NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2024, **2024(1)**, niae005
DOI: https://doi.org/10.1093/nc/niae005
Research article

Special Issue: Experiencing Well-Being

Deep CANALs: a deep learning approach to refining the canalization theory of psychopathology

Arthur Juliani[1,*], Adam Safron[2], Ryota Kanai[3]

[1]Microsoft Research, Microsoft, 300 Lafayette St, New York, NY 10012, USA
[2]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, 600 N Wolfe St, Baltimore, MD 21205, USA
[3]Neurotechnology R & D Unit, Araya Inc, 6F Sanpo Sakuma Building, 1-11 Kandasakumacho, Chiyoda-ku, Tokyo 101-0025, Japan

*Corresponding author. Microsoft Research, Microsoft, 300 Lafayette St, NY 10012, USA. E-mail: ajuliani@microsoft.com

Abstract

Psychedelic therapy has seen a resurgence of interest in the last decade, with promising clinical outcomes for the treatment of a variety of psychopathologies. In response to this success, several theoretical models have been proposed to account for the positive therapeutic effects of psychedelics. One of the more prominent models is “RElaxed Beliefs Under pSychedelics,” which proposes that psychedelics act therapeutically by relaxing the strength of maladaptive high-level beliefs encoded in the brain. The more recent “CANAL” model of psychopathology builds on the explanatory framework of RElaxed Beliefs Under pSychedelics by proposing that canalization (the development of overly rigid belief landscapes) may be a primary factor in psychopathology. Here, we make use of learning theory in deep neural networks to develop a series of refinements to the original CANAL model. Our primary theoretical contribution is to disambiguate two separate optimization landscapes underlying belief representation in the brain and describe the unique pathologies which can arise from the canalization of each. Along each dimension, we identify pathologies of either too much or too little canalization, implying that the construct of canalization does not have a simple linear correlation with the presentation of psychopathology. In this expanded paradigm, we demonstrate the ability to make novel predictions regarding what aspects of psychopathology may be amenable to psychedelic therapy, as well as what forms of psychedelic therapy may ultimately be most beneficial for a given individual.

Keywords: machine learning; psychedelics; psychiatry; altered states; theories and models

Introduction

Psychedelic therapy is seen as an increasingly promising avenue for the treatment of a variety of psychopathologies (Sessa, 2018). Given the clinical success of drugs such as psilocybin in the treatment of disorders ranging from depression and anxiety to addiction (Bogenschutz et al., 2015; Griffiths et al., 2016; Johnson and Griffiths, 2017; Goldberg et al., 2020), it is natural to consider that there may be an underlying common cause behind these pathologies that psychedelic therapy is helping to address. Building on the RElaxed Beliefs Under pSychedelics (REBUS) model of psychedelic action, Carhart-Harris and Friston (2019) have proposed that what they refer to as “excessive canalization” may constitute the primary factor underlying all psychopathology (Carhart-Harris et al., 2022). Within this “CANAL” model of psychopathology, canalization is the development of overly precise or rigid beliefs about the state of the world, with beliefs being defined within the context of a hierarchical predictive processing (HPP) framework (Keller and Mrsic-Flogel, 2018). These overlying precise beliefs act as sticky attractors in a dynamical system sense, becoming engaged under a variety of different contexts, regardless of their appropriateness to the situation. They are also difficult to unlearn due to the high precision assigned to them by the underlying generative models which support them.

The CANAL model draws on evidence from psychedelic research, combined with various broader clinical observations to support the proposition that most psychopathologies can be understood from the perspective of canalization. The most clear examples of pathologies that fit the canalization paradigm are depression, anxiety, obsessive-compulsive disorder, and addiction. In each instance, there is a specific behavioral, cognitive, or emotional “mental circuit” which has been reinforced such that it is triggered in a variety of contexts for which it is not appropriate.

Received 17 May 2023; Revised 15 January 2024; accepted 19 January 2024

These circuits are equivalent to trajectories through the belief landscape, which have a high likelihood of being taken and indeed often have been taken at many points in the past. Triggering of the circuit then results in maladaptive cognition, affect, or behavior and likewise typically also leads to experienced suffering on the part of the individual. This maladaptivity comes from a mismatch between the high-level goals and intentions of the individual and the result of deploying these mental circuits. Importantly, the individual may even have metacognitive awareness of the maladaptive nature of their behavior, but still be unable to modify the underlying mental circuits that support it.

The phenomenon of canalization can be seen to involve the reuse of mental circuits, which may have once been adaptive, but no longer are in the current context, individuals find themselves in. Early experiences leading to the development of canalized mental circuits may be related to trauma, stress, or other forms of hardship (Kessler et al., 2010). Mental circuits may also simply be no longer adaptive due to a significantly large shift in the context an individual inhabits. This may include moving to another city or ending a long-term romantic relationship. Given that our living contexts change over time, often dramatically, it is clear that circuits once adaptive might later become maladaptive. The development of maladaptive circuits may also be due to overwhelming environmental factors that act more directly on physiological mechanisms. Yet even then, an account of maladaptive learning through attempted coping may be powerfully explanatory and also beneficial in both reducing stigmatization and foregrounding our common humanity (Neff, 2023).

The canalization model of psychopathology clearly has significant explanatory power, but is it sufficiently elaborated to cover the broad spectrum of possible mental disorders? In particular, can it make sense of potential pathologies of too little canalization or provide recommendations for when psychedelic therapy should be avoided as a treatment option? HPP and complex systems theory provide a useful set of tools to understand psychopathology and mental health (Hipólito et al., 2023; Girn et al., 2023). At the same time, there exists a parallel research direction taking advantage of advances in machine learning, and deep learning in particular (LeCun et al., 2015), which has seen significant success in contributing to neuroscientific progress (Marblestone et al., 2016; Richards et al., 2019) and progress in psychiatry in particular (Huys et al., 2016), in recent years. In this article, we take initial steps toward a reconciliation of these two approaches by proposing a set of refinements to the CANAL model, which take insights from deep learning theory. We believe that these refinements allow for a more nuanced understanding of the ways in which canalization manifests within learning systems and its relationship to psychopathology. The field of deep learning is particularly fruitful because it has been a domain within which adaptive behavior has been studied for many decades in the context of continuous or lifelong learning (Parisi et al., 2019) and involves the study of complex non-linear optimization dynamics, which also underpins learning in living organisms (Rabinovich et al., 2006).

Concretely, using a simple recurrent neural network (RNN) model (Barak, 2017), we demonstrate that the construct of canalization can describe two independent phenomena in any complex learning system, including the human brain. These are overfitting and plasticity loss (Ying, 2019; Lyle et al., 2023). We then connect this distinction to other relevant constructs in the larger psychological literature and discuss the implications for the clinical efficacy of psychedelic therapy and our understanding of psychopathology more broadly. We demonstrate that within our expanded framework, there are pathologies of either too much or too little canalization along each of two independent axes. By elucidating this refined perspective, which we refer to as “Deep CANALs,” we hope that a clearer understanding of the contexts in which psychedelic therapy may or may not be effective for a given individual can be made possible. We believe that the additional conceptual complexity introduced into the CANAL model in this work serves an important purpose toward the goal of ultimately leading to better treatment outcomes in the futureand likewise avoiding the rare but distressing potential negative effects of psychedelic use (Johnson et al., 2008; Bremler et al., 2023).

The predictive coding account of psychedelic action

Among the various computational models of the brain, HPP has had significant success explaining and predicting a diverse array of empirical phenomena (Walsh et al., 2020; Draganov et al., 2023), both at low and high levels of abstraction (Friston and Kiebel, 2009; Kanai et al., 2015) (although for a critique of the framework, see Bowers and Davis (2012)). According to HPP, the brain (and the cortex in particular) can be understood as a collection of hierarchically organized generative models. Each of these models has the objective of predicting the activity of hierarchically lower models (i.e. closer to primary modalities) that serve as their input. To the extent that the predictions do not account for the incoming information from the lower levels, an error signal is produced. These errors in prediction are then passed to hierarchically higher levels (e.g. deep association, transmodal, and/or frontal cortices) as inputs.

The learning objective that guides this process is the minimization of variational free energy (Friston, 2010; Kirchhoff et al., 2018), which corresponds to the prediction errors generated by the system at each level of representation. The ultimate source of entropy comes from outside the central nervous system, either from external sensory information in the world or from interoceptive information gathered from the body of the organism (Seth, 2013). This simple objective of predicting the activity of other generative models within the brain is said to make possible the complex representational capacities that humans are endowed with, as each functionally distinct generative model must develop sophisticated representations to use as the basis of prediction. The hierarchical nature of the system results in models at the lower levels of the hierarchy developing representations that are more concrete and at smaller spatio-temporal scales than those further up the hierarchy. HPP, building on Bayesian theory, refers to these representations as beliefs and to the strength with which they are encoded as precision.

Within the context of the HPP framework, it is possible to consider what happens to the finely tuned set of beliefs under the effect of psychedelics. Psychedelics are believed to exert their subjective (and therapeutic) effects primary through 5-HT$_{2A}$ agonism (Vollenweider et al., 1998). This agonism results in postsynaptic excitation, which is understood at the level of neuronal populations to lead to desynchronization and thus impaired function. We can start by considering regions of the brain that have densely expressed 5-HT$_{2A}$ receptors in their neuronal populations (Beliveau et al., 2017). The regions of particular clinical interest are the thalamus, the prefrontal cortex, and the claustrum.

The thalamus is responsible for the gating of sensory information to the cortex. It has been hypothesized that psychedelics disrupt the normal gating function of the thalamus, resulting in additional sensory information entering the cortex during the acute phase of drug effects (Vollenweider and Geyer, 2001). In an

HPP account, thalamic gating normally serves to reduce the precision of incoming information as a means of focusing limited attentional resources on the most behaviorally relevant information. When this function is disrupted, higher-precision sensory evidence enters deeper/higher portions of the system, producing larger prediction errors (relative to unaltered conditions) at various generative models within the hierarchy. While this is generally discussed in terms of allowing in more information from the world (cf. cleansing the doors of perception), thalamic gating could also be understood in terms of informational routing, where disrupting these functions could also result in unusual combinations of beliefs, so producing novel states of mind.

The effect of increased prediction errors within the representational hierarchy is further catalyzed by the effect of 5-HT_{2A} agonism on the prefrontal cortex, which is thought to represent high-level beliefs about the world, including various self-referential beliefs (Miller and Cohen, 2001). By dramatically exciting neurons in the prefrontal cortex (PFC), psychedelics can desynchronize their collective activity. This desynchronization effectively disrupts the neuron's ability to contribute to the representation of coherent beliefs about the self or other high-level abstract beliefs that the PFC participates in supporting (Millière et al., 2018). According to the REBUS model, this disruption is understood to prevent high-level generative models from making coherent predictions over beliefs at other representational levels, while simultaneously reducing the precision of those higher-level beliefs, potentially resulting in complex cascades of belief alterations.

The inability of the brain's high-level generative models to predict beliefs at lower levels corresponds to an inability to suppress those lower-level beliefs. The result is that the incoming evidence from lower-level beliefs has a stronger influence in updating beliefs at higher levels. This lack of coherence of high-level representations of beliefs may be further exacerbated by a disruption of the normal activity of the claustrum, leading to typically functionally unconnected regions of the predictive hierarchy entering into communication with one another. Although the REBUS model does not specifically address the role of the claustrum in these effects, other recent models have given this region of the brain a more central consideration (Doss et al., 2022). The effects of this process are measurable through an increase in the complexity and entropy of neural activity, as well as changes in functional connectivity that have been observed in functional magnetic resonance imaging (fMRI) studies of individuals under the effect of LSD, psilocybin, and N,N-Dimethyltryptamine (DMT) (Tagliazucchi et al., 2016). These acute changes have been referred to as "Temperature of Entropy-Mediated Plasticity" in the original CANAL formulation.

According to the REBUS model, the effect of this change in cortical dynamics is to relax the set of belief landscapes that the high-level areas of the cortex normally instantiate. This relaxation is hypothesized to be most clinically relevant at the highest levels of the representational hierarchy in the regions of the brain responsible for self-referential beliefs (Letheby and Gerrans, 2017). Within the HPP framework, this relaxation corresponds to a reduction in the precision of high-level beliefs encoded by generative models of the brain. Lower precision in a belief means that the generative model is less confident in it, making it more amenable to updating by conflicting beliefs at other levels of the hierarchy. In this way, psychedelics can open a window of time within which high-level beliefs can be more easily updated by evidence from the brain-external world, from both the exteroceptive senses and the interoceptive world of the body. As the effects of the drug wear off, high-level neuronal populations are able to effectively synchronize their activity once again, and the belief landscape they instantiate gradually returns to a state of encoding higher-precision beliefs once again. Although acute relaxation of beliefs during psychedelic use (and the mechanisms by which this occurs) remains largely hypothetical, there is increasing empirical evidence for some relaxation in the strength of beliefs after psychedelic therapy (Zeifman et al., 2022).

Despite the hypothetical nature of these mechanisms, considerable attention has been paid to developing them for use in clinical and nonacademic contexts. Using an analogy from metallurgy, this entire process has been popularly described as "neural annealing" (Johnson, 2019; Gómez-Emilsson, 2021; Carhart-Harris et al., 2022). This perspective implies that "heating up" the brain by increasing the entropy of the neural activity is in some way comparable to increasing the temperature of a metal in order to have it worked on and subsequently recooled. The "working of the metal" that is performed while the individual is in the acute or postacute psychedelic state is psychotherapy (of one form or another). As the drug wears off, the system is then "recooled" and returns to a less pliable state. Robust positive outcomes from psychedelic use often depend on their use in a clinical, or at the very least, therapeutically conducive setting (Yaden et al., 2022). As such, the nature of the "work" done to malleable beliefs during psychedelic therapy is of significant importance. See Figure 1 for a simple schematic representation of the purported effects of psychedelics on an idealized abstract belief landscape according to the REBUS model.

Two canalizations for two optimization landscapes

Expanding the conceptual framework used to understand psychopathology and psychedelic action from HPP and complexity theory to the additional explanatory framework provided by deep

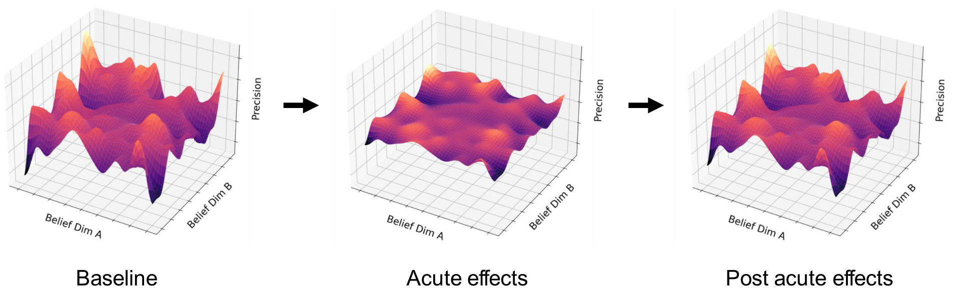


Figure 1. A diagram of the REBUS model of psychedelic action. The depth of the surface corresponds to precision of the set of beliefs. The acute effect of psychedelics is to relax the belief landscape by reducing the precision of the beliefs. This relaxation is hypothesized to persist to some degree during the postacute phase of the drug's effect as well, leading to lasting therapeutic benefit.

learning has the potential to provide a more nuanced understanding of both phenomena. A key example can be found in the recently introduced concept of "plasticity loss" in the literature on deep neural networks (Dohare et al., 2022; Lyle et al., 2023; Abbas et al., 2023). Plasticity loss is an empirical measure of the reduced ability of a neural network model to adapt to changes in the task distribution over time. Although often co-current, plasticity loss is distinct from overfitting (Ying, 2019), which refers to the extent to which a given model is incapable of generalization or of adequately responding in unfamiliar contexts. In empirical studies, plasticity loss has been shown to sometimes develop relatively early in the neural network learning process, leading to the characterization of "critical periods" or "primacy biases" in deep neural networks, both terms aptly borrowed from the literature on learning during early childhood development (Achille et al., 2019; Nikishin et al., 2022). Plasticity loss can also be understood as an inverse pathology to catastrophic forgetting (French, 1999; McCloskey and Cohen, 1989), which is the inability to retain critical information that was previously learned when learning new information.

In the original CANAL model, canalization is treated as a unified construct and is characterized as the primary causal mechanism that drives the presentation of a host of different psychopathologies (Carhart-Harris et al., 2022). We use a deep learning perspective to disambiguate between two different types of canalization. On the one hand, it is possible to interpret canalization as described in the original CANAL model as analogous to "overfitting" in machine learning. From this perspective, canalization is a measure of the rigidity of the belief landscape during inference, which may contain sticky attractors that are maladaptive, resulting in stereotyped pathological mental circuits. However, it is also possible to interpret canalization as described in the original CANAL model as analogous to "plasticity loss", in that it describes an inability of the underlying generative model to adapt or learn from new evidence over time. This would be characterized by the presence of difficult-to-escape local minima in the underlying optimization landscape of the model.

These two types of canalization are distinct because they describe properties of the topology of two related but unique optimization landscapes within a single dynamical system. We can illustrate this using a RNN as a simple model of the brain (or a network within the brain), as is often done in computational neuroscience (Mante et al., 2013; Barak, 2017).

In this simple model, the first type of canalization is a property of the optimization landscape induced by the activation pattern (h_t) in the hidden state of the RNN. This activation pattern is computed using both incoming sensory information (x_t) and the previous hidden state of the network (h_{t-1}). The process of computing this pattern within the predictive coding framework is equivalent to performing an inference procedure to minimize free energy, with h_{t-1} corresponding to the prior, x_t corresponding to the evidence, and h_t corresponding to the posterior in a given time step (t). Through this process, operative (and potentially subjectively experienced) beliefs are represented by the neural activity pattern at a given moment. The optimization landscape at this level is induced by the free energy minimization objective, with the depth of the landscape corresponding to the level of free energy for a given pattern of neural activation. We refer to this optimization landscape as Type A (Inference).

The second type of canalization is a property of the optimization landscape of the underlying synaptic weights (θ_t) that define the recurrence function of the RNN itself. Because these weights are being updated intermittently on the basis of the results of the inference process, we refer to this landscape as Type B (Learning). The depth of this landscape is determined on the basis of the free energy minimization objective, but at a longer timescale than the Type A landscape. This optimization landscape is also the one most familiar to machine learning practitioners, as it is the landscape within which traditional gradient descent through

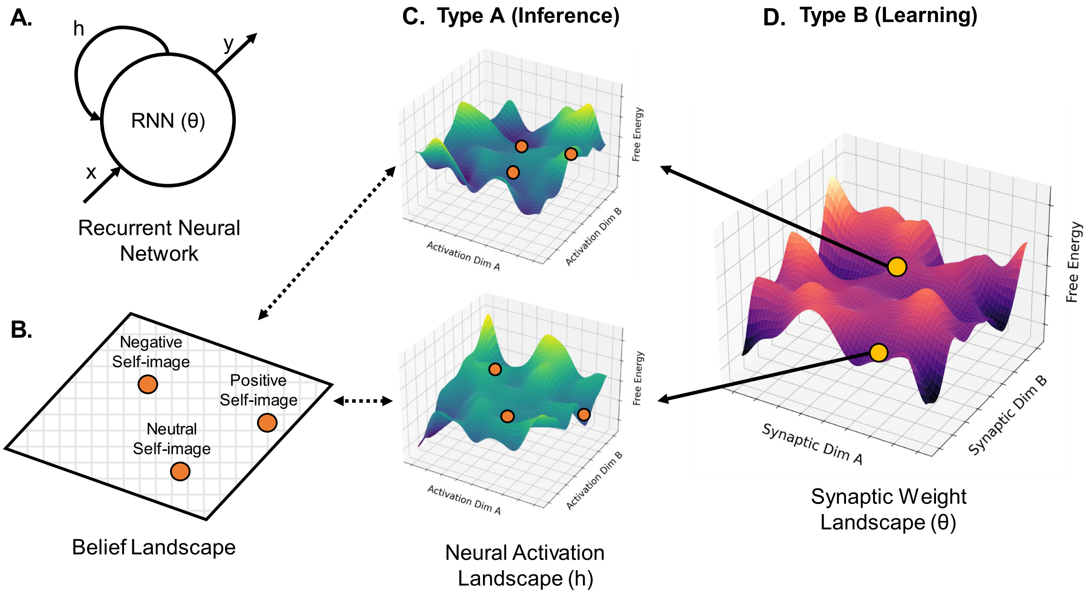


Figure 2. (a) In an RNN that performs free energy minimization, there are two different underlying optimization landscapes: one corresponding to the hidden activation pattern at a given time point (h) and another corresponding to the synaptic weight values which change with learning (θ). (b) Different points in an optimization landscape induced by a set of neural activity patterns will correspond to representing unique sets of beliefs, for example, a positive, neutral, or negative personal body image evaluation. (c) Actualizing these beliefs takes place through neural activity patterns (h), which are supported by the inference optimization landscape. The depth of the landscape corresponds to the free energy minimized by the set of instantiated beliefs. (d) The topology of the inference landscape is a function of the synaptic weight values (θ) defined by a learning optimization landscape. The depth of this landscape corresponds to the expected free energy minimized by the induced belief landscape.

backpropogation takes place. Although backpropogation is considered biologically implausible, there are other learning rules by which neurons can still instantiate a gradient-like learning process (Lillicrap et al., 2020).

In the context of an RNN, the nature of the induced landscape at the level of neural activity (Type A) is a function of both the current environmental context (made partially available to the model through a sequence of sensory observations x_t) and the state of the system within the synaptic weight landscape (θ_t). The inferred set of beliefs in a given instance are a function of the location of the system within the neural activity landscape (h_t). See Figure 2 for a diagram of these two optimization landscapes, a belief landscape, and their interaction.

The implications of this dissociation can be understood with a simple example. Consider an individual forming an evaluative belief about their body image over time. The position of neural activity in the Type A landscape at a given moment (h_t) corresponds to the immediate, currently instantiated belief about self-image. In a given Type A landscape, certain locations may correspond to beliefs that are more or less "canalized" in that they are represented by basins of steeper gradients. If neural activity enters one of these basins of "canalized belief," it is less likely to escape in the absence of a change in environmental context. This would correspond to the individual "getting stuck" in a particular self-evaluation, which is often consistent with ruminative thinking (Nolen-Hoeksema et al., 2008). This phenomena may also present itself in the cycling between a small subset of attractor basins, each of which represents a related canalized belief.

The topography of the Type A landscape is a function of both the environmental context and the location of synaptic weights (θ_t) in the Type B landscape. Because of this, the Type B landscape corresponds to the individuals' more general tendency to get stuck or not in specific kinds of self-evaluation. Canalization of this landscape then describes the longer-term tendency of the individual to make certain kinds of self-image evaluations. The implications of canalization in Type A and Type B landscapes can be broadly understood as a state-trait distinction (Fridhandler, 1986), with h determining the immediate-timescale state (instantiated belief) and θ determining the longer-timescale trait (set of possible beliefs).

The two types of canalization have different pathological profiles. Given an agent (artificial or natural) that continuously learns over time, being overfit to a previous context (Type A canalization) that is no longer realized will likely lead to both suboptimal behavior and some level of associated stress (Bennett et al., 2022). If the agent is able to adapt to the new context quickly enough to resolve the stress, then we would not consider it pathological. If, however, the agent both is overfit and suffers from plasticity loss (Type B canalization), this would present a greater difficulty because the former is simply the state of not being adapted to a new context, which may be addressed with learning, whereas the latter is the state of no longer being capable of adapting to new contexts. In cases in which both forms of canalization are present, we would describe the agent as being in a pathological situation. Such situations may be difficult to escape from without the aid of an external intervention of some sort, such as psychedelic-assisted psychotherapy.

The two types of canalization interact with one another in complex ways. For example, a learning system suffering from pathological Type A canalization would find it more difficult to obtain the experiences necessary to adapt to a new environmental context, even if there is no significant loss of plasticity. This is due to a reduction in the diversity of realizable mental circuits. An individual suffering from Type B canalization may be unable to learn more adaptive mental circuits even if their current repertoire is significantly flexible to enable acquisition of experiences that would allow an otherwise less canalized individual to learn the necessary information. More often, as happens in many cases of severe psychopathology, there is canalization of both Type A and Type B. This results in stereotypical and maladaptive mental circuits that cannot be corrected despite interventions that serve to provide the necessary evidence to update them, such as traditional psychotherapy. Within the context of deep learning theory, there is some evidence to suggest that, while dissociable, the two constructs are often correlated in practice, with smoother Type B optimization landscapes often producing models that are also better able to generalize (Li et al., 2018).

It is also worth noting that given the hierarchical nature of prediction within the brain, it is possible that different generative models (represented by distinct functional networks) at different levels of abstraction are more or less canalized to different extents. Because of this, it is inappropriate to describe the whole brain as canalized or not in the Type A or Type B sense. Additionally, canalization can have different clinical implications at different levels of the predictive hierarchy. For example, the low-level visual system is canalized in both the Type A and Type B senses in adults, and given the fact that the statistics of the visual world do not meaningfully change over the course of a lifetime, this is considered a desirable rather than pathological property. Indeed, when this canalization is disrupted, it can cause undesirable changes in visual perception such as hallucinogen-persistent perception disorder (Martinotti et al., 2018). In contrast, canalization of either type at higher levels of the predictive hierarchy is proposed by the CANAL model to be related to pathologies of self-referential processing and therefore a key target for psychedelic therapy (Letheby and Gerrans, 2017).

Relation to other psychological constructs

Delineating the two independent optimization landscapes that support belief representation in the brain allows us to make additional connections between the construct of canalization and other well-studied constructs in psychology and the cognitive sciences. There is a rich literature studying the relationship between psychopathology and personality theory (e.g. Clark (2005); Tackett (2006)). Although the original CANALs model does not draw explicit connections between personality constructs and canalization, it is possible to do so in our expanded framework.

We have reason to believe that there is not a simple linear relationship between decreases in canalization and positive mental health outcomes. To understand why, we examine the two types of optimization landscape through the lens of cybernetic personality theory (DeYoung, 2015; Safron and DeYoung, 2021), a popular model of personality factors, and their relationship to psychopathology that uses principles from cybernetic control theory. In this theoretical framework, the Big Five personality traits are grouped into two higher-level meta-traits: "stability" and "plasticity". Stability represents the shared variance of the three traits of neuroticism, agreeability, and conscientiousness, while plasticity represents the shared variance of extraversion and openness.

The stability factor is correlated with the topology of the Type A belief landscape. In our model, both overfitting and underfitting in this landscape type correspond to a decrease in trait stability. This results in a u-shaped relationship between canalization and stability. This may at first seem unintuitive, since a highly canalized belief landscape could be thought to produce

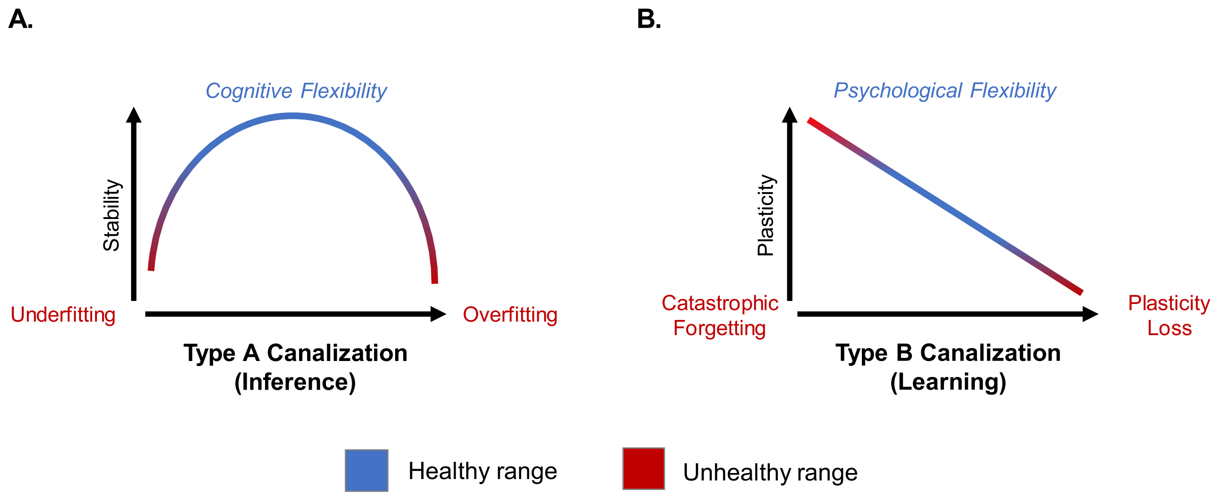


Figure 3. In the context of optimizing a deep neural network to perform optimally under a sequence of tasks, canalization exists along two independent dimensions. (a) Along the Type A (Inference) dimension, the pathologies of underfitting and overfitting present at either extreme. The meta-trait of "stability" corresponds to a balance between these two extremes and correlated with cognitive flexibility. (b) Along the Type B (Learning) dimension, catastrophic forgetting and plasticity loss are the pathologies manifest at either extreme. The meta-trait of "plasticity" is inversely related to canalization in this landscape type. The construct of psychological flexibility corresponds to a balance between the two extremes.

a "stable" set of encoded beliefs. Stability (the meta-trait), however, is better understood as the ability of the system to maintain a meta-stability between these two extremes of overfitting and underfitting and as a result be capable of responding flexibly to a variety of potential environmental contexts (Tognoli and Kelso, 2014; Hipólito et al., 2023). In the context of dynamical systems, meta-stability refers to a state where the system temporarily resides within a local energy minimum, showing resistance to perturbations, but can be transitioned out of this state by sufficiently large disturbances to reach a more globally stable state. Neural markers of meta-stability have also been inversely correlated with psychopathology (Lee et al., 2018), providing additional evidence for the central role of trait stability in mental health. Of the traits that make up stability, neuroticism is negatively correlated, while agreeableness and contentiousness are positively correlated.

The extent of canalization in Type B landscapes corresponds to the inverse of the plasticity meta-trait. Unlike stability, which follows a u-shaped curve in describing the balance between overfitting and underfitting, plasticity correlates with inverse Type B canalization in a straightforwardly linear fashion. The more the canalization that is present in a Type B landscape, the less likely an individual is to display meta-trait plasticity. We derive this linear relationship between these two factors based on the findings that psychedelic use is associated with increases in numerous measures of neural plasticity (Calder and Hasler, 2023) and increases in both trait openness (Lebedev et al., 2016; Erritzoe et al., 2019) and trait extraversion (Erritzoe et al., 2018).

Although decreases in Type B canalization (and corresponding increases in plasticity) may be considered desirable, a fine balance must be maintained (DeYoung and Krueger, 2018). Too little canalization in Type B landscapes would correspond to catastrophic forgetting, which may be related to the manifestation of certain forms of psychosis or depersonalization (for a more extensive discussion, see Section "Broader clinical implications"). In contrast, too much canalization in a Type B landscape corresponds to loss of plasticity, which is characterized most clearly in certain neurodegenerative disorders (Bossy-Wetzel et al., 2004; Vyas et al., 2016), but is practically associated with various psychopathologies as well. See Figure 3 for a diagrammatic representation of the relationship between the two meta-traits and the two optimization landscape types.

In addition to a connection to the two meta-traits of cybernetic personality theory, it is also possible to understand Type A and Type B canalization in light of another set of related constructs: cognitive and psychological flexibility (Ionescu, 2012; Kashdan and Rottenberg, 2010). Both measures have been shown to improve in people who undergo psychedelic therapy (Davis et al., 2020; Doss et al., 2021). Cognitive flexibility describes the ability of an individual to rapidly adapt to the current context by selecting the appropriate behavioral schema, which may or may not already have been learned in the past. In contrast, the construct of psychological flexibility describes the ability to adapt to changes in life circumstances over longer timescales. Cognitive flexibility can be associated with stability in the Type A optimization landscape. There is some preliminary evidence for the link between the meta-stability of neural dynamics and cognitive flexibility (Hellyer et al., 2015). Psychological flexibility can be linked to the maintenance of a healthy level of plasticity in the Type B landscape. Using the framework presented here, it is possible to begin to dissociate the potential mechanisms behind which changes to each measure take place.

Psychedelic action on optimization landscapes

Given the different substrates of the two optimization landscapes, the way in which psychedelics impact them likewise differs. In agreement with the original CANAL model, we hypothesize that the impact of psychedelics on the Type A landscape is the result of the acute effects of 5-HT_{2A} agonists producing excitation of cortical neurons along with the associated short-term changes that occur through Hebbian plasticity and follow this excitation. Rather than strictly relaxing the precision of encoded beliefs, these changes serve to destabilize belief representation (Hipólito et al., 2023). This destabilization can manifest itself acutely as a mixture of strengthened and relaxed beliefs over the course of the acute drug effect (Safron, 2020). These acute changes in belief

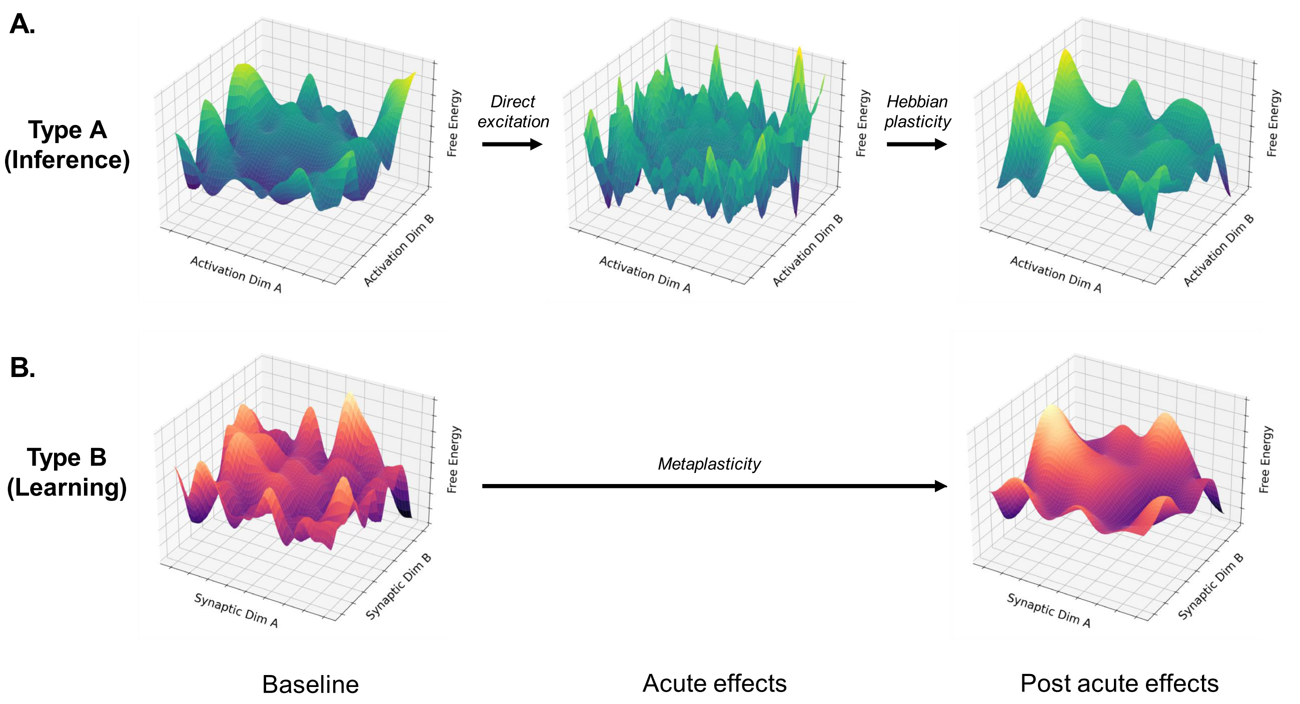


Figure 4. The impact of psychedelics on both types of optimization landscapes. (a) Acute effects of psychedelics perturb the neural activation (Type A) landscape, resulting in transiently strengthened and relaxed beliefs represented by the gradient magnitude across the optimization landscape. Postacute effects result in changes to landscape topology, which may involve an overall flattening of the landscape. (b) Postacute effects of psychedelics result in a smoothing of the synaptic weight (Type B) optimization landscape. This effect is mediated by changes in metaplasticity and dendritic growth.

representation are measurable in neural markers of entropy that have been consistently reported in the literature (Carhart-Harris et al., 2014). In our framework, these changes correspond to changes in the topology of Type A landscapes which are mediated by changes in synaptic weight parameters (θ), but not a change in the topology of the Type B landscape itself.

In agreement with the original CANAL model, we hypothesize that the effects of psychedelics on Type B optimization landscapes occur through postacute changes in metaplasticity caused by increases in various neurogenerative factors that affect cortical neurons, including increases in brain-derived neurotrophic factor (BDNF) and dendritic spine growth (Calder and Hasler, 2023). Metaplasticity describes the second-order plasticity in a system or the extent to which the level of plasticity in a network is itself plastic. Therefore, long-term changes in neural growth factor concentrations are likely to produce changes in metaplasticity. Due to the extended timescale at which these effects take place, it is likely that there is a limited role for potential changes in Type B landscapes in the acute phase of the drug effects, but the postacute effects can last for days to weeks. See Figure 4 for a diagram of the hypothesized effects on both Type A and Type B optimization landscapes that occur as a result of the use of psychedelics.

Due to the different mechanisms of action at different timescales, we can expect different impacts on the associated canalization of the two different types of landscape. For Type A landscapes, psychedelics serve to destabilize attractors, resulting in both a transient relaxation of previously canalized beliefs and the potential for transiently introduced novel beliefs. In our terminology, this implies that psychedelics are capable of reducing both overfitting and underfitting in Type A belief landscapes.

The evolution of the dynamics in this high-entropy state is significantly more context sensitive than the system would be when in a normal state, hence the need for an appropriate "set and setting," which influences the extent to which relaxation or strengthening of beliefs may take place and the nature of the strengthened beliefs when they do occur (Hartogsohn, 2016). It is this context sensitivity that may explain the tendency for more relaxation-like (and thus canalization-reducing) effects to be reported in the clinical literature relative to the recreational use literature (Ballentine et al., 2022).

The loss of plasticity in deep neural networks has been associated with the emergence of so-called "dead units," synaptic weights that no longer receive any useful gradient signal (Lyle et al., 2023). Given the connection between gradient-based learning and neural plasticity (Richards and Kording, 2023), we can interpret loss of dendritic volume in individuals with various psychopathologies as analogous to this development of "dead units" in artificial neural networks (Qiao et al., 2016). Similarly, we can understand the growth of dendritic volume after exposure to a psychedelic as enabling the flow of a learning gradient between synapses again. Interestingly, two of the most common approaches to addressing plasticity loss in deep neural networks are to reset the weights of the final layer or to shrink and perturb the weights of the network (Ash and Adams, 2020; Nikishin et al., 2022; Lyle et al., 2023). Both these can be understood to be inducing an analogous flattening of the optimization landscape, as may be taking place in humans under the effects of psychedelics.

In Type B landscapes, we hypothesize that psychedelics act directly to increase plasticity and thus reduce canalization. Given the second-order changes in plasticity that take place in the postacute phase of psychedelic use, there is also a potential role for

Table 1. A summary of the unique properties which characterize the Type A (Inference) and Type B (Learning) optimization landscapes along with the effects that psychedelics are hypothesized to have on each.

Property	Type A (Inference)	Type B (Learning)
Optimization landscape	Neural activity patterns	Synaptic weights
RNN equivalent term	Hidden state (h)	Network parameters (θ)
Over-canalization expressed as	Stereotyped mental circuits	Inability to learn or adapt mental circuits to context changes
Under-canalization expressed as	Inconsistently deployed mental circuits	Inability to retain previously adaptive mental circuits
Over-canalization described by	Overfitting	Plasticity loss
Under-canalization described by	Underfitting	Catastrophic forgetting
Associated meta-trait	Stability	Plasticity
Associated flexibility	Cognitive	Psychological
Psychedelic impact via	Acute effects	Postacute effects
Neural mechanism underlying psychedelic action	Direct excitation; Hebbian plasticity	Metaplasticity
Primary result of psychedelic action	Mixed relaxation and strengthening of canalization	Relaxation of canalization

the consideration of "set and setting" here as well, though one which takes place on a longer timescale. Increases in plasticity correspond to greater environmental sensitivity, and there is evidence for the reopening of critical learning periods in the weeks after psychedelic use (Nardou et al., 2023). This points to the importance of cultivating learning experiences in this period of time that are conducive to longer-term adaptedness and adaptivity for the individual even after plasticity levels have returned to baseline. The reopening of critical periods also brings with it the possibility of maladaptive learning as well, which likewise has the potential to be recanalized when the window closes. Because of this increased sensitivity, it may be valuable to explore the role of postacute experience in the outcomes of psychedelic-assisted therapy.

According to the original CANAL model, the degree to which metaplasticity was present in the system was associated with an increase in the learning rate (Carhart-Harris et al., 2022). From a deep learning perspective, the learning rate is independent of catastrophic forgetting or loss of plasticity. This is because both are related to the topology of the underlying optimization landscape, not to the speed at which the landscape is traversed during learning. We can understand increases in prediction errors, which are an acute effect of psychedelics, to produce increases in the magnitude of the gradient during learning, but this outcome could arise in the absence of any change in the learning rate. Indeed, increases or decreases in learning rate are likely to impact adaptation and may be related to certain psychopathologies, but they are not necessarily a function of canalization as we understand it here. More relevant to adaptation and generalization is the presence of local minima in Type A or Type B landscapes. Smoothing the topology of the optimization surface can then produce more successful learning, even when the learning rate is kept fixed.

In light of the dramatic plasticity-inducing effects of psychedelics, even in the absence of any subjective guidance (Ly et al., 2018), it is possible that structural changes in the brain (corresponding to changes in the topology of Type B landscapes) are a significant causal factor behind subsequent improvements in mental health. This perspective has guided interest in the development of non-hallucinatory psychedelic drugs (Cameron et al., 2021). However, this is complicated by the fact that the specific mental circuits involved during the acute and postacute effects of psychedelic therapy appear to be essential for meaningful positive outcomes (Yaden and Griffiths, 2020). Together, these pieces of evidence point to a critical therapeutic role for changes in both Type A and Type B optimization landscapes that support the representation of beliefs in the brain. Table 1 provides a summary of the two optimization landscapes that we have outlined here, as well as their relationship to the measures discussed earlier.

Returning to the "neural annealing" analogy, we can attempt to reinterpret it in light of the distinct optimization landscape types of our framework. Given the connection between the entropy of neural activity and the "heating" of the system, it initially seems the most simple to interpret neural annealing as applying to the Type A landscape. On the other hand, the aspect of the metaphor related to a structural crystallization of desirable system dynamics may best apply to changes in synaptic weights brought about by metaplasticity and downstream changes in neurotrophic factors. This would suggest an application to Type B landscapes. The mixed nature of this metaphor is acknowledged in the original CANAL model, where it is suggested that acute increases in entropy lead to postacute increases in metaplasticity (Carhart-Harris et al., 2022). This complexity points to the fact that although "neural annealing" has been useful in educating the general public, it may lack the sophistication necessary to fully capture the complexity underlying the therapeutic effects of psychedelics.

Destabilization of belief representation

According to the REBUS model, psychedelics alter the belief landscapes of the brain in a very specific way: they flatten them. Although there is considerable evidence for the relaxation of beliefs in the context of other psychotherapeutic interventions, such as meditation (Laukkonen and Slagter, 2021), there is mixed evidence in the case of psychedelics. An alternative framework (Altered Beliefs Under Psychedelics) has been proposed that hypothesizes that high-dose psychedelics act to destabilize beliefs through transient perturbation during the acute action of the drug (Safron, 2020). Altered Beliefs Under Psychedelics predicts that although destabilization can potentially produce on average an effect of relaxation, it is not guaranteed that this will be the outcome, particularly at particularly large doses. Although empirical evidence is still being accumulated, the framework of destabilization as a basis for understanding alterations in belief representation with psychedelics has also recently been proposed elsewhere (Hipólito et al., 2023).

Understanding psychedelic action through the lens of belief destabilization in HPP allows for a wider range of possible outcomes, with respect to both the subjective effects of the drug and the ultimate clinical outcomes. These include the recognition that, under the influence of psychedelics, people have the potential to acquire new beliefs that they had not previously held (Griffiths et al., 2019). Returning for a moment to the analogy provided by "neural annealing" (Johnson, 2019), rather than heating a metal to the point of becoming more malleable, high-dose psychedelic action on the Type A optimization landscape may correspond to heating a metal to the point of boiling and spontaneously taking on novel configurations in the process. We can utilize the framework of two optimization landscapes to further develop this account in a more theoretically grounded manner.

Examples of undesirable subjective effects resulting from psychedelic use are often more consistent with belief destabilization than idealized therapeutic outcomes of ego dissolution or mystical experience (Bremler et al., 2023). One subjectively challenging manifestation of destabilization is so-called cyclic thinking (Watkins, 2008), in which an individual may repeat a single chain of thought multiple times over the course of minutes. This experience is accompanied by a felt sense of inability to control one's' thoughts, as well as a negative affect. Cyclic thinking can be modeled as an instance of strong transient canalization being induced in the Type A optimization landscape, with a novel attractor developing which the neural dynamics are unable to escape from, unless further perturbation of the landscape takes place.

Another example of psychedelic-induced canalization in the Type A landscape is the development of strongly felt novel metaphysical beliefs, such as the belief in supernatural or mystical phenomena. For example, it is common for individuals to report the development of beliefs such as "the universe being made of love," which are described with a confidence ascribed to them that is not typical, even of beliefs that are normally strongly felt during normal conscious experience (Griffiths et al., 2019). It is difficult to reconcile the felt conviction with which individuals report the cognitive and affective content of mystical experiences with a perspective in which the precision of high-level beliefs is simply relaxed. The development or strengthening of supernatural beliefs has been of enough potential concern to be a recent topic of serious philosophical scholarship (Letheby, 2021). Although it may be the case that, on average or in the aftermath of a psychedelic experience, the Type A belief landscape has been flattened, it is likely that a more nuanced account than REBUS is necessary to capture the full range of reported phenomena.

Differences in defining psychedelic action in terms of relaxation, strengthening, altering, or destabilizing have meaningful clinical implications. Despite this, the possibility that psychedelics can be understood as perturbing belief landscapes and destabilizing previous attractor dynamics in addition to (or instead of) flattening landscapes does not diminish the therapeutic potential of this drug class in any way. Similar to relaxation, the destabilization of a belief landscape allows an individual to explore different possible sets of beliefs and behavioral, cognitive, and emotional circuits that one may have never been in a position to conceptualize before. Although it may feel jarring for an individual to find themselves experiencing perceptual, cognitive, or affective beliefs with high precision that they had not previously experienced as such, if the belief enables a more adaptive relationship between the individual and their environment, then the clinical outcome may be positive.

Even in cases where an individual might represent high-precision beliefs that can be distressing, the experience of the transience of these beliefs coupled with sufficient psychotherapeutic support can also contribute to ultimately positive outcomes (Letheby, 2021; Yaden and Griffiths, 2020). In fact, recent evidence suggests that therapeutic outcomes of psychedelic use are predicted by reductions in experiential avoidance (Zeifman et al., 2023), and confronting difficult experiences directly during the acute phase of the drug is a likely mechanism behind this effect.

Destabilization can also allow positively valenced and adaptive beliefs to be instantiated and ultimately recanalized through the process of psychological integration. It is this property of belief perturbation that enables beneficial psychotherapeutic interactions during and after psychedelic use. Engagement with a trained therapist can allow movement through the belief landscape to be interpreted and grounded. Postacute integration can allow more exotic beliefs, which may develop, such as "archetypal" experiences, or encounters (or even identification) with devas or spirits (Lutkajtis, 2021), to be incorporated into a beneficial and healthy set of later recanalized mental circuits which serve the individual in their normal everyday life.

Psychopathologies of reduced stability

The CANAL model proposes that excessive canalization is potentially the *p*-factor (Caspi et al., 2014) or primary factor underlying all psychopathology. If we understand canalization as describing the stability and predictability in the development and deployment of mental circuits over time, there is a class of psychopathologies that can be understood not to result from too much canalization, but rather from too little (DeYoung and Krueger, 2018). Examples of this include schizophrenia, borderline personality disorder, and bipolar disorder among these. In each case, the individual suffers from a lack of coherent, stable, and adaptive behavioral, cognitive, and emotional circuits (MacKinnon and Pies, 2006; Winterer et al., 2006; Schmack et al., 2015; Lozano et al., 2016). In the original CANAL model, the stereotyped beliefs associated with psychosis were used to suggest that even these pathologies may be understood to arise from a kind of canalization. Using the Deep CANALs framework described earlier, we can consider this proposal in light of the unique properties of both Type A and Type B optimization landscapes. Applying this lens, we find that a psychopathology such as schizophrenia may indeed be characterized by some form of canalization in certain Type A landscapes, corresponding to overfitting (and subsequently stereotyped deployment of mental circuits). However, it can also be described from the perspective of the Type B landscape as indicating insufficient canalization. In individuals experiencing psychosis, there is a catastrophic forgetting of the repertoire of previously adaptive mental circuits that an individual is no longer capable of deploying or in some cases of ever recovering.

This lack of stability results in a maladaptivity to the environment and corresponding stress for the individual. For such individuals, psychedelic therapy, if it truly acted to unilaterally reduce canalization, could potentially exacerbate their symptoms instead of alleviating them. Indeed, the clinical literature contains a number of cases of individuals with predispositions to these pathologies of undercanalization developing symptoms of psychosis or dissociation as a result of a psychedelic experience (Krebs and Johansen, 2013; Bremler et al., 2023). The careful screening used in clinical trials of psychedelic therapies can greatly reduce the likelihood of these negative outcomes. However, as the availability of psychedelic drugs increases in the coming years, it is critical that a theory of the action of these drugs on the brain is able to explain their effects both within and outside of clinical contexts. Indeed, one of the criteria by which the robustness of a theory can be measured is its ability to capture the full spectrum of possible phenomena it claims to account for (Wimsatt et al., 1981), and we believe that a theory of psychedelic action should meet such a bar, even if the phenomena to be explained are relative outliers.

There are also individuals who are psychologically healthy and do not have a severe psychopathology that requires clinical intervention. If we examine these individuals, we find that rather than lacking stable and reliable beliefs, we could instead describe them as canalized in healthy and adaptive ways to their environmental context (Olaru et al., 2023). This sculpted set of beliefs is often the result of a complex series of carefully tuned environmental factors, including supportive parenting, education, socialization, and lack of childhood trauma over the course of development. For these individuals, it may be undesirable to dramatically disrupt their carefully crafted belief landscapes, as any change has the potential to produce a less adaptive set of mental circuits than those they currently possess. Although it could be argued that such highly functioning individuals could enjoy even greater states of well-being by using psychedelics (cf. "betterment of well people"), the potential for disrupting existing adaptedness is a possibility that deserves at least some consideration.

We can understand this potential risk from an optimization perspective. Within the HPP framework, a psychologically healthy and adapted individual has arrived at a set of tuned beliefs which are at near-global minima in the optimization landscape underlying those beliefs. As such, movement in any direction from the minima may result in decreased adaptability. This danger is especially apparent when there is the real possibility of inducing trauma (and thus significantly pushing the individual away from the global minima) through a so-called "bad trip." Although such highly adverse experiences can sometimes be the seed of positive change, potentially involving therapeutic experiences of "surrender" as a kind of deep acceptance, risks should be considered seriously (including with respect to seemingly positive stances such as those involving radical acceptance) (Safron and Johnson, 2022). The likelihood of such an event can be significantly reduced through a supportive therapeutic context, but even in these cases, there is little reason to believe that the changed belief landscape will strictly be more optimal than it was before the intervention.

Potential benefits of a plasticity bias

If we assume that the environment within which individuals live is completely fixed, then it is reasonable to conclude that anyone who is perfectly adapted (near global minima in their belief landscape) is in need of no additional psychological support. However, from the perspective of an environmental context that is susceptible to change or disruption over time, it becomes desirable for an individual to be not only adapted but also adaptable (Stanley and Lehman, 2015). This means being biased toward slightly less canalization in order to be receptive to the changes which one will certainly eventually encounter during a lifetime. In the postindustrial world, social, professional, and cultural dynamics are changing at a rapid pace, which is unique in human history. In this context, some amount of adaptability is almost certainly necessary in order to ever hope to arrive near the global minima of the belief landscape in any given future context. Importantly, the adaptability in which we are interested corresponds as much to the absence of plasticity loss in the Type B landscape as to a lack of overfitting in the Type A landscape.

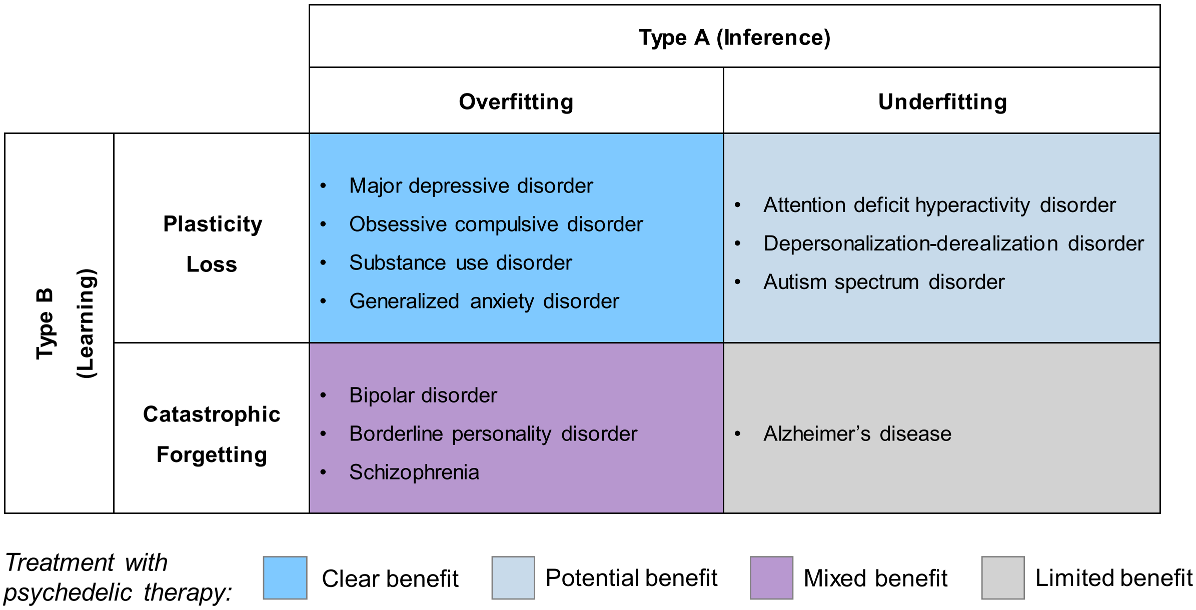
We can see the benefits of a bias toward adaptation in the deep reinforcement learning literature, where maximum entropy learning is provably more optimal than simple reward maximization (Ziebart et al., 2008; Levine, 2018) and has likewise become a common component of state-of-the-art systems (Schulman et al., 2017; Haarnoja et al., 2018). In maximum-entropy learning, an agent optimizes both the objective of accumulating the largest possible return and the objective of maximizing the entropy in the distribution over possible actions. Importantly, the agent attempts to maximize these objectives not only with respect to the immediate future but also over a long time horizon. Balancing these two objectives during the optimization process ensures that if and when the environmental context changes, the agent is less likely to have overfit to the previous context, resulting in better learning performance over time.

As discussed previously, the kinds of radical changes in beliefs that can occur with full-dose psychedelic therapy may not necessarily be helpful for many mentally healthy individuals. In such cases, it may actually be that only a minimal decrease in the canalization of the Type B landscape is necessary, one that balances the local greedy optimization of the belief landscape for the given context with the reality that the context will likely shift in the future. If we understand psychedelics to act as relaxers of Type B belief landscapes through the downstream effects of 5-HT$_{2A}$ agonism, this bias of the optimization process toward entropy could be brought about by microdosing a 5-HT$_{2A}$ agonist or taking a full dose of a 5-HT$_{2A}$ partial agonist such as Ariadne (Cunningham et al., 2023) at a regular interval. Such a minor intervention may be sufficient to tip the balance toward a healthy level of plasticity that allows the preservation of the adaptive circuits which an individual has developed over time. This kind of intervention has the additional benefit of reducing the risk of introducing undesirable instability into an individual's Type A belief landscapes.

The larger goal of psychotherapy enterprises is not to decrease canalization per se, but rather to enable the development of healthy and adaptive behavioral, cognitive, and emotional circuits that are plastic enough to meaningfully change with an individual as they grow and move through their life, but stable enough to provide a meaningful ground on which the individual can rely (Jedlicka et al., 2022). From this perspective, psychedelic therapy is just one of many possible tools that a therapist might employ in order to aid an individual toward the development and nurturing of such stable and adaptive circuits. We can imagine that it is likely desirable to employ psychedelic therapy in cases of severe depression or addiction, when other gentler methods of belief sculpting are unable to make enough of an impact, due to excessive canalization in either the Type A or Type B optimization landscapes. In contrast, it may be undesirable to use psychedelic therapy in cases of psychopathologies of insufficient canalization of Type B landscapes or even unnecessary in some psychologically healthy adults. Contextualizing the usefulness of psychedelic therapy allows the integration of canalization theory into a much larger literature on the process of the psychotherapeutic enterprise that has been accumulating for more than a century (Kazdin, 2007).

Broader clinical implications

Using our revised conceptual framework, we can reassess the potential suitability of psychedelic therapy based on the types of canalization an individual may experience along each of the two dimensions. It is important to reiterate that these two types of optimization landscape may be canalized to different extents within different functional networks of the brain. Given the heterogeneity in the presentation of symptoms associated with psychopathologies of various kinds, the classification attempted here



		Type A (Inference)	
		Overfitting	**Underfitting**
Type B (Learning)	**Plasticity Loss**	• Major depressive disorder • Obsessive compulsive disorder • Substance use disorder • Generalized anxiety disorder	• Attention deficit hyperactivity disorder • Depersonalization-derealization disorder • Autism spectrum disorder
	Catastrophic Forgetting	• Bipolar disorder • Borderline personality disorder • Schizophrenia	• Alzheimer's disease

Treatment with psychedelic therapy: 🟦 Clear benefit ◻ Potential benefit 🟪 Mixed benefit ⬜ Limited benefit

Figure 5. Hypothesized representative psychopathologies associated with either extreme of canalization along each of the two optimization landscapes underlying belief representation. Cell shading corresponds to potential efficacy of psychedelic therapy to address associated pathologies based on current empirical research and predictions of our model.

is largely provisional and a proof of concept. The examples presented later are intended to demonstrate the potential applicability of deep learning concepts to understanding psychopathology, not as a proposed tool for diagnoses. With this in mind, we examine each of the four combinations of over-canalization or under-canalization in the two optimization landscape types and their relationship to psychopathology. See Figure 5 for an overview of representative psychopathologies and the potential efficacy of psychedelic therapy to treat them.

In cases where there is canalization in both types of optimization landscape (upper left quadrant), psychedelic therapy is likely to produce the most beneficial clinical outcomes. This is because increasing plasticity in Type B landscapes can help to enable an increase in stability by making it more possible to reduce overfitting. Likewise, in cases where the Type A landscape is already heavily overfit, destabilization of that landscape will be more likely to lead away from canalization rather than toward it, especially in a supportive therapeutic context. We believe that this action on both types of belief landscape is the mechanism by which consistently compelling results have been demonstrated for the treatment of major depressive disorder, substance use disorder, body image disorders, and others (Johnson and Griffiths, 2017). It is also for individuals in this context that the REBUS model applies most directly (Carhart-Harris and Friston, 2019), as the likely result of psychedelic therapy (provided in a stable and supportive environment) would be a reduction in both types of canalization and a corresponding reduction in symptoms.

We can next consider the quadrant in which individuals may be under-canalized in one or more Type A landscapes but over-canalized in Type B landscapes (upper right). Psychopathologies that are potentially consistent with this configuration include attention deficit hyperactivity disorder (Hauser et al., 2016), autism spectrum disorder (Pellicano and Burr, 2012; Rogers et al., 2022) (although see Van de Cruys et al. (2013)), and depersonalization-derealization disorder (Seth et al., 2012; Ciaunica and Safron, 2022). All three of these disorders can be characterized by an inconsistent deployment of mental circuits (attention in the case of Attention deficit hyperactivity disorder, interpersonal behavioral strategies in the case of autism, and body schema in the case of depersonalization), as well as an inability or difficulty in learning or changing these circuits over time. In these cases, there is a potential benefit from psychedelic therapy, although it is less straightforward than what is expected from the first group described earlier. For these disorders, there have been some preliminary explorations into the efficacy of psychedelic therapy (Hutten et al., 2019; Markopoulos et al., 2022). The potential mechanism of this benefit would come from a reduction in Type B canalization, affording individuals a greater capacity to learn more adaptive mental circuits over time. The malleability introduced in Type A landscapes may also contribute to greater flexibility in rehearsing and enacting more adaptive mental circuits. Of particular interest may be the micro-dosing regime, where beliefs represented in the Type A landscape may be strengthened slightly, thus producing a decrease in underfitting. At the same time, micro-dosing has the potential to have a nontrivial effect on neurotrophic factors, thus also contributing to increases in plasticity in Type B landscapes.

The third quadrant (bottom right) corresponds to a pathological lack of canalization in both Type A and Type B belief landscapes. This manifests itself as both underfitting and catastrophic forgetting. In the clinical literature, we find that Alzheimer's disease matches this profile most closely (Parasuraman and Haxby, 1993). Individuals suffering from this disease show a breakdown in the deployment of adaptive mental circuits, as well as a permanent loss of these mental circuits over time as the disease progresses. Despite our classification of Alzheimer's in the diagonal quadrant as diseases such as major depressive disorder, there has been significant interest in the use of psychedelics to help treat the symptoms of this disease (Vann Jones and O'Kelly, 2020; Garcia-Romeu et al., 2021). In particular, interest has been paid to the value of micro-dosing as a means of increasing mental acuity in Alzheimer's patients. In our framework, this would correspond to using these drugs to transiently strengthen belief representation in Type A landscapes and thus reduce underfitting. In contrast, the neurotrophic effects of psychedelics on Type B landscapes may have some positive impact in reducing symptom severity, but are unlikely to significantly counter the severe neurodegenerative effects of the disease, especially in later stages.

The fourth quadrant (bottom left) describes individuals who may be over-canalized in one or more Type A landscapes but under-canalized with respect to Type B landscapes. Examples of this configuration include bipolar disorder, borderline personality disorder, and schizophrenia. In each case, there is the presence of highly overfit mental circuits that are deployed in often severely maladaptive contexts. There is also an inability to retain or reuse adaptive circuits that had been previously developed. In the case of bipolar disorder, this manifests itself as a cycling between two states of significant overfitting (mania and depression). In borderline personality disorder, this manifests itself as a drift between multiple different low-entropy maladaptive policies. In schizophrenia, this presents as the phasic entry into psychosis between periods of remission. In these cases, there is at best a mixed benefit from psychedelic therapy. Although psychedelic use can enable a reduction in overfitting through an increase in the malleability of Type A landscapes, it has the potential to further contribute to the under-canalization of Type B landscapes, which would fail to address some of the underlying mechanisms of these psychopathologies. For example, in cases of bipolar disorder, there is evidence for the possible exacerbation of mania and a suggestion to proceed with caution when using psychedelic therapy to treat this population (Gard et al., 2021). In borderline personality disorder, there is early work to examine the risks and benefits of psychedelic therapy (Zeifman and Wagner, 2020) for this population of individuals.

We can also consider a fifth case: that of individuals who are neither over- nor under-canalized along either dimension. If an individual is already high on the meta-trait of stability, then an increase in plasticity induced by psychedelics may lead to an undesirable downstream decrease in stability through the introduction of a potentially pathological underfitting or overfitting. However, given the correct supporting environment and proper preparation, increasing plasticity may enable greater adaptability going forward or the ability for slightly overfit individuals to arrive at a more optimal set of mental circuits. As noted in the original CANAL model, this profile fits a large number of relatively functional adults seeking mental health assistance in industrialized nations (Carhart-Harris et al., 2022). This benefit for healthy individuals who receive psychedelics in a properly supportive clinical setting has growing empirical validation (Kraehenmann et al., 2015; Gandy, 2019).

Summary and conclusion

Psychedelic therapy has tremendous potential to improve the lives of a great number of people who suffer from currently poorly treatable psychopathologies. To realize its potential, the mechanisms by which psychedelic therapy acts to change the brain must be well understood at multiple levels of analysis. This requires as a first step the deployment of useful conceptual frameworks by which to reason about the class of drugs' action. Both the REBUS model of psychedelics and the CANAL model of psychopathology have been proposed as such frameworks. Using additional theoretical insights from the field of deep learning, we have demonstrated a few refinements to these models. We believe that these refinements have the potential to increase the theory's robustness with respect to clinical observations in addition to its usefulness in the generation of experimental hypothesis.

Our primary contribution is a refinement of the CANAL model which delineates two distinct forms of canalization that a dynamical learning system may develop based on the topology of distinct optimization landscapes underlying belief representation. These are Type A (inference) and Type B (learning). Type A belief landscapes may suffer from pathologies of over- or underfitting. Type B belief landscapes may suffer from either catastrophic forgetting or plasticity loss. The distinction between these two landscapes also maps onto a larger literature on meta-traits in personality theory, with Type A corresponding to "stability" and Type B corresponding to "plasticity." Furthermore, the constructs of cognitive and psychological flexibility each describe the absence of pathology in the Type A and Type B landscapes, respectively.

This expanded framework enables the identification of a class of psychopathologies that can be understood to arise from reduced stability in Type A belief landscapes or too much plasticity in Type B belief landscapes, complicating the mapping between canalization and the *p*-factor of psychopathology. In this work, we examine the therapeutic role that destabilization of beliefs may play in psychedelic therapy, as well as the potential value of the micro-dosing regimen for healthy individuals. Importantly, psychedelics act on these two optimization landscapes in different ways, with different clinical implications. We end the work by laying out a preliminary means of predicting the efficacy of psychedelic therapy based on the topological features of both Type A and Type B belief landscapes.

We recognize that many of the ideas presented in this paper will require further theoretical elaboration and experimental validation. Despite the presence of some speculative hypotheses, we believe that "Deep CANALs" has the potential to serve as a foundation for a rigorous theoretical analysis of the pathologies that arise in learning systems. Methodological maturity in the field of deep learning provides for a wide array of techniques for empirical investigation of overfitting, plasticity loss, and their complex non-linear relationship. Likewise, as has been the case in other subfields of the brain sciences (Saxe et al., 2021), we have reason to believe that a deep learning perspective can help inform the development of experiments at both the level of basic research and the clinical level. Rather than acting as an alternative to the paradigm of hierarchical predictive coding, this perspective has the opportunity to enable a broader dialogue between fields, ultimately leading to a mutual benefit by which each is enhanced. Most importantly, such a broadened perspective may one day lead to the development of more sophisticated clinical protocols or even novel psychedelic drugs developed in a more patient-centered way. Such an outcome would serve the goals of a mature precision or computational psychiatry (Montague et al., 2012; Fernandes et al., 2017) and ultimately contribute to more positive mental health outcomes for those most in need.

Acknowledgements

The authors thank Colin DeYoung, Laura Graesser, Nate Hess, and Curt Tigges for their valuable feedback on an earlier draft of this paper. We furthermore thank Yad Konrad, Michael Johnson, Karl Friston, Matt Johnson, and Zahra Sheikhbahaee for their insightful comments when discussing initial forms of some of the ideas eventually presented here.

Data availability

No original data were used in the preparation of this manuscript.

References

Abbas Z, Zhao R, Modayil J *et al.* Loss of plasticity in continual deep reinforcement learning, preprint, arXiv:2303.07507, 2023.

Achille A, Rovere M, Soatto S. Critical learning periods in deep neural networks. In: *International Conference on Learning Representations* New Orleans, Louisiana, 2019.

Ash J, Adams RP. On warm-starting neural network training, *Advances in Neural Information Processing systems* 2020;**33**:3884–94.

Ballentine G, Friedman SF, Bzdok D. Trips and neurotransmitters: discovering principled patterns across 6850 hallucinogenic experiences, *Science Advances* 2022;**8**:eabl6989.

Barak O. Recurrent neural networks as versatile tools of neuroscience research, *Current Opinion in Neurobiology* 2017;**46**:1–6.

Beliveau V, Ganz M, Feng L *et al.* A high-resolution in vivo atlas of the human brain's serotonin system, *Journal of Neuroscience* 2017;**37**:120–8.

Bennett D, Davidson G, Niv Y. A model of mood as integrated advantage, *Psychological Review* 2022;**129**:513–41.

Bogenschutz MP, Forcehimes AA, Pommy JA *et al.* Psilocybin-assisted treatment for alcohol dependence: a proof-of-concept study, *Journal of psychopharmacology* 2015;**29**:289–99.

Bossy-Wetzel E, Schwarzenbacher R, Lipton SA. Molecular pathways to neurodegeneration, *Nature medicine* 2004;**10**:S2–S9.

Bowers JS, Davis CJ. Bayesian just-so stories in psychology and neuroscience, *Psychological bulletin* 2012;**138**:389–414.

Bremler R, Katati N, Shergill P *et al.* Focusing on the negative: cases of long-term negative psychological responses to psychedelics, 2023.

Calder AE, Hasler G. Towards an understanding of psychedelic-induced neuroplasticity, *Neuropsychopharmacology* 2023;**48**:104–12.

Cameron LP, Tombari RJ, Lu J *et al.* A non-hallucinogenic psychedelic analogue with therapeutic potential, *Nature* 2021;**589**:474-479.

Carhart-Harris R, Chandaria S, Erritzoe D *et al.* Canalization and plasticity in psychopathology, *Neuropharmacology* 2022;**226**:109398.

Carhart-Harris R, Friston K. Rebus and the anarchic brain: toward a unified model of the brain action of psychedelics, *Pharmacological reviews* 2019;**71**:316–44.

Carhart-Harris RL, Leech R, Hellyer PJ *et al.* The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs, *Frontiers in Human neuroscience* 2014;**8**:1662–5161.

Caspi A, Houts RM, Belsky DW *et al.* The p factor: one general psychopathology factor in the structure of psychiatric disorders?, *Clinical psychological science* 2014;**2**:119–137.

Ciaunica A, Safron A. Disintegrating and reintegrating the self–(in) flexible self-models in depersonalisation and psychedelic experiences, 2022. preprint, psyarxiv:mah78

Clark LA. Temperament as a unifying basis for personality and psychopathology, *Journal of Abnormal psychology* 2005;**114**:505–21.

Cunningham MJ, Bock HA, Serrano IC *et al.* Pharmacological mechanism of the non-hallucinogenic 5-ht2a agonist ariadne and analogs, *ACS Chemical Neuroscience* 2023;**14**:119–35.

Davis AK, Barrett FS, Griffiths RR. Psychological flexibility mediates the relations between acute psychedelic effects and subjective decreases in depression and anxiety, *Journal of Contextual Behavioral science* 2020;**15**:39–45.

DeYoung CG. Cybernetic big five theory, *Journal of Research in personality* 2015;**56**:33–58.

DeYoung CG, Krueger RF. A cybernetic theory of psychopathology, *Psychological Inquiry* 2018;**29**:117–38.

Dohare S, Sutton RS, Mahmood AR. Continual backprop: Stochastic gradient descent with persistent randomness. In: *International Conference on Learning Representations*, 2022.

Doss MK, Madden MB, Gaddis A *et al.* Models of psychedelic drug action: modulation of cortical-subcortical circuits, *Brain* 2022;**145**:441–56.

Doss MK, Považan M, Rosenberg MD *et al.* Psilocybin therapy increases cognitive and neural flexibility in patients with major depressive disorder, *Translational psychiatry* 2021;**11**:574.

Draganov M, Galiano-Landeira J, Doruk Camsari D *et al.* *Noninvasive modulation of predictive coding in humans: causal evidence for frequency-specific temporal dynamics*, *Cerebral Cortex*, 2023, bhad127.

Erritzoe D, Roseman L, Nour M *et al.* Effects of psilocybin therapy on personality structure, *Acta Psychiatrica Scandinavica* 2018;**138**:368–78.

Erritzoe D, Smith J, Fisher PM *et al.* Recreational use of psychedelics is associated with elevated personality trait openness: Exploration of associations with brain serotonin markers, *Journal of Psychopharmacology* 2019;**33**:1068–75.

Fernandes BS, Williams LM, Steiner J *et al.* The new field of 'precision psychiatry', *BMC medicine* 2017;**15**:1–7.

French RM. Catastrophic forgetting in connectionist networks, *Trends in Cognitive sciences* 1999;**3**:128–35.

Fridhandler BM. Conceptual note on state, trait, and the state–trait distinction, *Journal of Personality and Social Psychology* 1986;**50**:169–74.

Friston K. The free-energy principle: a unified brain theory?, *Nature Reviews neuroscience* 2010;**11**:127–38.

Friston K, Kiebel S. Predictive coding under the free-energy principle, *Philosophical Transactions of the Royal Society B: Biological sciences* 2009;**364**:1211–21.

Gandy S. Psychedelics and potential benefits in "healthy normals": a review of the literature, *Journal of Psychedelic Studies* 2019;**3**:280–7.

Garcia-Romeu A, Darcy S, Jackson H *et al.* Psychedelics as novel therapeutics in Alzheimer's disease: rationale and potential mechanisms, *Disruptive Psychopharmacology* 2021;**56**:287–317.

Gard DE, Pleet MM, Bradley ER *et al.* Evaluating the risk of psilocybin for the treatment of bipolar depression: a review of the research literature and published case studies, *Journal of Affective Disorders Reports* 2021;**6**:100240.

Girn M, Rosas FE, Daws RE *et al.* *A complex systems perspective on psychedelic brain action*, Trends in Cognitive Sciences, 2023.

Goldberg SB, Pace BT, Nicholas CR *et al.* The experimental effects of psilocybin on symptoms of anxiety and depression: a meta-analysis, *Psychiatry research* 2020;**284**:112749.

Gómez-Emilsson A (2021). *Healing trauma with neural annealing* https://qri.org/blog/neural-annealing (15 May 2023, date last accessed).

Griffiths RR, Hurwitz ES, Davis AK *et al.* Survey of subjective god encounter experiences: comparisons among naturally occurring experiences and those occasioned by the classic psychedelics psilocybin, LSD, ayahuasca, or DMT, *PLoS One* 2019;**14**: e0214377.

Griffiths RR, Johnson MW, Carducci MA *et al.* Psilocybin produces substantial and sustained decreases in depression and anxiety in patients with life-threatening cancer: a randomized double-blind trial, *Journal of psychopharmacology* 2016;**30**:1181–97.

Haarnoja T, Zhou A, Abbeel P *et al.* Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: *International conference on machine learning* Stockholm Sweden, 2018, pp. 1861–70.

Hartogsohn I. Set and setting, psychedelics and the placebo response: an extra-pharmacological perspective on psychopharmacology, *Journal of Psychopharmacology* 2016;**30**:1259–67.

Hauser TU, Fiore VG, Moutoussis M *et al.* Computational psychiatry of ADHD: neural gain impairments across marrian levels of analysis, *Trends in neurosciences* 2016;**39**:63–73.

Hellyer PJ, Scott G, Shanahan M *et al.* Cognitive flexibility through metastable neural dynamics is disrupted by damage to the structural connectome, *Journal of Neuroscience* 2015;**35**:9050–63.

Hipólito I, Mago J, Rosas FE *et al.* Pattern breaking: a complex systems approach to psychedelic medicine, *Neuroscience of Consciousness* 2023;**2023**:niad017.

Hutten NR, Mason NL, Dolder PC *et al.* Self-rated effectiveness of microdosing with psychedelics for mental and physical health problems among microdosers, *Frontiers in psychiatry* 2019;**10**:672.

Huys QJ, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications, *Nature neuroscience* 2016;**19**:404–13.

Ionescu T. Exploring the nature of cognitive flexibility, *New Ideas in psychology* 2012;**30**:190–200.

Jedlicka P, Tomko M, Robins A *et al.* Contributions by metaplasticity to solving the catastrophic forgetting problem, *Trends in Neurosciences* 2022;**45**:656–66.

Johnson M (2019). Neural annealing: toward a neural theory of everything. https://opentheory.net/2019/11/neural-annealing-toward-a-neural-theory-of-everything (15 May 2023, date last accessed).

Johnson MW, Griffiths RR. Potential therapeutic effects of psilocybin, *Neurotherapeutics* 2017;**14**:734–40.

Johnson MW, Richards WA, Griffiths RR. Human hallucinogen research: guidelines for safety, *Journal of psychopharmacology* 2008;**22**:603–20.

Kanai R, Komura Y, Shipp S *et al.* Cerebral hierarchies: predictive processing, precision and the pulvinar, *Philosophical Transactions of the Royal Society B: Biological Sciences* 2015;**370**:20140169.

Kashdan TB, Rottenberg J. Psychological flexibility as a fundamental aspect of health, *Clinical Psychology review* 2010;**30**:865–78.

Kazdin AE. Mediators and mechanisms of change in psychotherapy research, *Annual Review of Clinical Psychology* 2007;**3**:1–27.

Keller GB, Mrsic-Flogel TD. Predictive processing: a canonical cortical computation, *Neuron* 2018;**100**:424–35.

Kessler RC, McLaughlin KA, Green JG *et al.* Childhood adversities and adult psychopathology in the who world mental health surveys, *The British Journal of psychiatry* 2010;**197**:378–385.

Kirchhoff M, Parr T, Palacios E *et al.* The markov blankets of life: autonomy, active inference and the free energy principle, *Journal of The Royal Society interface* 2018;**15**:20170792.

Kraehenmann R, Preller KH, Scheidegger M *et al.* Psilocybin-induced decrease in amygdala reactivity correlates with enhanced positive mood in healthy volunteers, *Biological psychiatry* 2015;**78**:572–81.

Krebs TS, Johansen P.- Ø.. Psychedelics and mental health: a population study, *PLoS One* 2013;**8**:e63972.

Laukkonen RE, Slagter HA. From many to (n) one: Meditation and the plasticity of the predictive mind, *Neuroscience and Biobehavioral Reviews* 2021;**128**:199–217.

Lebedev AV, Kaelen M, Lövdén M *et al.* Lsd-induced entropic brain activity predicts subsequent personality change, *Human Brain mapping* 2016;**37**:3203–13.

LeCun Y, Bengio Y, Hinton G. Deep learning, *nature* 2015;**521**:436–44.

Lee WH, Doucet GE, Leibu E *et al.* Resting-state network connectivity and metastability predict clinical symptoms in schizophrenia, *Schizophrenia research* 2018;**201**:208–16.

Letheby C. *Philosophy of psychedelics*. Oxford, United Kingdom: Oxford University Press, 2021.

Letheby C, Gerrans P. Self unbound: ego dissolution in psychedelic experience, *Neuroscience of Consciousness* 2017;**2017**:nix016.

Levine S. Reinforcement learning and control as probabilistic inference: Tutorial and review, preprint, arXiv:1805.00909, 2018.

Lillicrap TP, Santoro A, Marris L *et al.* Backpropagation and the brain, *Nature Reviews Neuroscience* 2020;**21**:335–46.

Li H, Xu Z, Taylor G *et al.* Visualizing the loss landscape of neural nets. In: *Advances in Neural Information Processing Systems* Monreal, Canada, Vol. 31, 2018.

Lozano V, Soriano MF, Aznarte JI *et al.* Interference control commonalities in patients with schizophrenia, bipolar disorder, and borderline personality disorder, *Journal of Clinical and Experimental neuropsychology* 2016;**38**:238–50.

Lutkajtis A. Entity encounters and the therapeutic effect of the psychedelic mystical experience, *Journal of Psychedelic Studies* 2021;**4**:171–8.

Ly C, Greb AC, Cameron LP *et al.* Psychedelics promote structural and functional neural plasticity, *Cell reports* 2018;**23**:3170–82.

Lyle C, Zheng Z, Nikishin E *et al. Understanding plasticity in neural networks. International Conference on Machine Learning* Honolulu Hawaii, 2023.

MacKinnon DF, Pies R. Affective instability as rapid cycling: theoretical and clinical implications for borderline personality and bipolar spectrum disorders, *Bipolar disorders* 2006;**8**:1–14.

Mante V, Sussillo D, Shenoy KV *et al.* Context-dependent computation by recurrent dynamics in prefrontal cortex, *nature* 2013;**503**:78–84.

Marblestone AH, Wayne G, Kording KP. Toward an integration of deep learning and neuroscience, *Frontiers in Computational neuroscience* 2016;**10**:94.

Markopoulos A, Inserra A, De Gregorio D *et al.* Evaluating the potential use of serotonergic psychedelics in autism spectrum disorder, *Frontiers in Pharmacology* 2022;3341.

Martinotti G, Santacroce R, Pettorruso M *et al.* Hallucinogen persisting perception disorder: etiology, clinical features, and therapeutic perspectives, *Brain Sciences* 2018;**8**:47.

McCloskey M, Cohen NJ. Catastrophic interference in connectionist networks: the sequential learning problem. In: *Psychology of learning and motivation*. Elsevier, 1989, 24, 109–165.

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function, *Annual Review of neuroscience* 2001;**24**:167–202.

Millière R, Carhart-Harris RL, Roseman L *et al.* Psychedelics, meditation, and self-consciousness, *Frontiers in psychology* 2018;**9**:1475.

Montague PR, Dolan RJ, Friston KJ *et al.* Computational psychiatry, *Trends in Cognitive sciences* 2012;**16**:72–80.

Nardou R, Sawyer E, Song YJ *et al.* Psychedelics reopen the social reward learning critical period. *Nature* 2023;**618**:790–8.

Neff KD. Self-compassion: Theory, method, research, and intervention, *Annual Review of Psychology* 2023;**74**:193–218.

Nikishin E, Schwarzer M, D'Oro P *et al.* The primacy bias in deep reinforcement learning. In: *International conference on machine learning* Baltimore, Maryland, 2022, pp. 16828–47.

Nolen-Hoeksema S, Wisco BE, Lyubomirsky S. Rethinking rumination, *Perspectives on Psychological science* 2008;**3**:400–24.

Olaru G, van Scheppingen MA, Bleidorn W *et al.* The link between personality, global, and domain-specific satisfaction across the adult lifespan, *Journal of Personality and Social Psychology.* 2023;**125**:590–606.

Parasuraman R, Haxby JV. Attention and brain function in alzheimer's disease: A review, *Neuropsychology* 1993;**7**:242–72.

Parisi G.I, Kemker R, Part JL *et al.* Continual lifelong learning with neural networks: a review, *Neural networks* 2019;**113**:54–71.

Pellicano E, Burr D. When the world becomes 'too real': a bayesian explanation of autistic perception, *Trends in Cognitive sciences* 2012;**16**:504–10.

Qiao H, Li M.- X, Xu C *et al.* Dendritic spines in depression: what we learned from animal models, *Neural plasticity* 2016;**2016**:1–26.

Rabinovich M.I., Varona P, Selverston A.I *et al.* Dynamical principles in neuroscience, *Reviews of Modern physics* 2006;**78**:1213.

Richards BA, Kording KP. The study of plasticity has always been about gradients, *The Journal of Physiology.* 2023;**601**:3141–9.

Richards BA, Lillicrap TP, Beaudoin P *et al.* A deep learning framework for neuroscience, *Nature neuroscience* 2019;**22**:1761–70.

Rogers MA, Elison JT, Blain SD *et al.* A cybernetic theory of autism: autism as a consequence of low trait plasticity. *Journal of Personality* 2022;**91**:1035–50.

Safron A. On the varieties of conscious experiences: altered beliefs under psychedelics (ALBUS), 2020.

Safron A, DeYoung CG. Integrating cybernetic big five theory with the free energy principle: a new strategy for modeling personalities as complex systems. In: *Measuring and modeling persons and situations*. Elsevier, 2021, 617–649.

Safron A, Johnson M. Classic psychedelics: past uses, present trends, future possibilities, 2022.

Saxe A, Nelli S, Summerfield C. If deep learning is the answer, what is the question?, *Nature Reviews Neuroscience* 2021;**22**:55–67.

Schmack K, Schnack A, Priller J *et al.* Perceptual instability in schizophrenia: probing predictive coding accounts of delusions with ambiguous stimuli, *Schizophrenia Research: Cognition* 2015;**2**:72–7.

Schulman J, Wolski F, Dhariwal P *et al.* Proximal policy optimization algorithms, 2017, preprint, arXiv:1707.06347.

Sessa B. The 21st century psychedelic renaissance: heroic steps forward on the back of an elephant, *Psychopharmacology* 2018;**235**:551–60.

Seth AK. Interoceptive inference, emotion, and the embodied self, *Trends in Cognitive sciences* 2013;**17**:565–73.

Seth AK, Suzuki K, Critchley HD. An interoceptive predictive coding model of conscious presence, *Frontiers in psychology* 2012;**2**:395.

Stanley KO, Lehman J. *Why Greatness Cannot be planned: The Myth of the objective*. New York NY: Springer, 2015.

Tackett JL. Evaluating models of the personality–psychopathology relationship in children and adolescents, *Clinical Psychology Review* 2006;**26**:584–99.

Tagliazucchi E, Roseman L, Kaelen M *et al.* Increased global functional connectivity correlates with lsd-induced ego dissolution, *Current biology* 2016;**26**:1043–50.

Tognoli E, Kelso JS. The metastable brain, *Neuron* 2014;**81**:35–48.

Van de Cruys S, de Wit L, Evers K *et al.* Weak priors versus overfitting of predictions in autism: reply to Pellicano and Burr (TICS, 2012), *i-Perception* 2013;**4**:95–7.

Vann Jones SA, O'Kelly A. Psychedelics as a treatment for alzheimer's disease dementia, *Frontiers in Synaptic neuroscience* 2020;**12**:34.

Vollenweider FX, Geyer MA. A systems model of altered consciousness: integrating natural and drug-induced psychoses, *Brain Research bulletin* 2001;**56**:495–507.

Vollenweider FX, Vollenweider-Scherpenhuyzen MF, Bäbler A *et al.* Psilocybin induces schizophrenia-like psychosis in humans via a serotonin-2 agonist action, *Neuroreport* 1998;**9**:3897–902.

Vyas S, Rodrigues AJ, Silva JM *et al.* Chronic stress and glucocorticoids: from neuronal plasticity to neurodegeneration, *Neural plasticity* 2016;**2016**:1–15.

Walsh KS, McGovern DP, Clark A *et al.* Evaluating the neurophysiological evidence for predictive processing as a model of perception, *Annals of the New York Academy of Sciences* 2020;**1464**: 242–68.

Watkins ER. Constructive and unconstructive repetitive thought, *Psychological bulletin* 2008;**134**:163–206.

Wimsatt WC, Brewer M, Collins B. Robustness, reliability, and overdetermination, *Characterizing the Robustness of Science. Boston Studies in the Philosophy of Science* 1981;**292**:61–87.

Winterer G, Musso F, Beckmann C *et al.* Instability of prefrontal signal processing in schizophrenia, *American Journal of Psychiatry* 2006;**163**:1960–8.

Yaden DB, Earp D, Graziosi M *et al.* Psychedelics and psychotherapy: cognitive-behavioral approaches as default, *Frontiers in psychology* 2022;**13**:1604.

Yaden DB, Griffiths RR. The subjective effects of psychedelics are necessary for their enduring therapeutic effects, *ACS Pharmacology and Translational Science* 2020;**4**:568–72.

Ying X. An overview of overfitting and its solutions. *Journal of Physics: Conference Series* 2019;**1168**:022022.

Zeifman RJ, Spriggs MJ, Kettner H *et al. From relaxed beliefs under psychedelics (rebus) to revised beliefs after psychedelics (rebas): preliminary development of the relaxed beliefs questionnaire (reb-q)*, 2022.

Zeifman RJ, Wagner AC. Exploring the case for research on incorporating psychedelics within interventions for borderline personality disorder, *Journal of Contextual Behavioral Science* 2020;**15**: 1–11.

Zeifman RJ, Wagner AC, Monson CM *et al.* How does psilocybin therapy work? an exploration of experiential avoidance as a putative mechanism of change, *Journal of Affective Disorders* 2023;**334**:100–12.

Ziebart BD, Maas AL, Bagnell JA *et al.* Maximum entropy inverse reinforcement learning, In: *Aaai* Chicago, Illinois, 2008, Vol. 8, pp. 1433–8.

Neuroscience of Consciousness, 2024, **2024(1)**, 1–15
DOI: https://doi.org/10.1093/nc/niae005
Research article
Received 17 May 2023; Revised 15 January 2024; accepted 19 January 2024