and outputs simulating benchmark (non-hallucinatory) veridical perceptual phenomenology. **(A)** Schematic of model architecture and information flow for a single iteration. An initial arbitrary input image is passed forward through the DCNN, which extracts the visual features of the image across the layers of the network (Blue arrow). Using the Winner-Take-All error function, the neuron that responds maximally to the input image neuron is selected within the highest categorical layer of the DCNN. Using GenerativeAM the categorical information held within the selected target neuron is passed to the DGN, which updates the latent space of the DGN and creates a new synthetic image that maximises the activity of the selected target neuron (Red arrow). **(B)** The initial input images (5 images, left column) and visualisations of the network following 10, 50, 100 and 1,000 iterations.

3.2 Simulating neurodegenerative visual hallucinations

Next, we simulated the perceptual phenomenology associated with complex Neurodegenerative VHs (PD and LBD). Simulating complex Neurodegenerative VHs, here, we found the model's visualisations displayed high veridicality (Figure 5), producing a relatively high Inception Score (19.92) when compared to the benchmark (22.59). In addition, due to the use of the Fixed error function and selection of an experimenter-determined target neuron, the resulting synthetic VHs were unrelated to the input image. Finally, due to selecting the highest layer of DCNN for Generative AM to terminate, the synthetic VHs contained complex recognisable forms. Together, these synthetic VHs display the key phenomenological characteristics of high veridicality, spontaneity and complexity that are typical of complex Neurodegenerative VHs.

3.3 Simulating visual hallucinations due to visual loss (CBS)

Next, we simulated the perceptual phenomenology of simple and complex CBS VHs by introducing a central 'blur' to the input image (see Figure 6) and keeping all other parameters the same as complex Neurodegenerative VHs (Figure 5). We found that the resulting synthetic VHs displayed high veridicality, which was confirmed by a high Inception Score (22.40) when compared to both our benchmark (22.59) and Neurodegenerative (19.92) simulations. Together, these synthetic VHs are in line with typical reports of complex VHs in CBS (Schultz and Melzack, 1991; Teunisse et al., 1996; Ffytche and Howard, 1999; Menon et al., 2003; Abbott et al., 2007), displaying high veridicality and spontaneity.

To simulate simple CBS VH, we restricted the level within the DCNN that AM terminates to a lower layer (conv4) and allowed the DGN to synthesise new images based on the activity of a randomly selected target neuron within this layer (Figure 1, bottom). The resulting synthetic VHs contained low-level colours and textures associated with natural real-world objects, similar to the flashes of light, abstract shapes and repeating patterns commonly reported in simple CBS VHs (Teunisse et al., 1996; Ffytche and Howard, 1999; Menon et al., 2003; Abbott et al., 2007). Due to the predominance of low-level visual content of these synthetic VHs the resulting IS was as expected lower (6.68) than both benchmark and other simulations of complex VHs.

3.4 Simulating psychedelic visual hallucinations

Using the same model parameters as in our previous simulations of psychedelic hallucinatory phenomenology (Suzuki et al., 2017), we were able to simulate the reduced veridicality and

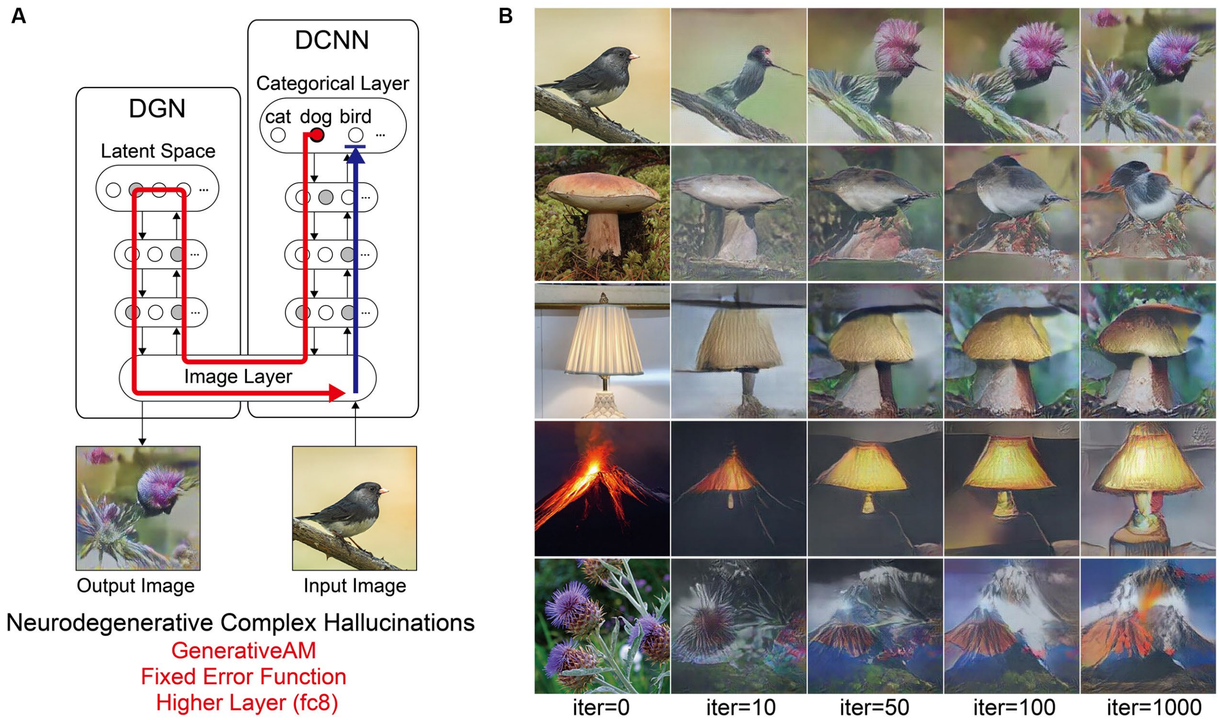

FIGURE 5
Schematic of the model architecture and synthetic VHs simulating Neurodegenerative complex VHs. **(A)** Schematic of model architecture for a single iteration. An arbitrary input image (e.g., bird), is passed forward through the DCNN (Blue arrow). Irrespective of the input image, an experimenter-determined target neuron is selected within the categorical layer of the DCNN (e.g., flower) using the Fixed error function. Using GenerativeAM this information is passed to the DGN and updates the latent space so that the generated image increases the activation of the target neuron (Red arrow). **(B)** The initial input images (left column) and synthetic VHs of the network following 10, 50, 100 and 1,000 iterations. From top to bottom, the experimenter-determined target neuron used in the Fixed error function was Flower, Bird, Mushroom, Lamp and Volcano.

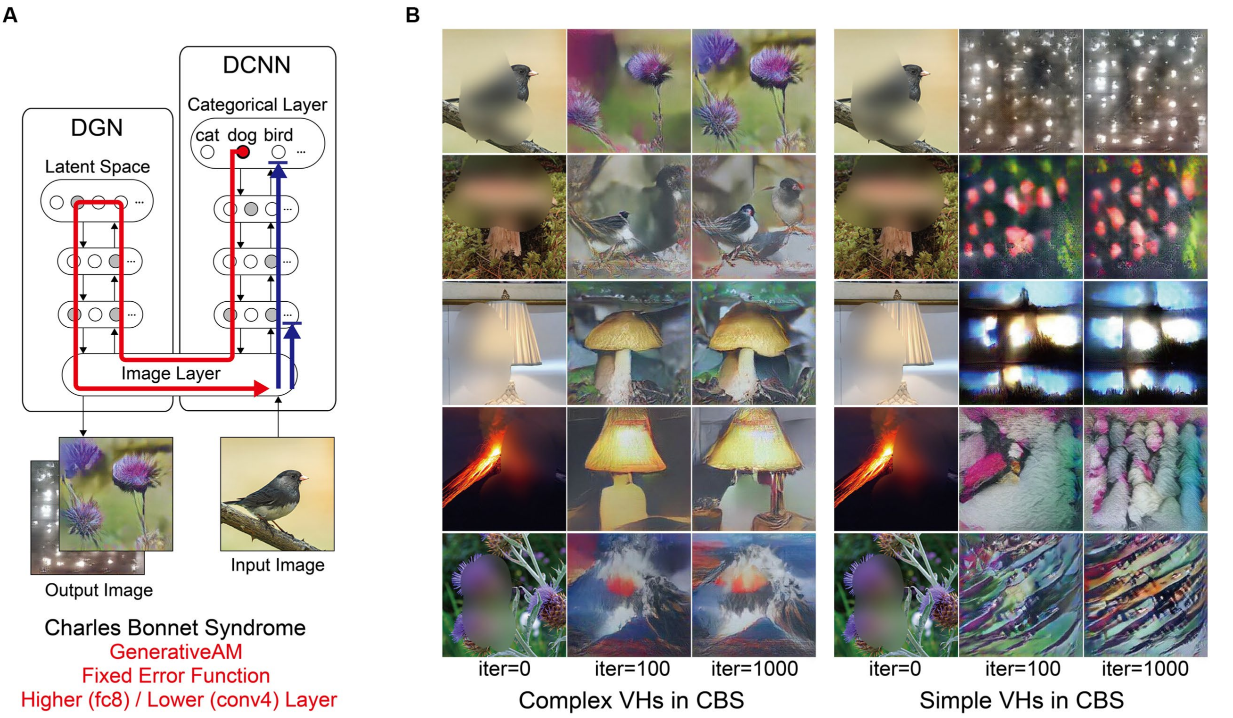

FIGURE 6
Schematic of the model architecture and synthetic VHs simulating simple and complex CBS VHs. **(A)** Schematic of the model architecture used in a single iteration. An arbitrary input image (e.g., bird) with representative features of the visual deficits associated with CBS (central blur), is passed forward through the DCNN (Blue arrow). A randomly-determined target neuron is selected in the DCNN output layer (e.g., flower) using the Fixed error function. With each iteration the DGN generates a new image that maximally activates the target neuron (red arrow). For simple CBS VHs, we restricted the level within the DCNN that AM terminates to a lower layer (Conv4). **(B)** Simulations of complex CBS VHs, iterating from an input image (left column) and outputs following 100 and 1,000 iterations. From top to bottom, the Fixed error function maximises the activity of the 5 randomly selected DCNN target neuron representing flower, bird, mushroom, lamp and volcano. **(C)** Simulations of simple CBS VHs iterating from an input image (left column) and outputs following 100 and 1,000 iterations. A randomly-determined target neuron is selected in the lower layer (conv4) of the DCNN using the Fixed error function. With each iteration, the DGN alters the image based on the learnt features specified by the selected target neuron in the conv4 layer of the DCNN.

dependency on sensory input (reduced spontaneity) associated with complex psychedelic VHs. The use of a consistent set of input images across all simulations in this paper enables the comparison of synthetic psychedelic VHs with aetiologically distinct synthetic VHs.

Compared to benchmark, Neurodegenerative and CBS synthetic VHs, simulations of complex psychedelic VHs displayed low veridicality, reflected by a low Inception Score (5.52). In addition, the complex hallucinatory content within these synthetic VHs can be seen as being driven by visual 'seeds' within the input image (Figure 7). These synthetic VHs also display hallucinatory transformations of the input image, for example, an input image of a mushroom or flower (2nd and 5th rows) was transformed into 'fish' or 'bird-like' hallucinatory content that still somewhat conforms to the global structural properties of the input image (Figure 7). Together, these synthetic VHs are in line with anecdotal reports of psychedelic complex VHs displaying lower veridicality and being driven by visual 'seeds' within an observed scene.

Finally, to simulate simple psychedelic VHs, we used the same parameters as above but restricted the level within the DCNN that AM terminates (conv4; see Figure 8). In contrast to simulations of simple CBS VHs (Figure 6C), the resulting visualisations prominently included geometric shapes, patterns and rhythmic kaleidoscopic imagery similar to those often reported during psychedelic experiences (Tyler, 1978; Bressloff et al., 2001; Díaz, 2010; Cowan, 2015; Nichols, 2016). The marked 'hallucinatory' quality of these visualisations was reflected by a reduced Inception Score (3.42), as compared to simple CBS VHs (6.68). Similar to our simulations of complex psychedelic VHs, these visualisations were also transformations of existing sensory input. As shown in Figure 8, some of the geometric features within each visualisation are clearly driven by the corresponding properties of the input image.

3.5 Clinical semi-structured phenomenological interview

A semi-structured phenomenological interview was used to enquire about the visual phenomenology associated with CBS and Neurodegenerative VHs (see Table 3) (See text footnote 2). Assessing the veridicality of complex VHs, we found that 92% of Neurodegenerative and 90% of CBS patients reported that their VHs appeared 'as real' or 'similar to' their normal visual experiences. In terms of spontaneity, 18 out of 20 participants (2 CBS participants were severely visually impaired) reported that their VHs always occurred spontaneously, in the absence of

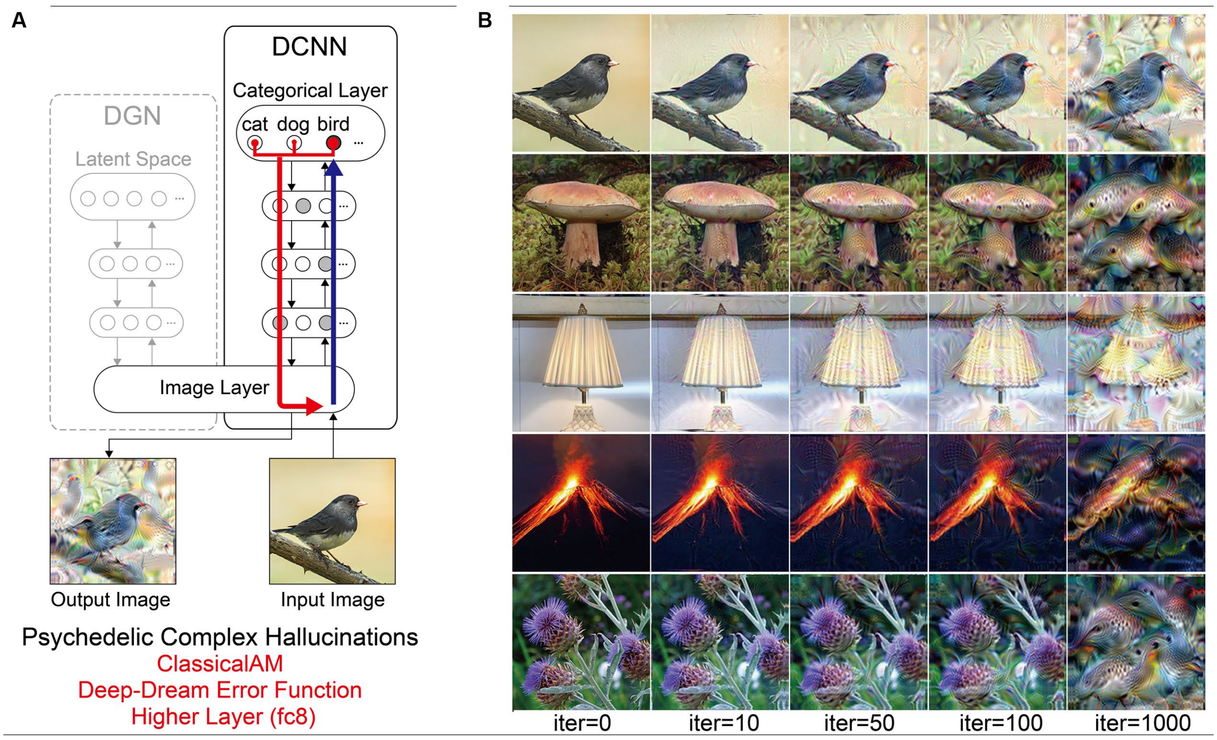

FIGURE 7
Schematic of model architecture and synthetic VHs simulating psychedelic simple VHs. Model. **(A)** Model architecture for a single iteration: the level that ClassicalAM terminates was restricted to a lower layer (conv4). An initial input image is passed forward through the DCNN. Using the Deep-Dream error function, errors are returned for all neurons within the conv4 layer of the DCNN that were activated by the input image. These errors are transmitted via backpropagation to alter the colour of each pixel within an input image to maximise activity within the activated target neurons. **(B)** Input images (left column) and visualisations of the network restricted to the conv4 layer of the DCNN, following 10, 50, 100 and 1,000 iterations. Note that this architecture is identical to that used in Suzuki et al. (2017).

sensory cues, with two PD participants reporting that infrequently their VHs could be described as transformations of existing visual information within the observed scene. For example, one PD participant described an occasion in which their slippers had turned into rats that ran across the bathroom floor. Examining the complexity of the reported hallucinatory content, we found that 92% of Neurodegenerative participants reported that their VHs were complex in nature, with only one participant reporting both simple and complex VHs. In contrast, 90% of CBS participants reported experiencing both simple and complex VHs, with only one participant reporting only complex VHs.

Next, to assess our model's ability to produce representative visualisations of Neurodegenerative and CBS VHs we showed participants a range of synthetic VHs generated using GenerativeAM or ClassicalAM and an increasing number of iterations and asked them to select which (if any) was the most representative of their hallucinatory experience (see Supplementary Figure S12 for images shown). Due to severe visual or cognitive impairment three CBS and two LBD patients were unable to complete this task. Supporting the validity of our synthetic VHs, we found that Neurodegenerative and CBS participants only selected images with high veridicality, that is, created using GenerativeAM, as being representative of their complex hallucinatory experience (Figure 9).

Next, we predicted that due to the high reported veridicality of Neurodegenerative and CBS VHs that participants would select the images generated by GenerativeAM with low iteration numbers. Pooling the images selected by Neurodegenerative participants and averaging each image's iteration number resulted in an average iteration value of 3 ($N = 10$; SD = 4.3), suggesting that for this group of Neurodegenerative participants their VHs displayed high veridicality. In contrast, we found an average iteration value of 18.6 for CBS complex VHs ($N = 7$, SD = 36.1), suggesting that for these participants their complex hallucinatory experience did not display as high a degree of veridicality as Neurodegenerative patients. Pooling data across both groups resulted in the synthetic VHs produced following 10 iterations as being the closest representation of both Neurodegenerative and CBS complex VHs (see Supplementary Figures S12; Figure 9).

Investigating participants' simple hallucinatory experience, we again found that all participants only selected the Neurodegenerative synthetic VHs (GenerativeAM) as being representative of their simple VHs (Figure 9). In contrast to complex VHs, we predicted that participants who experienced simple VHs would select simple synthetic VHs with a high number of iterations, due to the abstract low-level visual features requiring more iterations to develop fully in the synthetic VHs. Indeed, on average, the single PD and four CBS patients who experienced simple VHs and were able to perform the task selected synthetic VHs following 760 iterations ($N = 5$; SD = 260) as being the most accurate representation of their simple VHs (Figure 9).

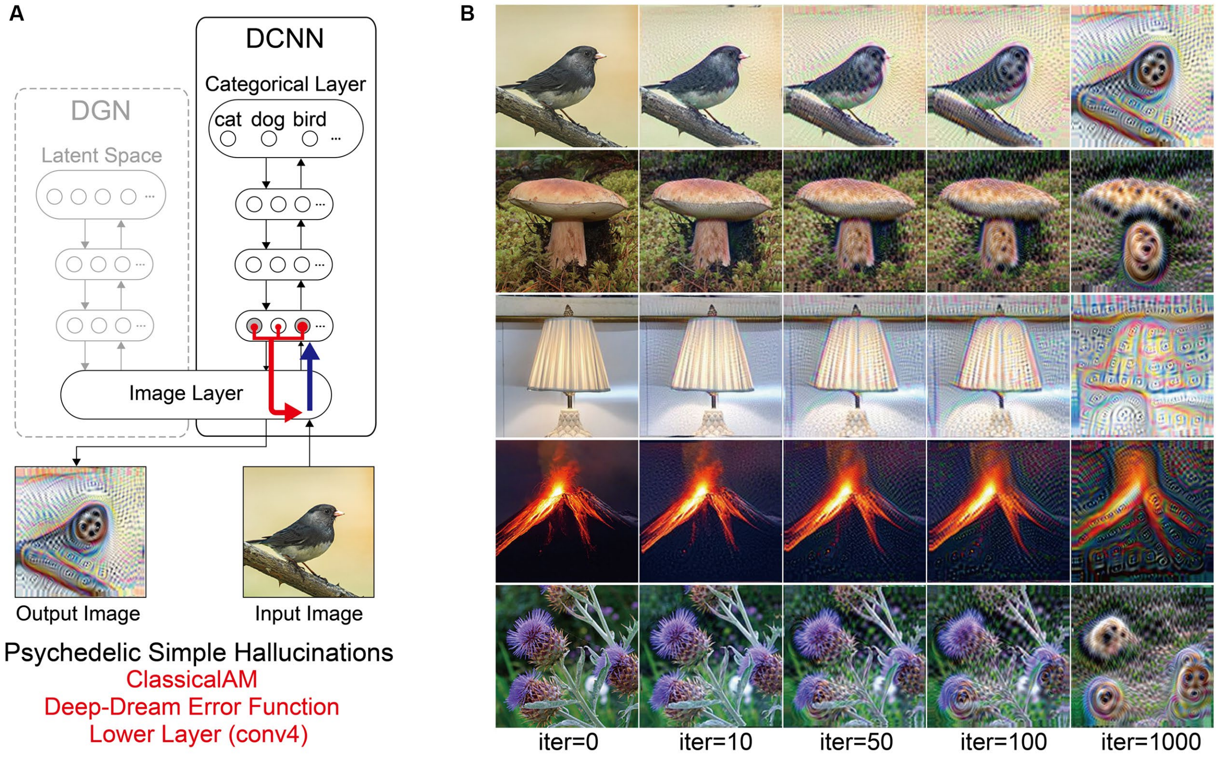

FIGURE 8
Schematic of the model architecture and synthetic VHs simulating psychedelic complex VHs. **(A)** Model architecture used in a single iteration, using only the DCNN and the Deep-Dream error function. An initial input image is passed forward through the DCNN extracting the visual features of the image across the layers of the network. Using the Deep-Dream error function, errors are returned for all neurons within the highest layer (fc8) of the DCNN that were activated by the input image. Using Classical AM these errors are transmitted via backpropagation to alter the colour of each pixel within an input image to maximise activity within the activated target neurons. **(B)** Simulations of psychedelic complex VHs. The initial input images (left column) and generated synthetic VHs following 10, 50, 100 and 1,000 iterations. Note that this architecture is identical to that used in Suzuki et al. (2017).

3.6 Psychedelic survey

3.6.1 Pre-registered analyses

The number of participants for this study did not reach the pre-registered sample size of 200, therefore the following analysis is exploratory. No participants were excluded based on the pre-registered exclusion criteria for this study. Applying the post-hoc exclusion criteria (less than 4 correct answers to the catch trials), four participants were excluded (4.9% of all participants) out of the 81 participants who completed the online survey.

Out of the remaining 77 participants, 46 indicated that their reports related to the ingestion of psilocybin, 24 LSD, and 7 DMT. When asked to rate the subjective potency of their hallucinatory experience on a scale of 0 (not potent at all) to 5 (as potent as my most intense psychedelic experience), average potency ratings were: psilocybin 3.6 (SD = 1.28), LSD 3.5 (SD = 1.1) and DMT 4.3 (SD = 0.76; Figure 10).

To investigate if synthetic images generated by ClassicalAM were representative of psychedelic experience, we first examined the results of the image selection task from the psychedelic survey (Figure 11). We found that participants who reported complex VHs chose the images generated by ClassicalAM significantly more frequently than the images generated by GenerativeAM [one-tailed *t* test, $t(76) = 13.02$, $p < 0.01$, Cohen's d = 3.80, $BF_{10} = 1.92 \times 1{,}018$; Figure 11A]. Similarly, participants who reported simple VHs selected the images generated by ClassicalAM significantly more frequently than the images generated by GenerativeAM [$t(76) = 12.26$, $p < 0.01$, Cohens' d = 3.75, $BF_{10} = 9.51 \times 1{,}016$; Figure 11B]. These results suggest that, for this sample, synthetic VHs generated using ClassicalAM were most representative of both simple and complex psychedelic phenomenology.

3.6.2 Exploratory analyses

Next, to explore if the additional parameters used in the generation of the synthetic psychedelic VHs—the number of iterations, and DCNN level where AM terminated—affected the likelihood of them being selected as representative of participant's psychedelic experience we compared these factors for each category of synthetic VH: complex-psychedelic, and simple-psychedelic, using 2 separate ANOVAs. For complex VHs the factors used in the ANOVAs were iteration level (3; 10, 100, 1,000) and AM Type (2; Figure 11A). For simple VHs the factors used were iteration level (3; 10, 100, 1,000) and the layer of DCNN (2; Figure 11B).

Results of the ANOVAs revealed a significant main effect for iteration level only for Simple-psychedelic VHs [$F(81,2) = 29.28$ $p < 0.01$, $\eta 2 = 0.12$]. Additional post-hoc tests revealed that the

TABLE 3 **Results of semi-structured clinical interviews, showing demographic information and characteristics of Neurodegenerative (PD and LBD) and CBS VHs.**

Demographics	PD/LBD	CBS
N =	12	10
Age	69 ± 8.61	72.1 ± 21.1
Gender	M = 7/*F* = 5	M = 2/*F* = 8

Veridicality	PD/LBD	CBS
Normal visual experience (9–10)	67%	30%
Similar (5–8)	25%	60%
Less clear (<5)	8%	10%

Spontaneity	PD/LBD	CBS
Spontaneous	83%	100%
Visual transformation	17%	0%

Complexity of VHs	PD/LBD	CBS
Simple only	0%	0%
Complex only	83%	10%
Both	17%	90%

Reality monitoring		
Preserved	50%	75%
Not preserved	25%	25%
Variable	25%	0%

Frequency of VHs	PD/LBD	CBS
Daily	58%	60%
Weekly	17%	0%
Monthly	25%	40%

How long VHs have occurred	PD/LBD	CBS
<1 year	42%	0%
1–10 years	50%	70%
>10 years	8%	30%

Duration of VHs	PD/LBD	CBS
1–5 s	8%	0%
5–60 s	33%	30%
1–60 min	58%	70%

Content	PD/LBD	CBS
Adults	92%	90%
Children	33%	10%
Animals	50%	50%
Scenes	8%	10%
Other objects	0%	30%

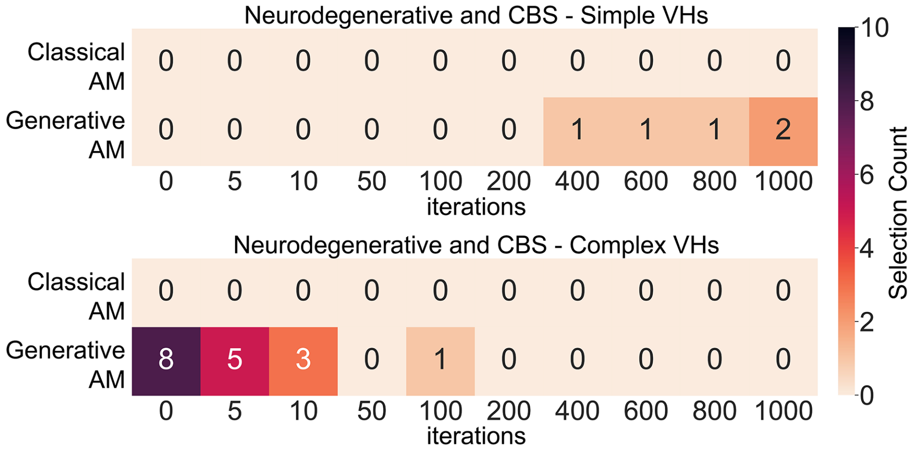

FIGURE 9
Matrices displaying the selection frequency of synthetic VHs by Neurodegenerative and CBS participants that displayed the closest visual similarity to their hallucinatory experience within the clinical interview. Based on the type of hallucinations reported by the participant, they were shown a matrix of either simple, complex or both types of synthetic VHs generated using either ClassicalAM or GenerativeAM with increasing numbers of iterations (see Supplementary Figure S12 for the full set of images). They were then asked to choose which column of synthetic VHs, if any, was similar in visual quality to their experience of simple or complex VHs. The top matrix displays the selection count for simple synthetic VHs generated using ClassicalAM and GenerativeAM and the number of iterations (01,000). All participants only selected 'Neurodegenerative' synthetic VHs (GenerativeAM) as being visually most similar to their simple VHs. We observed an influence of iteration number on participants selection preference for simple synthetic VHs, with the majority of participants selecting synthetic VHs with high iteration numbers as being most representative of their simple hallucinatory experience, suggesting that for this group their simple VHs displayed the low-level colours and textures associated with natural real-world objects. The bottom matrix shows the same information for Complex VHs. Again, all participants only selected 'Neurodegenerative' synthetic VHs (GenerativeAM) as being visually most similar to their VHs. Most participants selected the input image or synthetic VHs with low iteration numbers as being most representative of their complex hallucinatory experience, suggesting that for this group their complex VHs displayed high veridicality.

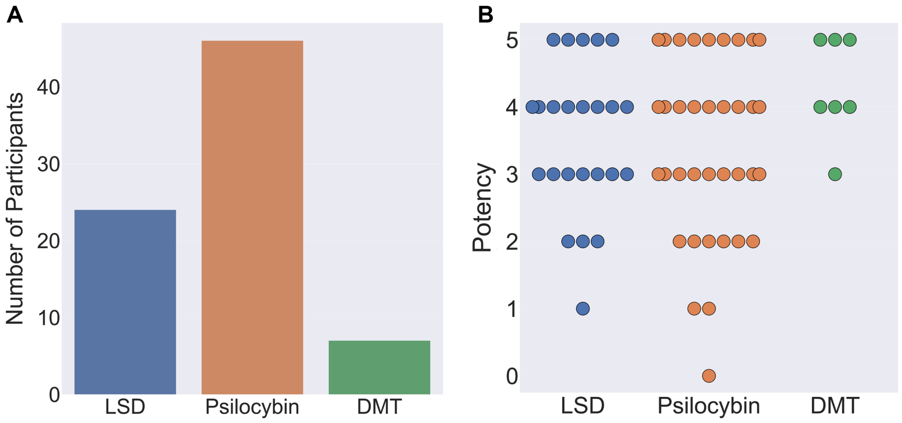

FIGURE 10
(A) Classical hallucinogen taken by participants during their chosen psychedelic experience (left). **(B)** Reported potency of participants chosen psychedelic experience on a scale of 0 (not potent at all) to 5 (as potent as my most intense psychedelic experience ever). Each circle denotes an individual participant response.

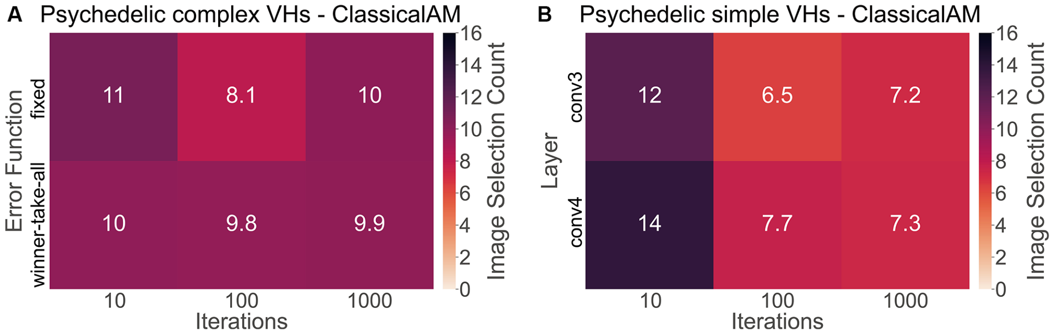

FIGURE 11
(A) Matrix displaying the effect of the type of error function (Winner-Take-All, Fixed) and the number of iterations (10,100,1,000) used to generate complex synthetic VHs on image selection preferences within the psychedelic survey. As can be seen for complex synthetic VHs, these parameters did not dramatically affect image selection preferences, which were relatively similar across parameter combinations. **(B)** Matrix displaying the effect of the layer at which ClassicalAM terminates (conv3, conv4) and the number of iterations (10,100,1,000) used to generate simple synthetic VHs on image selection preferences within the psychedelic survey. We observed an influence of these parameter values on selection frequency for simple synthetic VHs. Synthetic VHs were selected more frequently as being representative of psychedelic experience that were generated using low iteration numbers, suggesting that for this group their simple VHs were relatively 'mild'. All matrices display the selection frequency averaged across participants and the two respective blocks (simple, complex) for each class of image. Note that the maximum number of each cell is 16.

ratio of responses was significantly higher for 10 compared to both 100 [$t(76) = 6.82$, $p_{\text{bonf}} < 0.01$, Cohen's $d = 0.77$] and 1,000 iterations [$t(76,2) = 6.41$, $p_{\text{bonf}} < 0.01$, Cohen's $d = 0.70$; see Figure 11B], suggesting that for these participants the majority of their simple VHs were relatively subtle in nature, consisting of low-level visual distortion and not the colourful kaleidoscopic patterning typically reported under psychedelics. See Supplementary material 1.3 for full results examining the effects of the number of iterations and error function on image selection frequency.

Examining the correlation between the reported potency of the chosen psychedelic experience and how frequently different types of synthetic VH were selected, we found a significant positive correlation between potency and the iteration level preference (the slope value of the regression line of image selection frequency against iteration level): participants with highly potent psychedelic experiences were more likely to select psychedelic complex synthetic VHs with higher iteration levels (Pearson's $r = 0.33$, $p = 0.003$, $BF_{10} = 10.34$). We found the same pattern of results for simple synthetic VHs: a significant positive correlation between potency and the iteration level preference for simple synthetic psychedelic VHs (Pearson's $r = 0.31$, $p = 0.007$, $BF_{10} = 5.344$). These results suggest that increasing the number of iterations used to produce synthetic VHs captures the visual characteristics of differing subjective intensities of psychedelic experience, with higher potency psychedelic experiences being better characterised by synthetic VHs produced using greater numbers of iterations.

Next, we assessed the veridicality of psychedelic complex VHs, by asking participants to rate how similar their VHs were to their normal visual experiences on a scale of 1 (identical, high veridicality) to 10

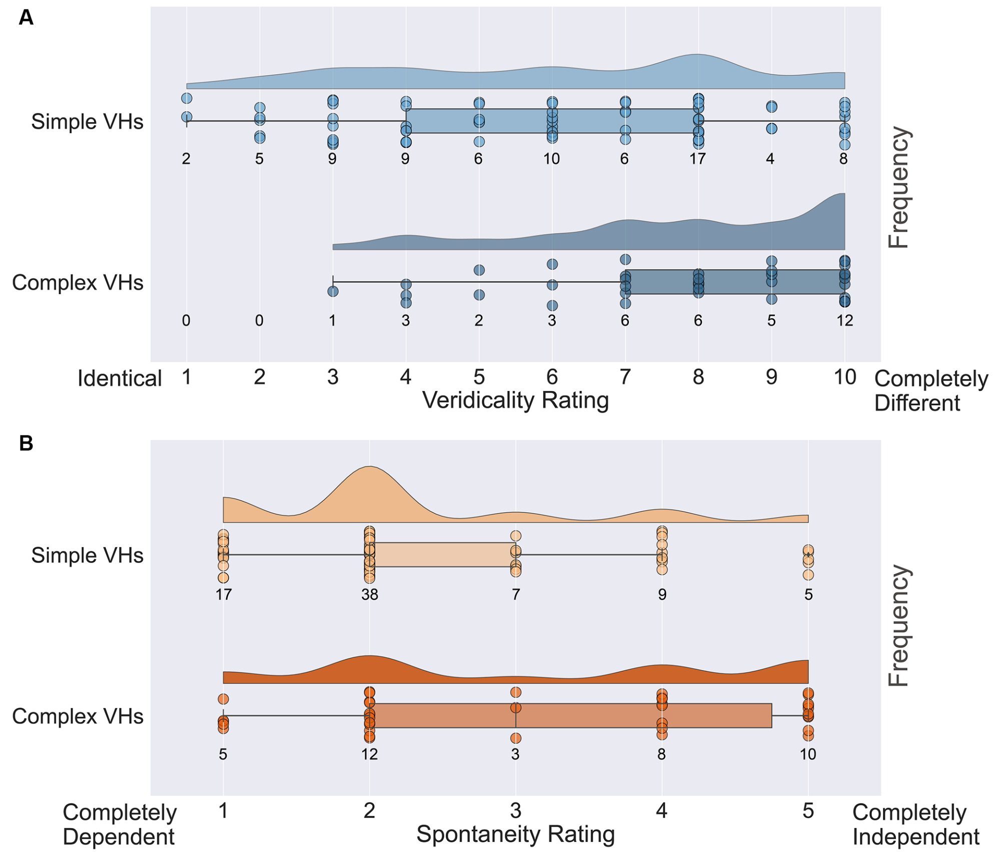

FIGURE 12
Raincloud plots displaying veridicality and spontaneity ratings for all participants in the psychedelic survey for both simple and complex psychedelic VHs. **(A)** Average veridicality ratings for simple and complex psychedelic VHs. Participants were asked: On a scale from 1 (Identical) to 10 (Completely Different), how similar were your complex (and simple) hallucination experiences (perception of identifiable forms: faces; objects; figures; landscapes; scenery) to your normal visual experiences. **(B)** Average spontaneity ratings for simple and complex psychedelic VHs. Participants were asked: On the following scale from 1 (Completely Dependent) to 5 (Completely Independent), please indicate the extent to which the complex (and simple) aspects of your psychedelic hallucinatory experience (perception of identifiable forms: faces; objects; figures; landscapes; scenery) were dependent upon or independent of existing visual content. Each circle indicates the data point for a single participant.

(completely different, low veridicality). Half the participants (38/77) reported that their psychedelic experience had included complex VHs and provided an average veridicality rating of 7.8 (SD = 2.1; Figure 12A). In line with previous findings (Studerus et al., 2011; Carhart-Harris et al., 2016; Preller and Vollenweider, 2018; Sanz et al., 2018) these results suggest that for these participants their complex psychedelic VHs displayed relatively low veridicality.

We also asked the same participants to report the spontaneity of their complex VHs using a scale of 1 (completely dependent) to 5 (completely independent). We found an average spontaneity rating of 3.16 (SD = 1.50; Figure 12B), suggesting that psychedelic complex VHs can be both dependent on, and independent of, sensory input, with a slight tendency to be rated more spontaneous in nature. We found no significant correlations between the potency of the chosen psychedelic experience and the reported veridicality (Pearson's r = 0.27, p = 0.10, BF_{10} = 0.76) or spontaneity ratings (Pearson's r = 0.19, p = 0.27, BF_{10} = 0.37).

Next, we assessed the veridicality and spontaneity of the reported simple VHs using the same method (Figure 12A). All participants reported that they experienced simple VHs (one participant's data was missing for these questions). We found an average rating for veridicality of 6.0 (SD = 2.6), suggesting that the majority of psychedelic simple VHs were distinct from normal visual experiences of simple patterning (Figure 12B). The average spontaneity rating was 2.3 (SD = 1.1), suggesting that psychedelic simple VHs were more likely to be rated as transformations of existing features within the visual scene, than as spontaneous in nature. As with complex VHs, we found no significant correlations between the potency of the chosen psychedelic experience and ratings of veridicality (Pearson's r = 0.22, p = 0.06, BF_{10} = 0.80) or spontaneity (Pearson's r = 0.17, p = 0.15, BF_{10} = 0.40).

3.7 Comparing the veridicality of complex VHs between clinical and psychedelic groups

To explore the veridicality of complex VHs between psychedelic (n = 38) and Neurodegenerative/CBS (n = 22) participants, we performed an exploratory analysis of the mean veridicality ratings between these groups. An independent sample t test revealed that Neurodegenerative/

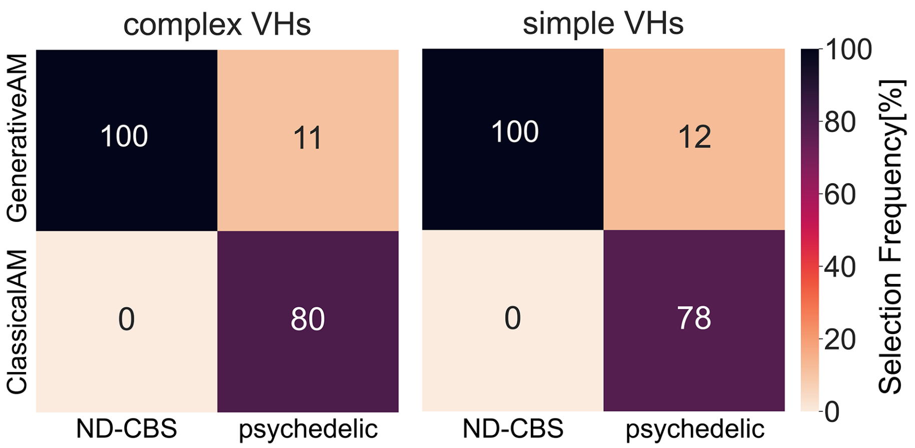

FIGURE 13
A confusion matrix of the selection frequency of simple and complex synthetic VHs, shown as a percentage, generated using ClassicalAM or GenerativeAM for the clinical interview (ND-CBS) and psychedelic survey. In the clinical interview, for both simple and complex VHs participants were shown a grid of images (see Supplementary Figure S12 for the full set of images) generated using either GenerativeAM (Neurodegenerative) or ClassicalAM (psychedelic) and were instructed: 'please select the column of synthetic VHs, if any, that displays the closest visual similarity to your experience of visual hallucinations'. Both Neurodegenerative and CBS participants selected only synthetic 'neurodegenerative' VHs (GenerativeAM) as displaying the closest visual similarity to their hallucinatory experience. For the psychedelic survey, the percentages were calculated as an average count of selection frequency in the forced choice image selection task between ClassicalAM and GenerativeAM. As can be seen, participants with recent psychedelic experience chose synthetic 'psychedelic' VHs (ClassicalAM) with a much higher frequency than 'Neurodegenerative' VHs (GenerativeAM) for both simple and complex synthetic VHs. Note that within the forced choice image selection task participants had the option to skip a trial if none of the images presented matched their chosen psychedelic experience. Therefore, the selection frequency for this group does not equal 100%.

CBS participants rated the veridicality of their complex VHs significantly higher (M = 8.3, SD = 1.8) compared to the psychedelic participants [M = 3.2, SD = 2.1; $t(58) = 9.505$, $p < 0.001$, Cohens'd = 2.55, $BF_{10} = 2.43e + 10$]. Note that the veridicality rating scale for the psychedelic group was inverted to match the scale used in the clinical interview.

3.8 Do the relevant synthetic VHs accurately capture clinical and psychedelic hallucinatory experience?

In two separate studies we investigated variations in hallucinatory experience in Neurodegenerative-CBS patients and people with recent psychedelic experience. We wanted to assess how closely the synthetic VHs were able to capture the specific visual characteristics of each group's VHs. Using methods specific to each population, we asked participants to select which class of synthetic VH (ClassicalAM or GenerativeAM) most accurately captured their hallucinatory experience. Using a clinical interview, we found that Neurodegenerative and CBS participants only selected synthetic 'Neurodegenerative' simple and complex VHs (GenerativeAM) as displaying the closest visual similarity to their hallucinatory experience (Figure 13; Supplementary Table S7). In contrast, using a forced choice task method, participants with recent psychedelic experience selected synthetic 'psychedelic' simple and complex VHs (ClassicalAM) with a much higher frequency than 'Neurodegenerative' VHs as being representative of their chosen psychedelic experience (Figure 13; Supplementary Table S7). Together these results establish that not only are we able to capture specific aspects of hallucinatory phenomenology (veridicality, spontaneity, and complexity) for each aetiology, but also that the relevant synthetic VHs are rated as being most representative of each group's hallucinatory experience, compared to other synthetic VHs produced by the model.

4 Discussion

Visual hallucinations offer fascinating insights into the mechanisms underlying perceptual experience, yet relatively little work has focused on understanding the differences in the phenomenology of VHs associated with different aetiologies. Using a computational phenomenology approach we first identified three dimensions of hallucinatory phenomenology which broadly characterise variations in VHs arising from Neurodegenerative, CBS and psychedelic origins: their veridicality, spontaneity, and complexity. Using a coupled DCNN-DGN neural network architecture, we generated synthetic VHs that captured the differences in these three phenomenological dimensions between Neurodegenerative, CBS, and psychedelic VHs. Each class of synthetic VH was generated by tuning the parameters of the visualisation algorithm to match the phenomenological dimensions represented in each group's VHs. The parameters we manipulated were the inclusion or omission of the natural image prior of the DGN (corresponding to veridicality), altering the error function used to select the target neuron(s) within the DCNN (corresponding to spontaneity), and restricting the level within the DCNN that AM terminates (corresponding to complexity).

We verified the validity of this approach experimentally in two separate studies that investigated variations in hallucinatory experience in Neurodegenerative-CBS patients and people with recent psychedelic experience. Both studies first verified that the three phenomenological dimensions usefully distinguished the different kinds of hallucinations, and then asked whether the appropriate synthetic VHs were able to capture specific visual aspects of hallucinatory phenomenology for each aetiology. Critically, we found that Neurodegenerative-CBS patients only selected Neurodegenerative and not psychedelic synthetic VHs as displaying the closest visual similarity to both their simple and complex hallucinatory experience. In contrast, we found that people with recent psychedelic experience selected synthetic 'psychedelic' VHs with a much higher frequency than 'Neurodegenerative' VHs as being representative of their chosen simple and complex psychedelic experiences. Together, our findings highlight how deep neural network architectures can be used to shed light on the computational mechanisms underpinning atypical perceptual phenomenology.

4.1 Simulating neurodegenerative and CBS VHs

Typically, Neurodegenerative and CBS complex VHs are reported as being similar in visual quality to normal perception, that is, they display high veridicality (Teunisse et al., 1996; Fénelon et al., 2000; Frucht and Bernsohn, 2002; Mosimann et al., 2006; Papapetropoulos et al., 2008). The results of our phenomenological interview confirmed this aspect of hallucinatory phenomenology, with 92% of Neurodegenerative and 90% of CBS participants rating their VHs as being 'as real' or 'similar to' their normal visual experiences (Table 3). We used GenerativeAM to simulate this specific aspect of hallucinatory experience, finding that the learned natural image prior of the DGN resulted in synthetic VHs that displayed high veridicality (Figures 5, 6), which was reflected by a comparable Inception Score to the benchmark simulation of non-hallucinatory perceptual phenomenology (Figure 4).

Complex Neurodegenerative and CBS VHs are also reported to occur spontaneously—i.e., they are not transformations of content within a perceived scene (Schultz and Melzack, 1991; Teunisse et al., 1996; Frucht and Bernsohn, 2002; Mosimann et al., 2006; Papapetropoulos et al., 2008). Indeed, we found that 83% of Neurodegenerative and 100% of CBS participants reported that their hallucinatory experiences occurred spontaneously (Table 3). We attempted to capture this phenomenological characteristic by using an error function for synthetic image generation that selects a target neuron independently of the input image (Fixed), resulting in the synthetic VH being based on the categorical information represented by the target neuron and not the visual features of the input image.

In terms of the complexity of the hallucinatory experience, Neurodegenerative VHs are most commonly reported as being complex in nature (Frucht and Bernsohn, 2002; Mosimann et al., 2006; Papapetropoulos et al., 2008). Our results again confirmed this aspect of hallucinatory phenomenology, finding that 83% of Neurodegenerative participants reported that their VHs consisted of only complex VHs (see Table 2). We simulated this aspect of hallucinatory phenomenology by restricting the hierarchical level used to create the synthetic image; specifically, by ensuring that AM terminates at the highest categorical layer in the DCNN (Figure 2). In contrast, the most commonly reported class of VHs reported in CBS are simple in nature (Schultz and Melzack, 1991; Teunisse et al., 1996; Ffytche and Howard, 1999; Santhouse et al., 2000; Abbott et al., 2007). We found that 90% of CBS participants reported not only simple but also complex VHs. Restricting the AM termination level within the DCNN to a lower layer (conv4) resulted in synthetic VHs that displayed many indicative features of simple CBS VHs (Figure 6). This was due to the overemphasis of the visual features learnt by neurons in this layer during training, together with the influence of the learned natural image prior from the DGN, i.e., low-level colours and textures associated with natural real-world objects.

Verifying the validity of our synthetic VHs, we found that both Neurodegenerative and CBS participants selected only synthetic VHs created using GenerativeAM, and not ClassicalAM, as providing the closest approximation of both their simple and complex hallucinatory experiences (Figure 13). In an exploratory analysis, we examined which iteration value used to generate the synthetic VHs participants selected as being most representative of their complex VHs, where increasing iteration value corresponds to decreasing veridicality of the synthetic VH. Here, we found that, on average, Neurodegenerative participants chose synthetic VHs with a low number of iterations (high veridicality, average iteration value of 3). In contrast, CBS participants, on average, selected a higher iteration value (average iteration value of 18.6) as being the most representative of their complex VHs. Together these findings fit the intuition that for these groups, Neurodegenerative participants experience hallucinatory phenomenology that is close to, but not identical to, their normal perceptual experience, whilst CBS participants have complex hallucinatory experiences that exhibit reduced veridicality compared to Neurodegenerative VHs.

Summarising the performance of our model in simulating Neurodegenerative and CBS VHs, we found that the addition of the natural image prior of the DGN, combined with the Fixed error function, and altering the hierarchical depth at which AM terminates, allowed us to simulate three specific aspects of Neurodegenerative and CBS hallucinatory phenomenology: their veridicality, spontaneity and complexity. The results of the semi-structured interviews with these patients verified that these properties were distinct aspects of the perceptual phenomenology associated with their VHs and that only the appropriate synthetic VHs, produced using GenerativeAM, provided representative examples of both their simple and complex hallucinatory experience.

4.2 Simulating psychedelic VHs

In contrast to complex Neurodegenerative or CBS VHs, psychedelic VHs are typically reported as having reduced veridicality, due to their dream-like qualities and the inclusion of visual distortion and unrealistic colours and patterning (Studerus et al., 2011; Kometer et al., 2013; Carhart-Harris et al., 2016; Kometer and Vollenweider, 2018; Preller and Vollenweider, 2018; Sanz et al., 2018; Timmermann et al., 2018).

To simulate the diminished veridicality associated with psychedelic VHs, we suppressed the influence of the learned natural image prior in the generation of synthetic VHs by removing the contribution of the DGN from the model (Figure 8). This resulted in the use of a visualisation method (ClassicalAM) that was identical to our previous simulations of psychedelic hallucinatory phenomenology (Suzuki et al., 2017). The resulting synthetic VHs displayed striking 'hallucinatory' qualities, bearing intuitive similarities to a wide range of psychedelic VHs reported in the literature (Studerus et al., 2011; Kometer et al., 2013; Carhart-Harris et al., 2016; Kometer and Vollenweider, 2018; Preller and Vollenweider, 2018; Timmermann et al., 2018).

The results of the psychedelic survey confirmed the diminished veridicality associated with psychedelic VHs, with participants rating their complex VHs as displaying low veridicality (Figure 12A). This finding was further supported by an exploratory analysis comparing the veridicality of complex VHs between psychedelic and Neurodegenerative/CBS groups, which found that complex psychedelic VHs were rated as displaying significantly lower veridicality than both Neurodegenerative and CBS VHs (see section 3.7). The reduction in veridicality of this class of synthetic VH (Figure 8) was also reflected by a substantially reduced Inception Score compared to both benchmark and simulations of Neurodegenerative and CBS complex VHs (Table 2).

Complex psychedelic VHs are also commonly reported as being less spontaneous than Neurodegenerative/CBS VHs, often appearing to be transformations of existing information within a visual scene (Kometer and Vollenweider, 2018; Preller and Vollenweider, 2018; Swanson, 2018). We simulated this aspect of hallucinatory phenomenology by using an error function that selected target neurons within the DCNN based on the visual features of the input image (Deep-Dream). However, the results of the online psychedelic survey revealed considerable heterogeneity in this aspect of hallucinatory phenomenology (Figure 12B). Psychedelic complex VHs were rated as being both dependent on or independent of sensory input, with a slight tendency to be rated more frequently as spontaneous in nature. In contrast, simple VHs were more likely to be rated as not being spontaneous, i.e., as being transformations of existing visual input. One possible explanation for this variability is that there may be a confound of dose-dependency, with both complexity and spontaneity of VHs increasing with dose. This would be in line with findings of the prevalence of complex hallucinations increasing with drug dose (Strassman et al., 1994; Shulgin, 1997; Strassmann, 2001; Shanon, 2002; Studerus et al., 2011; Leptourgos et al., 2020). Examining the number of iterations used to produce synthetic psychedelic VHs, we found a significant positive correlation between iteration number (of selected images) and the potency of psychedelic taken, suggesting that, not only did our synthetic VHs closely resemble psychedelic experience, but synthetic VHs produced using differing numbers of iterations were able to capture variations in the intensity of the visual phenomenology of psychedelic experience.

Psychedelic VHs occur in both complex and simple forms. We simulated simple psychedelic VHs by restricting the hierarchical level within the DCNN that AM terminates (i.e., the level at which AM terminates) to a lower layer (conv4; Figure 7). Compared to simulations of simple CBS VHs (Figure 6C), synthetic psychedelic VHs included a high prevalence of geometric shapes, colours and patterning, resembling typical reports of psychedelic simple VHs (Tyler, 1978; Bressloff et al., 2001; Díaz, 2010; Cowan, 2015; Nichols, 2016).

Verifying the validity of psychedelic synthetic VHs using an online forced choice image selection task we found that participants who reported complex VHs selected synthetic 'psychedelic' VHs with a much higher frequency than 'Neurodegenerative' VHs as being representative of their chosen psychedelic experience. We found that the image generation parameters (iterations, error function) did not significantly affect image selection frequency for complex synthetic VHs (Figure 11A).

Participants who reported simple VHs chose simple psychedelic synthetic VHs (ClassicalAM) significantly more frequently than synthetic Neurodegenerative VHs (Figure 13) as most closely resembling their psychedelic experience. Out of participants who indicated they had experienced simple VHs ($n = 81$), we found that on average they were most likely to choose synthetic VHs generated using a low number of iterations (10; see Figure 11B), suggesting that for this sample their simple hallucinatory phenomenology was relatively subtle in nature, consisting of low-level visual distortion, rather than the colourful kaleidoscopic patterning usually associated with classical hallucinogens.

Summarising the performance of our model in simulating simple and complex psychedelic VHs, we found that suppressing the natural image prior (DGN), combined with the Deep-Dream error function, and constraining the hierarchical depth at which AM terminates, produced synthetic VHs that resemble many aspects of psychedelic hallucinatory phenomenology, including their reduced veridicality, spontaneity, and differing levels of complexity. The results of the image selection task verified that synthetic VHs generated using ClassicalAM provided representative depictions of both simple and complex psychedelic hallucinatory experience. They also support the conclusion that psychedelic complex VHs display significantly reduced veridicality compared to Neurodegenerative or CBS VHs, and revealed that, unlike Neurodegenerative or CBS VHs, complex psychedelic VHs could be experienced as being both spontaneous and dependent on sensory input.

4.3 Predictive processing accounts of hallucinatory experience

By decoupling perceptual content from normal sources of sensory input, VHs exemplify the constructive nature of perceptual experience. They demonstrate that the brain is capable of generating rich, detailed and in some cases life-like perceptions irrespective of (or only partly constrained by) actual sensory input (dreams, of course, show this too). It is therefore natural to interpret VHs within 'predictive processing' (PP) accounts of perception and brain function (Rao and Ballard, 1999; Friston, 2005; Clark, 2013; Seth, 2014).

Indeed, considerable work has focused on understanding the computational origins of VHs within a PP framework. Such studies generally associate VHs—Neurodegenerative and clinical VHs in particular—with aberrant perceptual inference, usually appealing to overly strong perceptual priors (Friston, 2005; O'Callaghan et al., 2017; Powers et al., 2017; Corlett et al., 2019). For example, O'Callaghan et al. (2017) found that in PD, patients who experienced VHs displayed slow and inefficient sensory processing in a perceptual decision-making task. They interpreted this result within a Bayesian framework, suggesting that if visual information is accumulated slowly and inefficiently it will be deemed less informative, and may be down-weighted leading to an over-reliance on top-down expectations, resulting in the experience of VHs.

Whilst we acknowledge that the model used in this paper is not an explicit PP model, the architecture and visualisation methods used here display many similarities to PP models in terms of hierarchical information processing. For example, the DCNN used here is a purely feedforward network, which we employed to generate synthetic VHs using both ClassicalAM and GenerativeAM. Both instances of AM optimise the input image via backpropagation, which at least informally, mirrors certain aspects of the top-down signalling that is central to PP accounts of perception [see for example (Millidge et al., 2022), who show that predictive coding networks can implement an approximation of backpropagation].

Specific parallels to PP can also be identified in the parameters used to generate each type of synthetic VH. We simulated the phenomenology of Neurodegenerative/CBS VHs using GenerativeAM and a Fixed error function. These parameters resulted in the input image being overwritten by synthetic hallucinatory content, based on the features represented by a single target neuron. From a PP perspective, this process bears intuitive similarities to theories that posit that this type of VHs occurs due to overly strong perceptual priors overwhelming sensory prediction error signals (Powers et al., 2016; O'Callaghan et al., 2017).

Interpreting the contribution of the DGN within a PP framework, we emphasise that the natural image prior of the DGN should not be confused with perceptual priors. Instead, it can be thought of as a variety of hyperprior, which results in each layer and neuron of the network having a tendency to generate realistic interpretable images. Other examples of hyperpriors within a PP system include the physical constraints imposed by space and time, e.g., 'that there is only one object (one cause of sensory input) in one place, at a given scale, at a given moment', (Clark, 2013, p. 196). Such spatial and temporal hyperpriors narrow and restrict the range of possible hypotheses about the external objects causing incoming sensory input (Clark, 2013), increasing the tractability of (approximate) inference (Tenenbaum et al., 2011; Blokpoel et al., 2012; Clark, 2013; Kwisthout, 2014; Swanson, 2018; Baltieri, 2020). Similarly, the hyperprior of the DGN constrains the vast set of possible uninterpretable images that may excite a target neuron, forcing the network towards producing synthetic outputs that are both interpretable and that closely resemble natural images.

In contrast to clinical VHs, convergent evidence suggests that psychedelic VHs occur due to a reduction in top-down control and an increase in bottom-up information transfer (Muthukumaraswamy et al., 2013; Alonso et al., 2015; Timmermann et al., 2018; Carhart-Harris and Friston, 2019; Schartner and Timmermann, 2020). Such relaxation of high-level priors has been proposed to allow ascending prediction errors from lower levels of the visual system to reach higher-than-normal levels of the visual hierarchy and be integrated into conscious experience (Carhart-Harris and Friston, 2019). Our use of the Deep-Dream error function and ClassicalAM to generate synthetic psychedelic VHs intuitively approximates this disruption in the balance between hierarchical information processing. From a PP perspective, compared to the Fixed or Winner-Take-All, the Deep-Dream error function has the effect of relaxing the perceptual priors imposed on incoming sensory data. The selection of multiple target neurons based on the visual characteristics of the input image by this error function lowers the precision of the perceptual prior, effectively increasing bottom-up information transfer.

4.4 Alternative machine learning models for computational neurophenomenology

The coupled DGN-DCNN network we used served to capture some of the core features of a PP architecture in a way well suited to synthetic image generation. Whilst PP architectures such as predictive coding (PC) networks offer a model of cortical function that exhibits many similarities to our present understanding of human visual perception (Rao and Ballard, 1999; Friston, 2005), current limitations of this class of model, such as the scalability of local learning rules (Bartunov et al., 2018), their restriction to low-dimensional images and narrowly focused datasets, preclude their use in simulating visual phenomenology. In contrast, machine learning models such as DCNNs are highly scalable and, when coupled with DGNs, have the potential to shed new light on the neural mechanisms underlying visual phenomenology in a way that reflects key principles of theories of predictive perception.

DCNNs have long been utilised as models of the human visual system. Their utility is exemplified by close correspondences between activity within layers of DCNNs and activity in regions along the ventral visual stream. For example, a comparison of internal representational structures of trained DCNNs and the human visual system performing similar object recognition tasks has revealed similarities between the representational spaces in these two distinct systems (Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Grossman et al., 2019). However, others have argued that the correspondences between brain and model activity involve shared, not task-relevant, variance and more rigorous tests are required to provide a stricter measure of the correspondence between these two systems (Sexton and Love, 2022).

More fundamentally, the prevalence of top-down connectivity in the human visual system points to a fundamental difference with DCNNs, which implement top-down influence purely through the learning algorithm (typically backpropagation). A productive direction for future research would therefore be to address the current limitations in using PP-like architectures—such as predictive coding (PC)—for generating simulated phenomenology.

One promising possibility is raised by work demonstrating certain functional equivalences between backpropagation and predictive processing (Whittington and Bogacz, 2017; Millidge et al., 2022). Here, the bottom-up classification process employed by DCNNs are viewed as predictions from data to labels, and the error backpropagation that refines the network weights are viewed as 'prediction errors'. Based on this work, a recent 'hybrid predictive coding' network architecture has been proposed (Tschantz et al., 2023) in which predictions and prediction errors flow bidirectionally in hierarchical networks, to adaptively balance 'fast' and 'slow' inference. Hybrid predictive coding networks combine 'amortised inference'—which performs 'fast and cheap' approximate perceptual inference for familiar data—with the iterative inference implemented in standard predictive coding. By combining fast and slow inference, hybrid predictive coding networks are well suited for shedding light on the biological relevance of the feedforward sweeps and iterative recurrent activity observed in the visual cortex during perceptual tasks. Future work could also examine how adaptations of hybrid predictive coding could be used to provide insights into aspects of hallucinatory phenomenology.

4.5 Limitations

There are several limitations worth highlighting in the current investigation. First, the benchmark performance of the model did not succeed in producing output images that were identical to the input images. This finding may have been due, in part, to the synthetic images generated by the DGN being based on the compressed latent vectors within higher layers of the DGN and not the pixel values of the input image. The compression of the features of the input image has the effect of diminishing the realism of the output image. Another contributing factor is likely to be the number of separate categories the DCNN has been trained to classify, which whilst relatively large (1,000 classes), is trivial compared to the human brain. This constraint means that when presented with an arbitrary (untrained) input image the model cannot generate an output that displays a fine-grained visual similarity to the input image. Using larger training-set data would ameliorate this problem.

Another important limitation that should be considered when interpreting these results is the relatively small sample sizes in both studies investigating the phenomenology of different hallucinatory experiences. We acknowledge the conclusions drawn here are limited to these specific groups of participants. Further work is required to establish if the phenomenological differences in VHs found here, both across and within different aetiological categories, apply to wider populations.

In an attempt to assess how representative our synthetic VHs were of aetiologically distinct hallucinatory experiences we employed different methodological approaches, enquiring about the general phenomenology of a person's VHs, including specific questions about the complexity, veridicality, and spontaneity of their VHs. Our model also allowed us to ask participants to directly rate the visual similarity between the appropriate synthetic VHs and their hallucinatory experiences. The flexibility of our model allowed us to create a wide range of images produced using differing number of iterations or constraining the hierarchical depth at which AM terminated. Whilst we highlight the potential of this approach, we also acknowledge the inherent challenges in assessing how closely an individual's private subjective experience is captured by a synthetic VHs.

Our use of the Fixed error function was intended to simulate the overly strong perceptual predictions thought to play a role in Neurodegenerative and CBS VHs (Powers et al., 2016; O'Callaghan et al., 2017). In a loose sense, our use of a Fixed target neuron could therefore be viewed as simulating the tendency of these populations to hallucinate specific categories of hallucinatory content, primarily people, with fewer reports of animals (Frucht and Bernsohn, 2002; Mosimann et al., 2006; Papapetropoulos et al., 2008). However, we acknowledge that this approach is an oversimplification of these populations' hallucinatory experience, as a Fixed target neuron would always generate the same synthetic VH, based on the categorical information held by the target neuron.

As discussed above, there is increasing evidence highlighting the lack of biological plausibility of the architecture and learning rules used by DCNNs. Future studies attempting to generate simulations of phenomenology should consider using PP-like architectures.

Finally, we highlight that our simulations were designed to capture the specific variations in phenomenology of VHs arising from neurodegenerative, CBS and psychedelic origins and as such were not designed to map directly between the pathological mechanisms and variations in hallucinatory experience observed in these different aetiologies. Future research should explore this interesting avenue of research using combinations of the biologically plausible models outlined here.

5 Conclusion

We have described a novel method for simulating altered visual phenomenology associated with Neurodegenerative, CBS and psychedelic VHs. Using a coupled deep convolutional neural network (DCNN) and deep generator network (DGN) we leveraged recent advances in visualising the learned representations of these networks to simulate three dimensions relevant to distinguishing aetiologically distinct VHs: their realism (veridicality), their dependence on sensory input (spontaneity), and their complexity. By selectively manipulating the parameters of the algorithm used to generate the synthetic VHs, we were able to capture differences in these phenomenological characteristics across distinct VH populations. Two experimental studies confirmed the existence of variations in these phenomenological dimensions across groups and critically, that the appropriate synthetic VHs were representative of the hallucinatory phenomenology experienced by Neurodegenerative-CBS and psychedelic groups.

Overall, our findings highlight the phenomenological diversity of VHs associated with distinct causal factors, and the need for a more fine-grained mapping between computational mechanisms and specific hallucinatory phenomenology. We have shown here how a coupled machine learning network and specific visualisation algorithms can be used as a step towards this goal, by demonstrating that such models can successfully capture the distinctive visual characteristics of hallucinatory experience. Future research could usefully apply this approach using more biologically plausible predictive processing architectures to test computational hypotheses about the origins of specific hallucinatory phenomenology and validate their results within appropriate hallucinatory populations. In this way, it would be possible to further close the loop between hallucinatory experiences and their computational mechanisms.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

The studies involving humans were approved by the University of Sussex, Research Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. Participants in both studies provided either written or verbal informed consent before participation.

Author contributions

KS conceived the study and developed the computational models. KS and DJS performed the experiments. All authors wrote the article and approved the submitted version.

Funding

All authors are grateful to the Dr. Mortimer and Theresa Sackler Foundation. AS is supported by European Research Council Advanced Investigator Grant CONSCIOUS (grant number 101019254).

Acknowledgments

The authors would like to thank Sarah Wolffe for assistance with data collection and Alexander Tschantz, Beren Millidge, and Cecelia Schwartzman for helpful discussions. We would like to thank Parkinson's UK and Esme's Umbrella for their invaluable assistance in helping recruit participants for this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2023.1159821/full#supplementary-material

References

Abbott, E. J., Connor, G. B., Artes, P. H., and Abadi, R. V. (2007). Visual loss and visual hallucinations in patients with age-related macular degeneration (Charles bonnet syndrome). *Invest. Ophthalmol. Vis. Sci.* 48, 1416–1423. doi: 10.1167/iovs.06-0942

Alonso, J. F., Romero, S., Angel Mañanas, M., and Riba, J. (2015). Serotonergic psychedelics temporarily modify information transfer in humans. *Int. J. Neuropsychopharmacol.* 18, 1–9. doi: 10.1093/ijnp/pyv039

Baltieri, M. (2020) A Bayesian perspective on classical control. In *2020 International Joint Conference on Neural Networks* (IJCNN) (pp. 1–8). IEEE.

Barnes, J., and David, A. S. (2001). Visual hallucinations in Parkinson's disease: a review and phenomenological survey. *J. Neurol. Neurosurg. Psychiatry* 70, 727–733. doi: 10.1136/jnnp.70.6.727

Bartunov, S., Santoro, A., Richards, B. A., Marris, L., Hinton, G. E., Brain, G., et al. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Adv Neural Inform Process Syst* 31

Bauer, S. M., Schanda, H., Karakula, H., Olajossy-Hilkesberger, L., Rudaleviciene, P., Okribelashvili, N., et al (2011). Culture and the prevalence of hallucinations in schizophrenia. *Compr Psychiatry.* 52, 319–325. doi: 10.1016/j.comppsych.2010.06.008

Blokpoel, M., Kwisthout, J., and van Rooij, I. (2012). When can predictive brains be truly Bayesian?. *Front. Psychology* 3:406. doi: 10.3389/fpsyg.2012.00406

Boubert, L., and Barnes, J. (2015). Phenomenology of visual hallucinations and their relationship to cognitive profile in Parkinson's disease patients. *SAGE Open* 5:215824401558582. doi: 10.1177/2158244015585827

Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., and Wiener, M. C. (2001). Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356, 299–330. doi: 10.1098/rstb.2000.0769

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for Core visual object recognition. *PLoS Comput. Biol.* 10:e1003963. doi: 10.1371/journal.pcbi.1003963

Carhart-Harris, R. L., and Friston, K. (2019). REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacol. Rev.* 71, 316–344. doi: 10.1124/pr.118.017160

Carhart-Harris, R. L., Kaelen, M., Bolstridge, M., Williams, T. M., Williams, L. T., Underwood, R., et al. (2016). The paradoxical psychological effects of lysergic acid diethylamide (LSD). *Psychol. Med.* 46, 1379–1390. doi: 10.1017/S0033291715002901

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., and Powers, A. R. (2019). Hallucinations and strong priors. *Trends Cogn. Sci.* 23, 114–127. doi: 10.1016/j.tics.2018.12.001

Cowan, J. D. (2015). "Geometric visual hallucinations and the structure of the visual cortex" in *The neuroscience of visual hallucinations first edit*. eds. D. Collerton, U. P. Mosimann and E. Perry (New Jersey, United States): John Wiley & Sons, Ltd), 219–254.

Díaz, J. L. (2010). Sacred plants and visionary consciousness. *Phenomenol. Cogn. Sci.* 9, 159–170. doi: 10.1007/s11097-010-9157-z

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Dosovitskiy, A., and Brox, T. (2016a). Generating images with perceptual similarity metrics based on deep networks. *Adv. Neural Inf. Proces. Syst.* 29, 658–666.

Dosovitskiy, A., and Brox, T. (2016b) Inverting visual representations with convolutional networks. In: *IEEE conference on Computer Vision & Pattern Recognition*.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. University of Montreal. 1341:1.

Fénelon, G., Mahieux, F., Huon, R., and Ziégler, M. (2000). Hallucinations in Parkinson's disease. Prevalence, phenomenology and risk factors. *Brain* 123, 733–745. doi: 10.1093/brain/123.4.733

Ffytche, D. H. (2005). Visual hallucinations and the Charles bonnet syndrome. *Curr. Psychiatry Rep.* 7, 168–179. doi: 10.1007/S11920-005-0050-3

Ffytche, D. H., and Howard, R. J. (1999). The perceptual consequences of visual loss: "positive" pathologies of vision. *Brain* 122, 1247–1260. doi: 10.1093/brain/122.7.1247

Fletcher, P. C., and Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* 10, 48–58. doi: 10.1038/nrn2536

Friston, K. J. (2005). Hallucinations and perceptual inference. *Behav. Brain Sci.* 28, 764–766. doi: 10.1017/S0140525X05290131

Friston, K., and Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Netw.* 22, 1093–1104. doi: 10.1016/j.neunet.2009.07.023

Frucht, S. J., and Bernsohn, L. (2002). Visual hallucinations in PD. *Neurology* 59:1965. doi: 10.1212/01.WNL.0000033279.00463.F8

Gould, C., Froese, T., Barrett, A. B., Ward, J., and Seth, A. K. (2014). An extended case study on the phenomenology of sequence-space synesthesia. *Front. Hum. Neurosci.* 8:433. doi: 10.3389/fnhum.2014.00433

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., et al. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* 10:4934. doi: 10.1038/s41467-019-12623-6

Harding, A. J., Broe, G. A., and Halliday, G. M. (2002). Visual hallucinations in Lewy body disease relate to Lewy bodies in the temporal lobe. *Brain* 125, 391–403. doi: 10.1093/brain/awf033

Hardstone, R., Zhu, M., Flinker, A., Melloni, L., Devore, S., Friedman, D., et al. (2021). Long-term priors influence visual perception through recruitment of long-range feedback. *Nat. Commun.* 12, 1–15. doi: 10.1038/s41467-021-26544-w

Hohwy, J., and Seth, A. K. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences.* 1:3.

Jeffreys, H. (1939). *Theory of Probability*, 1st ed. Oxford:The Clarendon Press.

Jin, A. (2014) Caffe: convolutional architecture for fast feature embedding. arXiv [Preprint].

Khaligh-Razavi, S.-M. M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10:e1003915. doi: 10.1371/journal.pcbi.1003915

Kluver, H. (1926). Mescal visions and eidetic vision. *Am. J. Psychol.* 37:502. doi: 10.2307/1414910

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Kometer, M., Schmidt, A., Jäncke, L., and Vollenweider, F. X. (2013). Activation of serotonin 2A receptors underlies the psilocybin-induced effects on α oscillations, N170 visual-evoked potentials, and visual hallucinations. *J. Neurosci.* 33, 10544–10551. doi: 10.1523/JNEUROSCI.3007-12.2013

Kometer, M., and Vollenweider, F. X. (2018). Serotonergic hallucinogen-induced visual perceptual alterations. *Curr. Top. Behav. Neurosci.* 36, 257–282. doi: 10.1007/7854_2016_461

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Kwisthout, J. (2014). Minimizing Relative Entropy in Hierarchical Predictive Coding, in Probabilistic Graphical Models, Lecture Notes in Computer Science. Eds GaagL. C. van der and A. J. Feelders. (Cham:Springer). 8754. doi: 10.1007/978-3-319-11433-0_17

Lee, T. S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20:1434. doi: 10.1364/josaa.20.001434

Leptourgos, P., Fortier-Davy, M., Carhart-Harris, R., Corlett, P. R., Dupuis, D., Halberstadt, A. L., et al. (2020). Hallucinations under psychedelics and in the schizophrenia Spectrum: an interdisciplinary and multiscale comparison. *Schizophr. Bull.* 46, 1396–1408. doi: 10.1093/schbul/sbaa117

Liechti, M. E. (2017). Modern clinical research on LSD. *Neuropsychopharmacology* 42, 2114–2127. doi: 10.1038/npp.2017.86

Lutz, A. (2002). Toward a Neurophenomenology as an account of generative passages: a first empirical case study. *Phenomenol. Cogn. Sci.* 1, 133–167. doi: 10.1023/A:1020320221083

Lutz, A., Lachaux, J. P., Martinerie, J., and Varela, F. J. (2002). Guiding the study of brain dynamics by using first-person data: synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proc. Natl. Acad. Sci.* 99, 1586–1591. doi: 10.1073/pnas.032658199

Mahendran, A., and Vedaldi, A. (2014). Understanding deep image representations by inverting them. ArXiv [Preprint], 1–9.

Mahendran, A., and Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vis.* 120, 233–255. doi: 10.1007/s11263-016-0911-8

Menon, G. J., Rahman, I., Menon, S. J., and Dutton, G. N. (2003). Complex visual hallucinations in the visually impaired: the Charles Bonnet Syndrome. *Surv Ophthalmol.* 48, 58–72. doi: 10.1016/s0039-6257(02)00414-9

Millidge, B., Tschantz, A., and Buckley, C. L. (2022) Predictive Coding Approximates Backprop Along Arbitrary Computation Graphs. *Neural Comput.* 34:13291368. doi: 10.1162/neco_a_01497

Mordvintsev, A., Olah, C., and Tyka, M. (2015). *Inceptionism: Going deeper into neural networks* Google Research Blog.

Mosimann, U. P., Rowan, E. N., Partington, C. E., Collerton, D., Littlewood, E., O'Brien, J. T., et al. (2006). Characteristics of visual hallucinations in Parkinson disease dementia and dementia with Lewy bodies. *Am. J. Geriatr. Psychiatry* 14, 153–160. doi: 10.1097/01.JGP.0000192480.89813.80

Muthukumaraswamy, S. D., Carhart-Harris, R. L., Moran, R. J., Brookes, M. J., Williams, T. M., Errtizoe, D., et al. (2013). Broadband cortical desynchronization underlies the human psychedelic state. *J. Neurosci.* 33, 15171–15183. doi: 10.1523/JNEUROSCI.2063-13.2013

Nair, A. G., Nair, A. G., Shah, B. R., and Gandhi, R. A. (2015). Seeing the unseen: Charles bonnet syndrome revisited. *Psychogeriatrics* 15, 204–208. doi: 10.1111/psyg.12091

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in Neural Information Processing Systems 29 (NIPS 2016).* Available at: https://proceedings.neurips.cc/paper/2016/hash/5d79099fcdf499f12b79770834c0164a-Abstract.html

Nichols, D. E. (2016). Psychedelics. *Pharmacol. Rev.* 68, 264–355. doi: 10.1124/pr.115.011478

O'Callaghan, C., Hall, J. M., Tomassini, A., Muller, A. J., Walpola, I. C., Moustafa, A. A., et al. (2017). Visual hallucinations are characterized by impaired sensory evidence accumulation: insights from hierarchical drift diffusion modeling in Parkinson's disease. *Biol Psychiatry Cogn Neurosci Neuroimag* 2, 680–688. doi: 10.1016/J.BPSC.2017.04.007

Papapetropoulos, S., Katzen, H., Schrag, A., Singer, C., Scanlon, B. K., Nation, D., et al. (2008). A questionnaire-based (UM-PDHQ) study of hallucinations in Parkinson's disease. *BMC Neurol.* 8:21. doi: 10.1186/1471-2377-8-21

Powers, A. R., Kelley, M., and Corlett, P. R. (2016). Hallucinations as top-down effects on perception. *Biol Psychiatry Cogn Neurosci Neuroimag* 1, 393–400. doi: 10.1016/j.bpsc.2016.04.003

Powers, A. R., Kelley, M. S., and Corlett, P. R. (2017). Varieties of voice-hearing: psychics and the psychosis continuum. *Schizophr. Bull.* 43, 84–98. doi: 10.1093/schbul/sbw133

Preller, K. H., and Vollenweider, F. X. (2018). Phenomenology, structure, and dynamic of psychedelic states. *Curr. Top. Behav. Neurosci.* 36, 221–256. doi: 10.1007/7854_2016_459

Ramstead, M. J., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., et al. (2022). From generative models to generative passages: a computational approach to (neuro) phenomenology. *Rev. Philos. Psychol.* 13, 829–857. doi: 10.1007/s13164-021-00604-y

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *IJCV* 115, 211–252. Available at: https://www.image-net.org/challenges/LSVRC/2012/index#cite

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X. (2016). "Improved techniques for training GANs," in *Proceedings of the 30th International Conference on Neural Information Processing Systems,* 2234–2242

Santhouse, A. M., Howard, R. J., and Ffytche, D. H. (2000). Visual hallucinatory syndromes and the anatomy of the visual brain. *Brain* 123, 2055–2064. doi: 10.1093/brain/123.10.2055

Sanz, C., Zamberlan, F., Erowid, E., Erowid, F., and Tagliazucchi, E. (2018). The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. *Front. Neurosci.* 12:7. doi: 10.3389/fnins.2018.00007

Schartner, M. M., and Timmermann, C. (2020). Neural network models for DMT-induced visual hallucinations. *Neurosci Conscious* 2020:niaa024. doi: 10.1093/nc/niaa024

Schmid, Y., Enzler, F., Gasser, P., Grouzmann, E., Preller, K. H., Vollenweider, F. X., et al (2015). Acute effects of lysergic acid diethylamide in healthy subjects. *Biol Psychiatry* 78, 544–553. doi: 10.1016/j.biopsych.2014.11.015

Schultz, G., and Melzack, R. (1991). The Charles bonnet syndrome: "phantom visual images". *Perception* 20, 809–825. doi: 10.1068/p200809

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn. Neurosci.* 5, 97–118. doi: 10.1080/17588928.2013.877880

Sexton, N. J., and Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Sci. Adv.* 8:eabm2219. doi: 10.1126/sciadv.abm2219

Shanon, B. (2002). Ayahuasca visualizations a structural typology. *J. Conscious. Stud.* 9, 3–30.

Shulgin, A. (1997). *TIHKAL: The continuation*. Berkeley, CA: Transform Press; (1997)

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. Iclr, 1. ArXiv [Preprint]. Available at: http://arxiv.org/abs/1312.6034

Strassmann, R. J., (2001) *DMT: The Spirit molecule. A Doctor's revolutionary research into the biology of near-death and mystical experience*. Rochester, VT: Park Street Press.

Strassman, R. J., Qualls, C. R., Uhlenhuth, E. H., and Kellner, R. (1994). Dose-response study of N,N-dimethyltryptamine in humans: II. Subjective effects and preliminary results of a new rating scale. *Arch. Gen. Psychiatry* 51, 98–108. doi: 10.1001/archpsyc.1994.03950020022002

Studerus, E., Kometer, M., Hasler, F., and Vollenweider, F. X. (2011). Acute, subacute and long-term subjective effects of psilocybin in healthy humans: a pooled analysis of experimental studies. *J. Psychopharmacol.* 25, 1434–1452. doi: 10.1177/0269881110382466

Suzuki, K., Roseboom, W., Schwartzman, D. J., and Seth, A. K. (2017). A deep-dream virtual reality platform for studying altered perceptual phenomenology. *Sci. Rep.* 7:15982. doi: 10.1038/s41598-017-16316-2

Swanson, L. R. (2018). Unifying theories of psychedelic drug effects. *Front Pharmacol* 9:172. doi: 10.3389/fphar.2018.00172

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* doi: 10.1109/CVPR.2015.7298594

Tani, J. (2016). *Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena (Oxford Series on Cognitive Models and Architectures)*, Oxford Univerisity press. Available at: https://www.amazon.co.jp/-/en/Jun-Tani/dp/0190281065

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Teunisse, R. J., Cruysberg, J. R., Hoefnagels, W. H., Verbeek, A. L., and Zitman, F. G. (1996). Visual hallucinations in psychologically normal people: Charles Bonnet's syndrome. *Lancet* 347, 794–797. doi: 10.1016/S0140-6736(96)90869-7

Timmermann, C., Roseman, L., Williams, L., Erritzoe, D., Martial, C., Cassol, H., et al. (2018). DMT models the near-death experience. *Front. Psychol.* 9:1424. doi: 10.3389/fpsyg.2018.01424

Tschantz, A., Millidge, B., Seth, A. K., and Buckley, C. L. (2023) Hybrid predictive coding: Inferring, fast and slow. *PLoS Computational Biology* 19:e1011280.

Tyler, C. W. (1978). Some new entoptic phenomena. *Vision Res.* 18, 1633–1639. doi: 10.1016/0042-6989(78)90255-9

Waters, F., Collerton, D., Ffytche, D. H., Jardri, R., Pins, D., Dudley, R., et al (2014). Visual hallucinations in the psychosis spectrum and comparative information from neurodegenerative disorders and eye disease. *Schizophr Bull.*40 Suppl 4:S233–245. doi: 10.1093/schbul/sbu036

Whittington, J. C. R., and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Comput.* 29, 1229–1262. doi: 10.1162/NECO_a_00949

Yacoub, R., and Ferrucci, S. (2011). Charles Bonnet syndrome. *Optometry* 82, 421–427. doi: 10.1016/j.optm.2010.11.014

Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* 10, 301–308. doi: 10.1016/j.tics.2006.05.002

Zarkali, A., Adams, R. A., Psarras, S., Leyland, L.-A., Rees, G., and Weil, R. S. (2019). Increased weighting on prior knowledge in Lewy body-associated visual hallucinations. *Brain Commun* 1:fcz007. doi: 10.1093/braincomms/fcz007

Glossary

Activation Maximisation: (ClassicalAM) - A method which maximises the activity of a particular neuron (group of neurons or an entire layer) within a (D)CNN by adjusting the network's input (e.g., through backpropagation), whilst keeping the weights and outputs of the network fixed. AM is a widely used approach to visualise the representations learned by a specific neuron (or neurons) within a (D) CNN. Because the form of AM that we use is restricted to a DCNN we refer to it as ClassicalAM.

Backpropagation - Short for 'backward propagation of errors'; backpropagation (or backprop) is a widely used algorithm for the training of artificial neural networks using gradient descent. Given a network and an error function, it enables the calculation of the gradient of the error function with respect to the weights in the network. Each weight is then adjusted in proportion to how much it contributes to the overall error so that the network improves its performance over successive iterations.

Computational (neuro)phenomenology - We use computational phenomenology to describe the use of computational modelling to model phenomenological properties, rather than (or in addition to) function or behaviour. This approach becomes computational neurophenomenology when the computational models can be related to neural mechanisms, either by interpretation or through generating experimental predictions (Suzuki et al., 2017; Ramstead et al., 2022).

(Deep) Convolutional Neural Network ((D)CNN) - A class of (deep) neural networks with an architecture specifically designed to process pixel-wise data of two-dimensional images. DCNNs are trained via backpropagation to classify the content of a set of training images into distinct categories, creating a feature map that summarises the presence of the detected features within an input image. DCNNs have been used primarily for applications such as object detection and image classification.

Deep Generator Network (DGN) - A class of deep neural network that is used to generate synthetic data. In this paper, we use the term DGN to refer to a specific type of deep generator network that has been pre-trained using the generative adversarial network (GAN) framework.

Deep Generator Network Activation Maximisation (GenerativeAM) - A novel form of AM proposed by Nguyen et al. (2016). Instead of directly optimising the image to maximise the activity of a target neuron within a DCNN as in ClassicalAM, Nguyen et al. (2016) used the learned natural image prior of a pre-trained DGN in combination with ClassicalAM to synthesise a new image that maximised activity in a target neuron within a DCNN, resulting in realistic and interpretable output images.

Error Function - A method used in machine learning to minimise the error between a predicted value and an actual value, which is commonly used to evaluate the performance of the model. For example, in supervised learning for image recognition, an error function calculates the discrepancy between the actual label of an image, and the model's estimation of this value. The network then attempts to minimise the error for each iteration using (for example) backpropagation.

Generative Adversarial Network (GAN) - The GAN is a learning framework in which a 'generator' and 'discriminator' network are pitted against each other in an adversarial manner. A generative neural network is used to generate synthetic data (e.g., images) that resemble a training data set. The discriminator network (e.g., a DCNN) is then trained to distinguish images produced by the DGN from those in the actual training data. The two networks are trained together in an adversarial zero-sum game in which the DGN tries to 'fool' the discriminator, whilst the discriminator tries to keep from being fooled. GANs are widely used in image processing to generate realistic synthetic images.

Gradient descent - A popular class of optimization algorithms in which the local minimum of a function is found by iteratively moving in the direction of the steepest descent as defined by the negative of the gradient. In machine learning, gradient descent is often used to update the parameters (weights) of a neural network, as for example in backpropagation.

Inception Score (IS) - A widely used objective approach to assess the realism of synthetic images produced by generative models, such as a GAN (Salimans et al., 2016).

Iteration - a term used in machine learning to indicate the number of times a function or process is repeated, in which the output of each step is used as the input for the next iteration.

Natural Image Prior - A common method used to constrain optimization of a (D)CNN towards generating synthetic images that resemble natural images, for example by minimising the colour difference between adjacent pixels (hand-designed natural image prior). Recently, Nguyen et al. (2016) used the learned natural image priors acquired by a DGN during GAN training to constrain the output of the DCNN to generate synthetic images that closely resembled natural images.