Deep CANALs:

A Deep Learning Approach to Refining the Canalization Theory of Psychopathology

Arthur Juliani

Microsoft Research, New York, NY, USA

ajuliani@microsoft.com

Adam Safron

Johns Hopkins University, Baltimore, MD, USA

Ryota Kanai

Araya Inc, Minato-ku, Tokyo, Japan

Abstract

Psychedelic therapy has seen a resurgence of interest in the past decade, with promising clinical outcomes for treatment of a variety of psychopathologies. In response to this success, a number of theoretical models have been proposed to account for psychedelic's positive therapeutic effects. One of the more prominent models is 'RElaxed Beliefs Under pSychedelics,' (REBUS) which suggests 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 REBUS by proposing that canalization (the development of overly rigid belief landscapes) may be a primary factor in psychopathology. Here we take inspiration from 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, 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, suggesting that at the very least canalization does not have a linear relationship with psychopathology. In this expanded paradigm, we demonstrate the ability to make novel predictions regarding which psychopathologies may be amenable to psychedelic therapy, as well as what forms of psychedelic therapy may ultimately be most beneficial for a given individual.

1 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; M. W. Johnson & Griffiths, 2017; Goldberg, Pace, Nicholas, Raison, & Hutson, 2020), it is natural to consider that there may be an underlying common cause behind these pathologies which psychedelic therapy is helping to address. Building on the RElaxed Beliefs Under pSychedelics (REBUS) model of psychedelic action (R. L. Carhart-Harris & Friston, 2019), Carhart-Harris and colleagues have proposed that what they refer to as 'excessive canalization' may constitute the primary factor underlying all psychopathology (R. 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 & Mrsic-Flogel, 2018). These overly precise beliefs act as sticky attractors in a dynamical systems sense, becoming engaged under a variety of different contexts, regardless of their appropriateness to the situation. Furthermore, they are difficult to unlearn due to the high precision assigned to them by the underlying generative models which support them.

The CANAL model draws upon evidence from psychedelic research, combined with various broader clinical observations to support the proposal that most psychopathologies can be understood from the perspective of canalization. The most clear examples of pathologies which fit the paradigm of canalization 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. We can understand these circuits to represent 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 outcome 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 change or replace the underlying mental circuits which 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 the individual finds themselves in. Such earlier times where these pathological mental circuits develop might be ones of early adversity marked by trauma, stress, or other kinds of hardship (Kessler et al., 2010). These mental circuits may also simply be no longer adaptive due to a significantly large shift in the context an individual inhabits, such as a move across the country, or the end of a long-term romantic relationship. Given the fact that the contexts in which we live change over time, sometimes dramatically, it follows that in many cases the circuits which may at some point have been adaptive might eventually cease to be so. 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 via attempted coping may be powerfully explanatory, and also beneficial in both reducing stigmatization and foregrounding our common humanity (Neff, 2023).

The frameworks of HPP, dynamical systems theory, and canalization all provide a useful set of tools for understanding psychopathology and mental health. An additional project is to critically examine these frameworks and bring them into dialogue with the broader paradigm of machine learning, and deep learning in particular (LeCun, Bengio, & Hinton, 2015), which has seen meaningful success in contributing to neuroscientific progress in recent years (Richards et al., 2019). In this article we take initial steps toward this goal by proposing a set of refinements to the CANAL model which take insights from deep learning theory. The field of deep learning is a particularly fruitful ground because it has been a domain within which problems of adaptive behavior have been studied for many decades within the context of continual or lifelong learning (Parisi, Kemker, Part, Kanan, & Wermter, 2019). Deep neural networks also require the study of complex nonlinear optimization dynamics, which likewise underpin learning in living organisms as well (Rabinovich, Varona, Selverston, & Abarbanel, 2006). Through elucidating this new perspective, we hope that a clearer understanding of the space of possible psychopathologies which could arise from malformations in the optimization landscapes of the brain will become possible. We furthermore hope that a better understanding of the dynamics of psychopathology will also help to make clearer an understanding of the contexts in which psychedelic therapy may or may not be particularly effective for a given individual, ultimately leading to better treatment outcomes in the future.

2 The predictive coding account of psychedelic action

Before proceeding to discussing refinements to the CANAL model, it is worth briefly reviewing the conceptual frameworks upon which it is based: hierarchical predictive processing (HPP) and the REBUS model of psychedelic action. Among the various computational models of the brain, HPP has seen significant success with respect to understanding and predicting a diverse array of empirical phenomena (Walsh, McGovern, Clark, & O'Connell, 2020; Draganov et al., 2023), both at low and high levels of abstraction (Friston & Kiebel, 2009; Kanai, Komura, Shipp, & Friston, 2015). According to HPP, the brain can be understood as a sequence of generative models making predictions about the world. Each of these models has the objective of predicting the activity of hierarchically lower models (i.e., closer to primary modalities), which serve as their input. The extent to which the predictions fail to account for the incoming information from lower levels produces an error signal. These errors in prediction are then passed to hierarchically higher levels (e.g. deep association, transmodal, and/or frontal cortices) as inputs. This simple objective of predicting the activity of other generative models within the brain is said to enable all the complex representational capacity which we humans seem to be endowed with. Within this framework the brain is acting to minimize variational free-energy (Friston, 2010; Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018), which can be thought of as equivalent to the entropy, surprise, or prediction errors of the system at each level of representation. Ultimately this surprise or potential entropy comes from the outside, either from

external sensory information in the world, or from interoceptive information gathered from the body of the organism (Seth, 2013). The hierarchical nature of these generative models is manifest in the fact that ones at lower levels of the hierarchy make predictions which are more concrete, and at smaller spatiotemporal scales. In contrast, further up the hierarchy the models make predictions which are more abstract and on larger spatiotemporal scales.

Given the HPP framework, we can then ask what happens to these finely-tuned belief landscapes under the effect of psychedelics. We can start to understand this action by considering the brain regions which have densely expressed 5-HT_{2A} receptors in their neuronal populations. One region of interest is the thalamus, which has been hypothesized to decrease in its ability to gate information's entry into the cortex under the effects of psychedelics (Vollenweider & Geyer, 2001). Under normal functioning, this gating would serve to reduce the amount (or precision) of incoming information as a means of focusing attentional resources on the most important details. When this function is disrupted, higher precision sensory evidence enters deeper/higher portions of the system, producing larger (relative to unaltered conditions) prediction errors at different levels of the generative model. 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.

This effect 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 & Cohen, 2001). By intensely exciting neurons in the PFC, psychedelics can desynchronize their collective activity, effectively disrupting their ability to contribute to the representation of coherent beliefs about the self or other high-level abstract beliefs which the PFC participates in supporting (Millière, Carhart-Harris, Roseman, Trautwein, & Berkovich-Ohana, 2018). In terms of HPP, this disruption could 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 the beliefs at lower levels corresponds to an inability to suppress those beliefs, with incoming evidence from lower-level beliefs serving to more greatly impact the updating of 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. While the REBUS model does not specifically address the claustrum's role in these effects, other recent models have given it more of a central position (Doss et al., 2022). These complex effects are all measurable via an increase in complexity of neural activity, as well as in changes in functional connectivity which have been seen in fMRI studies of individuals under the effect of LSD, psilocybin, and DMT (Tagliazucchi et al., 2016).

According to the REBUS model, the result of this change in cortical dynamics is to relax the overall belief landscape which the cortex normally represents. This relaxation is specifically hypothesized to take

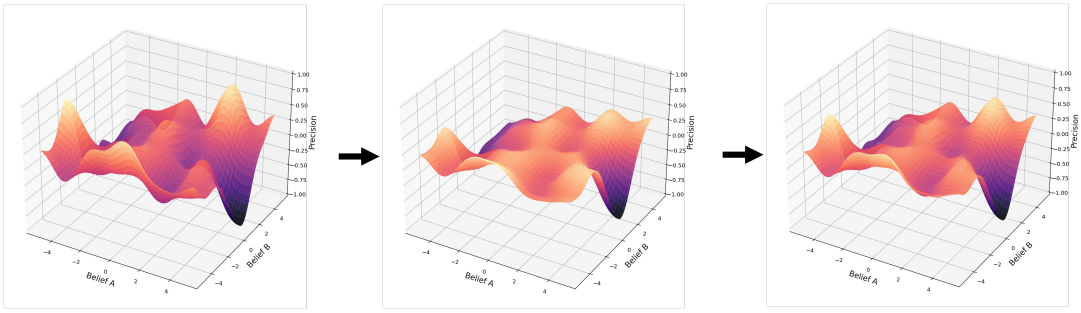

Figure 1: A diagram of the REBUS model of psychedelic action. The acute effect of psychedelics is to relax the belief landscape. This relaxation is understood to persist on some level during the post acute phase of the drug's effect as well.

place at the highest levels of the representational hierarchy in the regions of the brain responsible for self-referential beliefs. Within the HPP framework, this relaxation corresponds to a reduction in the precision of the high-level beliefs being encoded by the 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 from conflicting beliefs at other levels of the hierarchy. In this way, it is thought that psychedelics open up a window of time within which high-level beliefs can be more easily updated by evidence from the brain-external world, both from exteroceptive senses as well as the interoceptive world of the body. As the effects of the drug wear off, the belief landscape gradually returns to a state of higher precision beliefs once again, thus recalcifying the newly updated beliefs. A recently modified version of this model, along with empirical data provides some evidence for this relaxation of beliefs after psychedelic therapy (Zeifman et al., 2022). See Figure 1 for a representation of the purported effects of psychedelics on the belief landscapes of the brain.

3 Two canalizations for two optimization landscapes

Expanding the conceptual framework used to understand psychopathology and psychedelic action from HPP and active inference to the wider framework provided by deep 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 deep neural network literature (Lyle et al., 2023; Abbas, Zhao, Modayil, White, & Machado, 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. While often co-occurrent, plasticity loss is distinct from overfitting (Ying, 2019), which refers to the extent to which a given model is incapable of generalization or of appropriately responding to unfamiliar contexts. In empirical studies, plasticity loss has been shown to sometimes develop relatively early in the learning process of neural networks, leading to a characterization of "critical periods" or "primacy biases" in deep neural networks, both terms borrowed from the literature on human learning (Achille, Rovere, & Soatto, 2017; Nikishin, Schwarzer, D'Oro, Bacon,

& Courville, 2022). Plasticity loss can furthermore be understood as the inverse pathology to catastrophic forgetting (French, 1999; McCloskey & Cohen, 1989), which is the inability to retain critical information which was previously learned when learning new information.

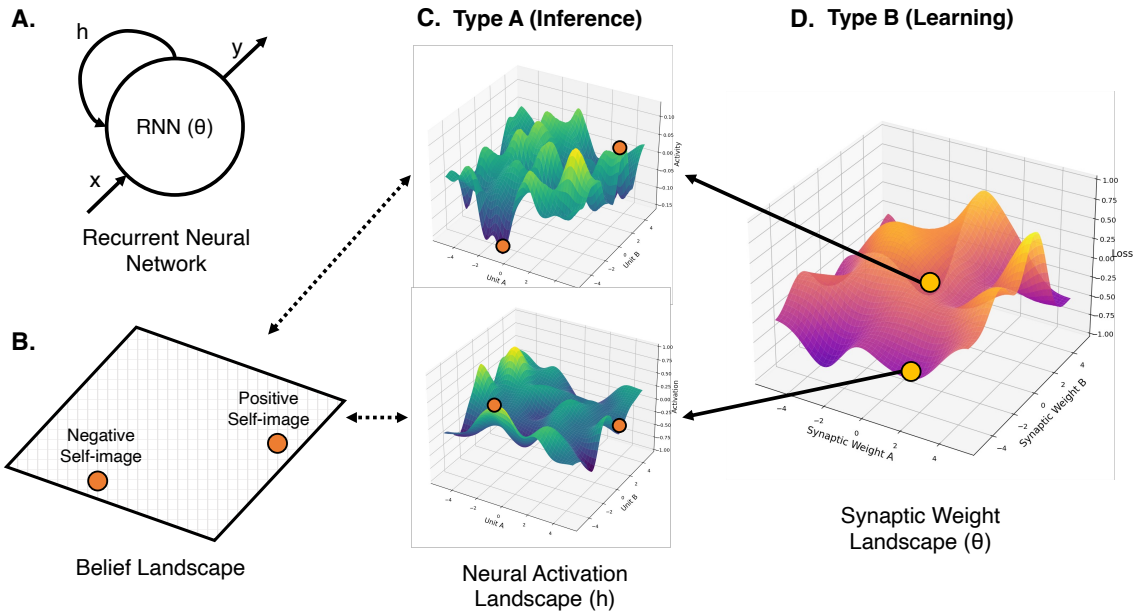

Figure 2: A. In an RNN performing free-energy minimization, there are two different underlying optimization landscapes, one corresponding to the activation pattern (h), and another corresponding to the synaptic weight values (θ). B. Different locations in a belief landscape will correspond to representing different beliefs, for example a positive or negative body image. C. The actualization of these beliefs takes place through the neural activity patterns, which are supported by the inference optimization landscape. D. The topology of the inference landscape is a function of the synaptic weight values which are defined by a learning optimization landscape.

In the original CANAL model, canalization is treated as a single construct and is described as a causal mechanism within the presentation of a host of different psychopathologies, all of which are said to arise from overly precise beliefs. We can use a deep learning perspective to disambiguate between two different kinds of canalization, which can correspond to either overfitting or plasticity loss. On one hand, it is possible to interpret canalization as described in the original CANAL model as being analogous to overfitting in machine learning. From this perspective, canalization is a measure of the rigidity of the belief landscape, which may contain sticky attractors that are maladaptive, resulting in stereotyped pathological mental circuits. On the other hand, it is possible to interpret canalization in the original CANAL model as being analogous to plasticity loss, in that it describes an inability of the underlying generative model to adapt or learn from new evidence over time.

These two types of canalization are distinct because they describe the topology of two related but unique optimization landscapes. We can illustrate this using a recurrent neural network (RNN) as a simple

model of the brain (or a network within the brain), as has often been done in computational neuroscience (Barak, 2017). Using this simple model, we can understand the first type of canalization as a property of the landscape of the activation patterns of the hidden state of an RNN. These activation patterns are the result of both incoming sensory information as well as the previous hidden state of the network, and the process of computing these patterns within a predictive coding framework can be understood as performing an inference procedure to minimize free energy. Through this process, operative (and potentially subjectively-experienced) beliefs are represented by the neural activity pattern at a given moment. 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 which define the recurrence function of the RNN itself. Because these weights are being updated intermittently based on the results of the inference process, we refer to this landscape as Type B (Learning). 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 as well as the state of the system within the synaptic weight landscape (Type B). Likewise, currently inferred beliefs are then a function of the location of the system within the neural activity landscape. See Figure 2 for a diagram of these two optimization landscapes, a belief landscape, and their interaction.

The two kinds of canalization have different clinical implications. Given an agent (artificial or natural) which is continually learning over time, being overfit to a previous context (Type A canalization) which is no longer realized will likely lead to both suboptimal behavior as well as associated stress (Bennett, Davidson, & Niv, 2022). If the agent is able to adapt to the new context quickly enough, then it is unlikely that we would classify the agent as pathological. If however the agent is both overfit and suffers from plasticity loss (Type B canalization), this would present a greater issue 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 not being capable of adapting to new contexts. It is in this latter case that we would describe the agent as being in a pathological situation which it may not be able to escape from without the aid of an external intervention of some sort. It follows that in such a case an intervention such as psychedelic therapy may indeed be most appropriate.

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 for adapting to a new environmental context, even if significant plasticity loss is not present, due to a reduced repertoire of realizable mental circuits. Inversely, 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 which would enable an otherwise less canalized individual to learn the necessary information. More often as happens in many cases of psychopathology there is both Type A and Type B canalization, resulting in maladaptive and stereotypical mental circuits which are unable to be corrected despite interventions which serve to provide necessary evidence to update them.

It is also worth noting that given the hierarchical nature prediction within the brain, it may be the case that different generative models (represented by distinct functional networks) at different levels of abstrac-

tion may be more or less canalized in different ways. As such, it is likely inappropriate to describe the whole brain as being either canalized or not in either the Type A or Type B sense. Furthermore, canalization may 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 sense in adults, and this is seen as a desirable rather than pathological property. Indeed, when this canalization is disrupted it can lead to undesirable changes to visual perception such as Hallucinogen persisting perception disorder (Martinotti et al., 2018). In contrast, canalization of either type at higher levels of the predictive hierarchy are thought to be related to pathologies of self-referential processing, and thus a key target for psychedelic therapy (Letheby & Gerrans, 2017).

4 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 large class of psychopathologies which may be understood to result not from too much canalization but rather from too little (DeYoung & Krueger, 2018). We can count various forms of schizophrenia, borderline, bipolar, and dissociative disorders among these. In each case, the individual suffers from a lack of coherent, stable, and adaptive behavioral, cognitive, and emotional circuits (MacKinnon & Pies, 2006; Winterer et al., 2006; Schmack, Schnack, Priller, & Sterzer, 2015; Lozano, Soriano, Aznarte, Gómez-Ariza, & Bajo, 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 framework described above, 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). On the other hand however, it may also be described from the perspective of the Type B landscape as corresponding to a lack of canalization, and thus the catastrophic forgetting of previously adaptive mental circuits which an individual is no longer capable of deploying, or in some cases of ever recovering.

This lack of stability results in real felt suffering in the individual, and a maladaptivity to the environment they find themselves in. For such individuals, psychedelic therapy, if it truly acted to unilaterally reduce canalization, could potentially lead to an exacerbation of their symptoms instead of relief. Indeed, the clinical literature contains a number of cases of individuals with predispositions to these pathologies of undercanalization who engage in psychedelic experiences only to have the result be the triggering of psychosis or a further worsening of their symptoms (Krebs & Johansen, 2013; Bremler, Katati, Shergill, Erritzoe, & Carhart-Harris, 2023). While the careful screening used in clinical trials of psychedelic therapies can greatly reduce the likelihood of these negative outcomes, as the availability of psychedelic drugs increases in coming years, it is critical that a theory of these drugs' action in the brain be able to explain

their effects both in 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, Brewer, & Collins, 1981), and we believe a theory of psychedelic action should meet such a bar.

There are also many individuals who are psychologically healthy and do not suffer from any severe psychopathology in need of clinical intervention. If we examine these individuals, we find that rather than lacking in stable and reliable beliefs, we could instead describe them as being canalized in healthy and adaptive ways for the environmental context that they find themselves in (Olaru, van Scheppingen, Bleidorn, & Denissen, 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. While it could be argued that such high-functioning individuals could potentially enjoy even greater states of well-being by using psychedelics (cf. the "betterment of well people"), the potential for disrupting existing flourishing is a possibility that merits consideration.

We can understand this 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 a near-global minima in the optimization landscape underlying those beliefs. As such, movement in any direction from that minima may result in decreased adaptability. This danger is especially apparent when the real possibility to induce trauma (and thus push the individual significantly away from the global minima) through a so-called "bad trip" is present. While 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 & 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.

5 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 a global minima in their belief landscape) is in need of no additional psychological support. From the perspective of an environmental context which is apt to change or be disrupted over time however, it becomes desirable for an individual to be not only adapted, but also adaptable (Stanley & Lehman, 2015). This means being biased towards slightly less canalization in order to be receptive to the changes which one will certainly eventually encounter during a lifetime. In our contemporary post-industrial world in which social, professional, and cultural dynamics are rapidly shifting, 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 we are interested

in 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 this bias towards adaptation in the context of deep reinforcement learning, where maximum-entropy learning is provably more optimal than simple reward maximization (Ziebart, Maas, Bagnell, Dey, et al., 2008; Levine, 2018), and has likewise become a common component of state of the art systems (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017; Haarnoja, Zhou, Abbeel, & Levine, 2018). In maximum-entropy learning, an agent optimizes both the objective of accumulating the largest possible return as well as 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 over a long time horizon as well. Balancing these two objectives during the optimization process ensures that if and when the environmental context shifts, the agent is less likely to have overfit to the previous context, resulting in better learning performance over time.

We can apply this maximum entropy perspective to learning during an individual's lifetime. As previously discussed, the kinds of radical disruption in beliefs which can be occasioned by 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 which balances the local greedy optimization of the belief landscape for the given context with the knowledge that the context will likely shift in the future. If we understand psychedelics to act as relaxers of the belief landscape through their 5-HT_{2A} agonism, such a biasing of the optimization process towards entropy might be brought about by engaging in the microdosing of a 5-HT_{2A} agonist or a full dose of a 5-HT_{2A} partial agonist such as Ariadne (Cunningham et al., 2022) on a regular interval. While speculative, such a minor intervention may be sufficient to tip the balance towards a healthy level of plasticity that allows the preservation of the adaptive circuits which an individual has developed over time. These kinds of interventions have the additional benefit of reducing the risk of introducing undesirable instability in individuals' belief landscapes, or potentially helping to mitigate the development of other kinds of maladaptive mental circuits as may be involved in psychotic disorders, or potentially some neurological conditions such as epilepsy.

With the above in mind, we can understand the broader goal of the psychotherapeutic enterprise to consist not of decreasing canalization per-se, but rather in enabling the development of healthy and adaptive behavioral, cognitive, and emotional circuits which are just plastic enough to meaningfully change with an individual as they grow and move through their life, but stable enough to provide a meaningful ground upon which the individual can rely (Jedlicka, Tomko, Robins, & Abraham, 2022). From this perspective, psychedelic therapy is just one of many possible tools that a therapist might employ in order to aid an individual towards 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 "lighter" methods of belief sculpting are unable to make enough of an impact, either due to excessive canalization in the Type A or Type B optimization landscapes. In contrast, it may be undesirable to utilize psychedelic therapy in cases of psychopathologies of insufficient canalization of the Type A or Type B landscapes, or even unnecessary in some psychologically healthy adults. Contextualizing the usefulness

of psychedelic therapy allows for the integration of the canalization theory into a much broader literature on the process of the psychotherapeutic enterprise that has been accumulating for over a century (Kazdin, 2007).

6 Psychedelic action on optimization landscapes

Underpinning the canalization theory of psychopathology is the REBUS model of psychedelics. According to the model, psychedelics change the belief landscapes of the brain in a very specific way: they flatten them. Multiple pieces of evidence suggest however that rather than flattening the belief landscape high-dose psychedelics act instead to destabilize it through transient perturbation during the acute action of the drug (Safron, 2020). While this destabilization can potentially produce on-average an effect of relaxation, it is not guaranteed that this will be the outcome. As a result relaxation may not be the most useful normative framework through which to view the action of psychedelics in the belief landscape, either acutely or post acutely.

Understanding psychedelic action through the lens of destabilization makes space for a broader range of possible outcomes, both with respect to the subjective effects of the drug, as well as to the ultimate clinical outcomes. These include an acknowledgement that under the influence of these drugs individuals will sometimes take on new beliefs which they did not previously hold, either during or after the experience (Griffiths, Hurwitz, Davis, Johnson, & Jesse, 2019). Returning to the neural annealing metaphor, rather than the idea of heating a metal to the point of simply being more malleable, a better metaphor for psychedelic's action on the Type A optimization landscape may be 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 provide further nuance to this account.

Perhaps tellingly, the examples of undesirable subjective effects resulting from psychedelic use seem to concord more often with a perspective in which beliefs are being destabilized than the idealized therapeutic experience where an individual simply enters into a state of mystical or unitive consciousness (Bremler et al., 2023). One such difficulty 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 with a felt sense of inability to control one's thoughts, as well as a negative attendant affect. Cyclic thinking is an instance of a strong kind of 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. It is common for example 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 normally strongly felt beliefs 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 are simply relaxed. While it may be the case that on-average, or in the aftermath of a psychedelic experience the belief landscape has been flattened (and this may be especially true for the Type B landscape), it is likely that a more nuanced account than REBUS is necessary to capture the full range of reported phenomena.

Rather than a pedantic exercise, the differences in framing psychedelic action in terms of relaxation, strengthening, altering, or destabilizing has meaningful clinical implications. Significantly, the possibility that psychedelics can be understood as perturbing belief landscapes and destabilizing previous attractor dynamics in addition to (or rather than) flattening landscapes does not in any way diminish the therapeutic outlook for these drugs. Similarly to flattening, the destabilization of a belief landscape allows an individual to explore different possible sets of beliefs and behavioral, cognitive, and emotional circuits which one may have never thought of, or even been in a position to conceptualize before. While it can sometimes feel jarring to find oneself experiencing perceptual, cognitive, or affective beliefs with high precision which one has 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 with high precision beliefs which can be distressing, the experience of the transience of these beliefs coupled with sufficient psychotherapeutic support may also contribute to ultimately positive outcomes (Letheby, 2021). Indeed, recent evidence suggests that therapeutic outcomes from psychedelic use are predicted by reductions in experiential avoidance (Zeifman, Wagner, Monson, & Carhart-Harris, 2023). Through the process of destabilization, more positively valenced and adaptive beliefs can be arrived at and ultimately recanalized through the process of integration. It is this property of belief perturbation that enables beneficial psychotherapeutic interactions during and after psychedelic usage. The process of interacting with a trained therapist can enable the movement through the belief landscape to be interpreted and grounded. This can enable more exotic beliefs–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.

Given the different substrates of the two optimization landscapes, The way in which psychedelics impact them likewise differs. We propose that the impact of psychedelics on the Type A landscape can be understood to result from the acute effects of 5-HT_{2A} agonists directly producing excitation of cortical neurons and the associated short timescale changes in Hebbian plasticity which results from this excitation. This corresponds to the destabilization of the belief landscape, and the accompanying increase in neural markers of entropy which have been consistently reported (R. L. Carhart-Harris et al., 2014). In contrast, the effects of psychedelics on the Type B optimization landscape can be better understood to be altered through post acute changes in metaplasticity brought on through increases in various neurogenerative factors effecting the cortical neurons, including increases in BDNF and growth of dendritic spines (Calder & Hasler, 2023). See Figure 3 for a diagram of these potential effects on the optimization landscapes during psychedelic use.

Because of the different mechanisms of action at different timescales, we can expect different impacts

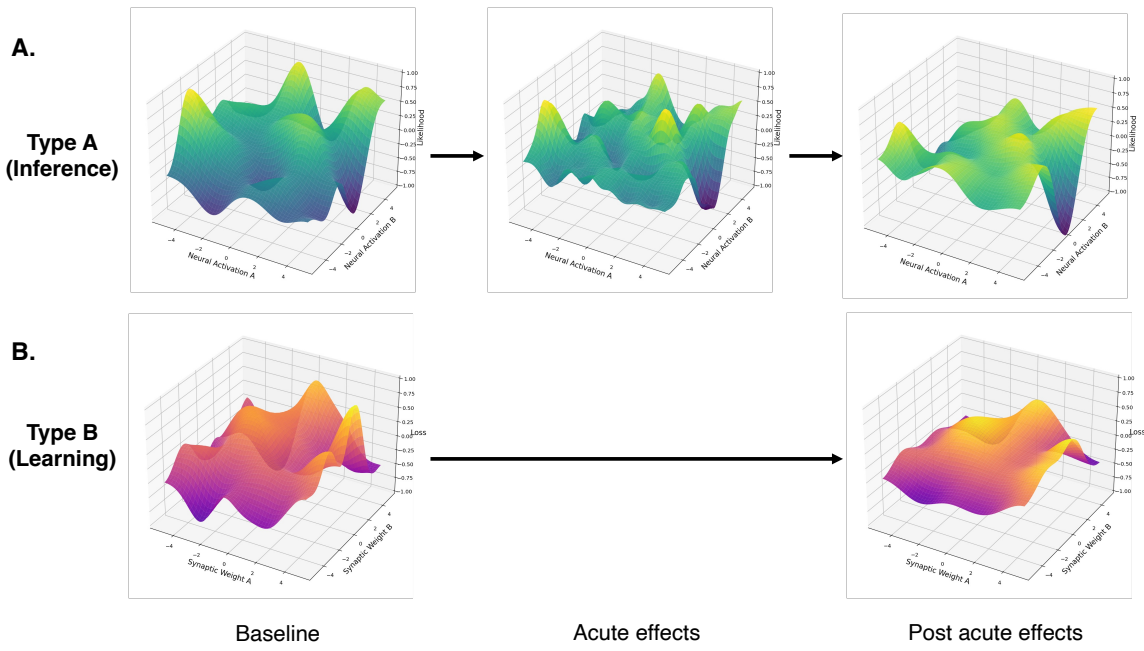

Figure 3: The impact of psychedelics on both types of optimization landscapes. A. Acute effects of psychedelics perturb the neural activation landscape, resulting in transiently strengthened and relaxed beliefs. Post acute effects result in changes to landscape topology. B. Post acute effects of psychedelics result in a smoothing of the synaptic weight optimization landscape.

on the attendant canalizations associated with each landscape. For Type A landscapes, psychedelics serve to destabilize attractors, resulting in both a transient relaxation of previously canalized beliefs, but also the potential for transiently introduced novel beliefs as well. 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, thus the need for a proper "'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.

For Type B landscapes, the effects of psychedelics can be understood to more straightforwardly correspond to decreasing the plasticity loss of the landscape, and thus a reduction in canalization. In deep neural networks, plasticity loss has been associated with the emergence of so-called "dead units," synaptic weights which no longer receive any useful gradient signal (Lyle et al., 2023). Given the recently proposed connection between gradient-based learning and neural plasticity (Richards & Kording, 2023), we can interpret the loss of dendritic volume in individuals with various psychopathologies are being potentially related to this development of "dead units" in DNNs (Qiao et al., 2016). Likewise, 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 either reset the weights of the final layer, or shrink and perturb the weights of the network (Ash & Adams, 2020; Nikishin et al., 2022; Lyle et al., 2023). Both of 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.

7 Broader clinical implications

As described above, 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 can look at these two optimization landscapes through the lens of cybernetic personality theory (DeYoung, 2015; Safron & DeYoung, 2021), a popular model of personality factors and their relationship to psychopathology. In this theoretical framework, the Big Five personality traits are grouped into two meta-traits: Stability and Plasticity. Stability represents the shared variance of the three traits of Neuroticism, Agreeableness, and Conscientiousness, while Plasticity represents the shared variance of Extraversion and Openness. Here we propose that the stability factor can be understood to reflect the topology of the Type A belief landscape. In our model, both overfitting and underfitting of this landscape would correspond to a decrease in trait stability. As such, stability can be understood as the ability of the system to maintain a metastability between these two extremes, and be capable of adaptively responding to a variety of potential environmental contexts. We can see this in the fact that the underlying traits neuroticism, agreeableness, and conscientiousness all describe certain tendencies with respect to in-context inference towards either over- or under-canalization.

The extent of canalization in the Type B landscape can then be understood to correspond to the inverse of the plasticity meta trait. Unlike stability which describes the balance between over- and under-fitting, plasticity correlates with inverse Type B canalization in a straightforwardly linear fashion. This connection can explain the finding that psychedelic use is associated with increases in trait openness, which is one of the two factors underlying the meta-trait plasticity (Lebedev et al., 2016; Erritzoe et al., 2019). Likewise, increases in extraversion (the other trait underlying plasticity) have also been found after psychedelic use (Erritzoe et al., 2018). While decreases in Type B canalization may seen desirable, there is a fine balance to be maintained (DeYoung & Krueger, 2018). Too little canalization in Type B would correspond to catastrophic forgetting, which we propose may manifest in cases of certain types of psychosis. In contrast too much canalization in Type B corresponds to plasticity loss, which in its purest sense is best characterized by certain neurodegenerative disorders, but is practically associated with various psychopathology as well. See Figure 4 for a diagramatic representation of the relationship between the two meta-traits and the two optimization landscape types.

It is worth discussing an additional note on the link between learning and canalization. In the original CANAL model, the extent to which metaplasticity was present in the system was associated with increases in learning rate (R. Carhart-Harris et al., 2022). From a deep learning perspective, learning rate is independent of either catastrophic forgetting or plasticity loss. This is because both of these 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 lead to overall increases in the magnitude of the gradient during learning, but such an outcome could

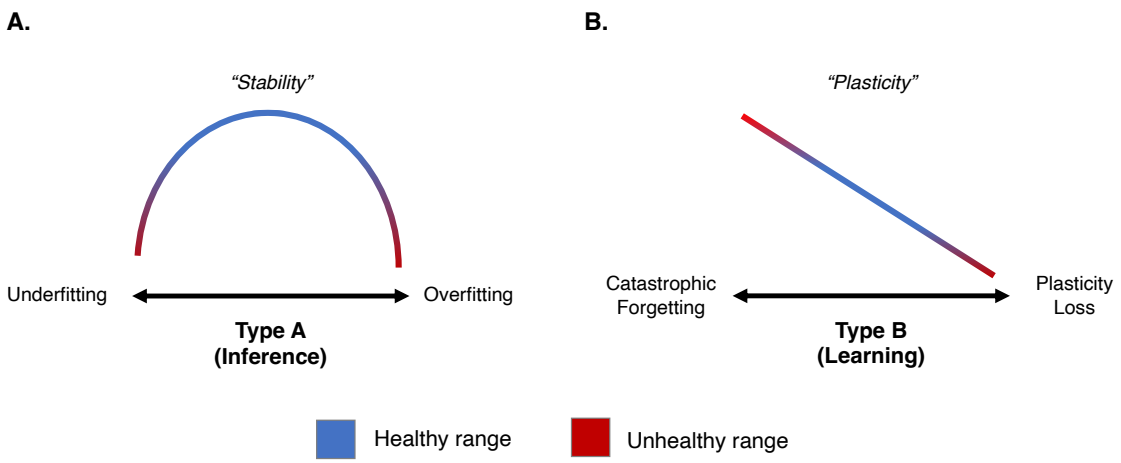

Figure 4: In the context of optimizing a deep neural network, canalization corresponds to different pathologies along two separate dimensions. A. Along the Type A dimension, this includes underfitting and overfitting. B. Along the Type B dimension, this includes catastrophic forgetting and plasticity loss. Along each dimension, optimal adaptability corresponds to maintaining a balance between extremes.

arise in the absence of any change in learning rate. Indeed increases or decreases in learning rate are likely to impact adaptation and may be linked to certain psychopathologies, but they are not necessarily relevant to the question of canalization per-se.

Using this revised conceptual framework, we can reassess the potential appropriateness of psychedelic therapy based on where an individual is along each of the two dimensions of canalization. It is important to reiterate that these two optimization landscapes may be canalized to different extents within different functional networks of the brain. Note that the examples presented below are meant for the purpose of demonstrating the applicability of the deep learning concepts to understanding psychopathology, rather than as a proposed tool for diagnoses. With this in mind, we can examine each of the four combinations of either over- or under-canalization in the two optimization landscapes. 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 of both types of optimization landscapes, then psychedelics are likely to result in the most beneficial clinical outcomes. This is because increasing plasticity 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 the landscape will be more likely to lead away from canalization rather than towards it. We believe that this is the mechanism by which the consistently compelling results for the treatment of major depressive disorder, substance use disorder, and body image disorders have been demonstrated (M. W. Johnson & Griffiths, 2017). This is also the situation in which the REBUS model most directly applies, as the likely result of psychedelic therapy would be a reduction in both types of canalization, and an attendant reduction in symptoms.

We can then turn to situations in which an individual may be under-canalized in a Type A landscape, but over-canalized in a Type B landscape. Psychopathologies potentially consistent with this configura-

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

Treatment with psychedelic therapy:

Likely beneficial | Possibly beneficial | Possibly unbeneficial | Likely unbeneficial

Figure 5: Representative psychopathologies associated with either extreme of canalization along each of the two optimization landscapes underlying belief inference. Color corresponds to potential efficacy of psychedelic therapy to address associated pathologies.

tion include attention deficit hyperactivity disorder and autism spectrum disorder (Rogers, Elison, Blain, & DeYoung, 2022). These are characterized by an inconsistent deployment of mental circuits, as well as an inability to learn or change these circuits over time. Here there is a possible benefit of psychedelic therapy, and indeed there has been some early research exploring such potential benefit (Hutten, Mason, Dolder, & Kuypers, 2019; Markopoulos, Inserra, De Gregorio, & Gobbi, 2022). The potential mechanism of this benefit would come from a reduction in Type B canalization, affording individuals with a greater capacity to learn more adaptive mental circuits over time. Still, it is less clear that such cases are ideally treated by psychedelic therapy, and this is evidenced by the relative lack of clinical attention to this treatment route compared to the treatment of major depressive disorder and other psychopathologies of a clear "dual canalization" type.

Next we can consider psychopathologies associated with under-canalization in both the Type A and Type B optimization landscapes. This would correspond to both catastrophic forgetting as well as underfitting being manifest in at least some relevant networks within the brain. Possible representative psychopathologies of this class would include borderline personality disorder as well as depersonalization-derealization disorder (Ciaunica & Safron, 2022). In both cases there is a lack of stable and reliably deployed mental circuits, as well as an inability to return to or redeploy previously successful circuits from the past. In these cases it is possibly harmful to utilize psychedelic therapy, or at the very least it should be approached with caution. This is because for individuals already high on plasticity (catastrophic forgetting) and low on stability (via underfitting) an increase in plasticity and a destabilization may not serve

to improve symptoms, and may in certain cases make them worse, as has been reported in cases where depersonalization disorder may be linked to psychedelic use (Bremler et al., 2023).

We can then consider 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 and schizophrenia, particularly cases of mania or psychosis. In these individuals there is the presence of highly overfit mental circuits being deployed in often severely maladaptive contexts. Likewise, there is an inability to retain or reuse previously adaptive circuits. In such cases psychedelic therapy is likely harmful–and is currently contraindicated–as increasing plasticity and decreasing stability has a significant chance to simply exacerbate or prolong symptoms. This has been clinically reported in cases where psychedelic use has been causally linked to the onset of psychosis in individuals with a genetic predisposition to schizophrenia (Krebs & Johansen, 2013).

	Type A (Inference)	**Type B (Learning)**
Optimization landscape	Neural activity patterns	Synaptic weights
RNN equivalent term	Hidden state	Network parameters
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	Post acute effects
Neural mechanism underlying psychedelic effect	Direct excitation; Hebbian plasticity	Metaplasticity
Primary result of psychedelic action	Mixed relaxation and strengthening of canalization	Relaxation of canalization

Table 1: A summary of the differences between the Type A and Type B optimization landscapes and the effects that psychedelics have on each.

We can finally also consider a fifth case, that of individuals who are neither over- nor under-canalized along either dimension. However, 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. Given the correct supporting environment and proper preparation however, 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. This benefit to healthy individuals who are administered psychedelics in a properly supporting clinical context has increasing support (Kraehenmann et al., 2015).

In addition to a connection to the two meta traits from cybernetic personality theory, it is also possible to understand Type A and B canalization in light of another set of psychological constructs: cognitive and psychological flexibility (Ionescu, 2012; Kashdan & Rottenberg, 2010). These two measures are related, and have each been shown to improve in individuals who undergo psychedelic therapy (Davis, Barrett, & Griffiths, 2020; Doss et al., 2021). Importantly however, they measure different underlying mechanisms. Cognitive flexibility measures the ability of an individual to rapidly adapt to the current context via the selection of the appropriate behavioral schema, which may have already been learned in the past. The construct of psychological flexibility on the other hand describes the ability to adapt to changes in life circumstances over longer timescales. We believe that the former can be associated with stability in the Type A optimization landscape, and the latter with 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. Table 1 provides a summary of the differences between the two optimization landscapes and their relationship to the measures discussed in this article.

8 Revisiting the neural annealing metaphor

In light of the above analysis, we can revisit the popular "neural annealing" theory of psychedelic activity (M. Johnson, 2019), which has served as a guiding metaphor for not only the REBUS and CANAL models, but also various popular culture descriptions of psychedelic therapy (Gómez-Emilsson, 2021). The neural annealing metaphor is one of metallurgy. It suggests that "heating up" the brain by increasing the entropy of the neural activity is in some way comparable to the increasing the temperature of a metal in order to have it worked on, and subsequently re-cooled. The "working of the metal" that is then done while the individual is in the acute or post acute psychedelic state is psychotherapy (of one form or another). Robust positive outcomes from psychedelic use often rely on them being taken in a clinical, or at the very least therapeutically conducive setting (Yaden et al., 2022). Because of this, we can understand the action of a psychedelic drug to not be directly providing therapeutic relief for the individual, but rather to be creating conditions under which maladaptive beliefs and mental circuits can be reworked.

This is somewhat in contrast however to an implicit aspect of the neural annealing metaphor. Within the metaphor is the assumption that the act of increasing the temperature of the system in-and-of-itself also produces a therapeutic effect. This is because in the actual act of metallurgy the heating and cooling of a metal results in its durability increasing due to the change in molecular organization of that compound during the process. The extent to which this aspect of the metaphor holds depends on the hypothesized mechanism behind the therapeutic action. Given the dramatic plasticity-inducing effects of these drugs, even in the absence of any subjective guidance (Ly et al., 2018), it is possible that the metaphor may apply to the extent to which these structural changes to the brain are themselves a significant causal factor

behind improvements in mental health. This is complicated by the fact that the specific mental circuits engaged during both the acute and post acute effects of the psychedelic therapy seem to be essential to meaningful positive outcomes (Yaden & Griffiths, 2020).

We can further ask on what optimization landscapes we can understand this metaphor to be operating. Given the direct connection between entropy of neural activity and the "heating" of the system, it seems at first to be most straightforward to interpret neural annealing as applying to the Type A landscape. On the other hand, the aspect of the metaphor related to a structural crystalization of desirable system dynamics seems more related to changes in the properties of the synaptic weights and thus the Type B landscape. The mixed nature of this metaphor is acknowledged in the original CANAL model, where it is suggested that acute increases in entropy lead to post-acute increases in metaplasticity (R. Carhart-Harris et al., 2022). By grounding these effects in specific neural mechanisms and separate optimization landscapes, we hope to enable a more nuanced dialogue about the metaphor in order to set better patient expectations regarding both the importance of the acute subjective experience during psychedelic therapy as well as the space of possible clinical outcomes which are possible.

9 Summary and conclusion

Psychedelic therapy has a tremendous potential to improve the lives of a great many individuals who suffer from currently poorly treatable psychopathologies. In order for it to realize its potential, the mechanisms by which it acts to change the brain must be understood. This requires as a first step the deployment of useful conceptual frameworks by which to reason about the class of drugs' action. The REBUS model of psychedelics and CANAL model of psychopathology have both been put forward as such frameworks. Utilizing the additional framework of machine learning, and deep learning in particular, we have demonstrated a few refinements to these models which we believe have the potential to increase their robustness with respect to clinical observations as well as their usefulness in experimental hypothesis generation.

Our main contribution is a refinement of the CANAL model which addresses the fact that there are two distinct forms of canalization which a dynamical learning system may develop: one of inference and one of learning. We also addressed the fact that there exists a class of psychopathologies which can be understood to arise from reduced stability in one's belief landscape, complicating the mapping between canalization and the p-factor of psychopathology. We then examined the therapeutic role of destabilization of beliefs during psychedelic use, suggesting the value of a maximum entropy perspective on adaptability and the potential value of micro dosing regimen for healthy individuals. Importantly, psychedelics act on these two optimization landscapes in different ways, with different clinical implications. The distinction between these two landscapes also maps onto a larger literature on meta-traits in personality theory, with inference corresponding to stability, and learning corresponding to plasticity. The two landscapes likewise map onto the different constructs of cognitive and psychological flexibility. Finally we examine the guiding 'neural annealing' metaphor, which is currently dominant, especially among general audiences (M. Johnson, 2019). We find that this metaphor may oversimplify the underlying mechanisms of action involved

in psychedelic therapy. Implicit in it is also the suggestion that psychedelic therapy will lead to de facto positive therapeutic outcomes, and this can be misleading, especially in non-therapeutic settings.

We believe that the framework described in this paper can serve as a foundation for a more rigorous theoretical analysis of pathologies which arise from learning systems. We acknowledge however that many of the ideas presented in this paper will require further theoretical elaboration as well as experimental validation. The maturity of the field of deep learning provides for a wide array of simulations which can empirically investigate overfitting, plasticity loss, and their complex nonlinear relationship. Likewise, as has been the case in other sub-fields of the brain sciences (Saxe, Nelli, & Summerfield, 2021), we have reason to believe that a deep learning perspective can help to inform the development of experiments both at the level of basic research, as well at 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 which are developed in a more patient-centered way. Such an outcome would serve the goals of a mature precision or computational psychiatry (Montague, Dolan, Friston, & Dayan, 2012; Fernandes et al., 2017), and ultimately contribute to more positive mental health outcomes for those most in need.

10 Acknowledgements

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

References

Abbas, Z., Zhao, R., Modayil, J., White, A., & Machado, M. C. (2023). Loss of plasticity in continual deep reinforcement learning. *arXiv preprint arXiv:2303.07507*.

Achille, A., Rovere, M., & Soatto, S. (2017). Critical learning periods in deep neural networks. *arXiv preprint arXiv:1711.08856*.

Ash, J., & Adams, R. P. (2020). On warm-starting neural network training. *Advances in neural information processing systems*, *33*, 3884–3894.

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, *46*, 1–6.

Bennett, D., Davidson, G., & Niv, Y. (2022). A model of mood as integrated advantage. *Psychological Review*, *129*(3), 513.

Bogenschutz, M. P., Forcehimes, A. A., Pommy, J. A., Wilcox, C. E., Barbosa, P. C., & Strassman, R. J. (2015). Psilocybin-assisted treatment for alcohol dependence: a proof-of-concept study. *Journal of psychopharmacology*, *29*(3), 289–299.

Bremler, R., Katati, N., Shergill, P., Erritzoe, D., & Carhart-Harris, R. (2023). Focusing on the negative: cases of long-term negative psychological responses to psychedelics.

Calder, A. E., & Hasler, G. (2023). Towards an understanding of psychedelic-induced neuroplasticity. *Neuropsychopharmacology*, *48*(1), 104–112.

Carhart-Harris, R., Chandaria, S., Erritzoe, D., Gazzaley, A., Girn, M., Kettner, H., ... others (2022). Canalization and plasticity in psychopathology. *Neuropharmacology*, 109398.

Carhart-Harris, R. L., & Friston, K. J. (2019). Rebus and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews*, *71*(3), 316–344.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., ... Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience*, 20.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... others (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical psychological science*, *2*(2), 119–137.

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

Cunningham, M. J., Bock, H. A., Serrano, I. C., Bechand, B., Vidyadhara, D., Bonniwell, E. M., ... others (2022). Pharmacological mechanism of the non-hallucinogenic 5-ht2a agonist ariadne and analogs. *ACS Chemical Neuroscience*.

Davis, A. K., Barrett, F. S., & Griffiths, R. R. (2020). Psychological flexibility mediates the relations between acute psychedelic effects and subjective decreases in depression and anxiety. *Journal of contextual behavioral science*, *15*, 39–45.

DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of research in personality*, *56*, 33–58.

DeYoung, C. G., & Krueger, R. F. (2018). A cybernetic theory of psychopathology. *Psychological Inquiry*, *29*(3), 117–138.

Doss, M. K., Madden, M. B., Gaddis, A., Nebel, M. B., Griffiths, R. R., Mathur, B. N., & Barrett, F. S. (2022). Models of psychedelic drug action: modulation of cortical-subcortical circuits. *Brain*, *145*(2), 441–456.

Doss, M. K., Považan, M., Rosenberg, M. D., Sepeda, N. D., Davis, A. K., Finan, P. H., ... others (2021). Psilocybin therapy increases cognitive and neural flexibility in patients with major depressive disorder. *Translational psychiatry*, *11*(1), 574.

Draganov, M., Galiano-Landeira, J., Doruk Camsari, D., Ramírez, J.-E., Robles, M., & Chanes, L. (2023). Noninvasive modulation of predictive coding in humans: causal evidence for frequency-specific temporal dynamics. *Cerebral Cortex*, bhad127.

Erritzoe, D., Roseman, L., Nour, M., MacLean, K., Kaelen, M., Nutt, D., & Carhart-Harris, R. (2018). Effects

of psilocybin therapy on personality structure. *Acta Psychiatrica Scandinavica*, *138*(5), 368–378.

Erritzoe, D., Smith, J., Fisher, P. M., Carhart-Harris, R., Frokjaer, V. G., & Knudsen, G. M. (2019). Recreational use of psychedelics is associated with elevated personality trait openness: Exploration of associations with brain serotonin markers. *Journal of Psychopharmacology*, *33*(9), 1068–1075.

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of 'precision psychiatry'. *BMC medicine*, *15*(1), 1–7.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128–135.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, *364*(1521), 1211–1221.

Goldberg, S. B., Pace, B. T., Nicholas, C. R., Raison, C. L., & Hutson, P. R. (2020). The experimental effects of psilocybin on symptoms of anxiety and depression: A meta-analysis. *Psychiatry research*, *284*, 112749.

Griffiths, R. R., Hurwitz, E. S., Davis, A. K., Johnson, M. W., & Jesse, R. (2019). 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*, *14*(4), e0214377.

Griffiths, R. R., Johnson, M. W., Carducci, M. A., Umbricht, A., Richards, W. A., Richards, B. D., ... Klinedinst, M. A. (2016). Psilocybin produces substantial and sustained decreases in depression and anxiety in patients with life-threatening cancer: A randomized double-blind trial. *Journal of psychopharmacology*, *30*(12), 1181–1197.

Gómez-Emilsson, A. (2021, May). *Healing trauma with neural annealing.* Retrieved from https://qri.org/blog/neural-annealing

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning* (pp. 1861–1870).

Hutten, N. R., Mason, N. L., Dolder, P. C., & Kuypers, K. P. (2019). Self-rated effectiveness of microdosing with psychedelics for mental and physical health problems among microdosers. *Frontiers in psychiatry*, *10*, 672.

Ionescu, T. (2012). Exploring the nature of cognitive flexibility. *New ideas in psychology*, *30*(2), 190–200.

Jedlicka, P., Tomko, M., Robins, A., & Abraham, W. C. (2022). Contributions by metaplasticity to solving the catastrophic forgetting problem. *Trends in Neurosciences*.

Johnson, M. (2019). Neural annealing: Toward a neural theory of everything. *URL https://opentheory.net/2019/11/neural-annealing-toward-a-neural-theory-of-ev erything*.

Johnson, M. W., & Griffiths, R. R. (2017). Potential therapeutic effects of psilocybin. *Neurotherapeutics*, *14*, 734–740.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: predictive processing, precision

and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140169.

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical psychology review*, *30*(7), 865–878.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annu. Rev. Clin. Psychol.*, *3*, 1–27.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron*, *100*(2), 424–435.

Kessler, R. C., McLaughlin, K. A., Green, J. G., Gruber, M. J., Sampson, N. A., Zaslavsky, A. M., ... others (2010). Childhood adversities and adult psychopathology in the who world mental health surveys. *The British journal of psychiatry*, *197*(5), 378–385.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, *15*(138), 20170792.

Kraehenmann, R., Preller, K. H., Scheidegger, M., Pokorny, T., Bosch, O. G., Seifritz, E., & Vollenweider, F. X. (2015). Psilocybin-induced decrease in amygdala reactivity correlates with enhanced positive mood in healthy volunteers. *Biological psychiatry*, *78*(8), 572–581.

Krebs, T. S., & Johansen, P.-Ø. (2013). Psychedelics and mental health: a population study. *PloS one*, *8*(8), e63972.

Lebedev, A. V., Kaelen, M., Lövdén, M., Nilsson, J., Feilding, A., Nutt, D. J., & Carhart-Harris, R. L. (2016). Lsd-induced entropic brain activity predicts subsequent personality change. *Human brain mapping*, *37*(9), 3203–3213.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Letheby, C. (2021). *Philosophy of psychedelics*. Oxford University Press.

Letheby, C., & Gerrans, P. (2017). Self unbound: ego dissolution in psychedelic experience. *Neuroscience of Consciousness*, *2017*(1), nix016.

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

Lozano, V., Soriano, M. F., Aznarte, J. I., Gómez-Ariza, C. J., & Bajo, M. T. (2016). Interference control commonalities in patients with schizophrenia, bipolar disorder, and borderline personality disorder. *Journal of clinical and experimental neuropsychology*, *38*(2), 238–250.

Lutkajtis, A. (2021). Entity encounters and the therapeutic effect of the psychedelic mystical experience. *Journal of Psychedelic Studies*, *4*(3), 171–178.

Ly, C., Greb, A. C., Cameron, L. P., Wong, J. M., Barragan, E. V., Wilson, P. C., ... others (2018). Psychedelics promote structural and functional neural plasticity. *Cell reports*, *23*(11), 3170–3182.

Lyle, C., Zheng, Z., Nikishin, E., Pires, B. A., Pascanu, R., & Dabney, W. (2023). Understanding plasticity in neural networks. *arXiv preprint arXiv:2303.01486*.

MacKinnon, D. F., & Pies, R. (2006). Affective instability as rapid cycling: theoretical and clinical implica-

tions for borderline personality and bipolar spectrum disorders. *Bipolar disorders*, *8*(1), 1–14.

Markopoulos, A., Inserra, A., De Gregorio, D., & Gobbi, G. (2022). Evaluating the potential use of serotonergic psychedelics in autism spectrum disorder. *Frontiers in Pharmacology*, 3341.

Martinotti, G., Santacroce, R., Pettorruso, M., Montemitro, C., Spano, M. C., Lorusso, M., ... Lerner, A. G. (2018). Hallucinogen persisting perception disorder: etiology, clinical features, and therapeutic perspectives. *Brain Sciences*, *8*(3), 47.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Elsevier.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167–202.

Millière, R., Carhart-Harris, R. L., Roseman, L., Trautwein, F.-M., & Berkovich-Ohana, A. (2018). Psychedelics, meditation, and self-consciousness. *Frontiers in psychology*, *9*, 1475.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80.

Neff, K. D. (2023). Self-compassion: Theory, method, research, and intervention. *Annual Review of Psychology*, *74*, 193–218.

Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P.-L., & Courville, A. (2022). The primacy bias in deep reinforcement learning. In *International conference on machine learning* (pp. 16828–16847).

Olaru, G., van Scheppingen, M. A., Bleidorn, W., & Denissen, J. J. (2023). The link between personality, global, and domain-specific satisfaction across the adult lifespan. *Journal of Personality and Social Psychology*.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural networks*, *113*, 54–71.

Qiao, H., Li, M.-X., Xu, C., Chen, H.-B., An, S.-C., & Ma, X.-M. (2016). Dendritic spines in depression: what we learned from animal models. *Neural plasticity*, *2016*.

Rabinovich, M. I., Varona, P., Selverston, A. I., & Abarbanel, H. D. (2006). Dynamical principles in neuroscience. *Reviews of modern physics*, *78*(4), 1213.

Richards, B. A., & Kording, K. P. (2023). The study of plasticity has always been about gradients. *The Journal of Physiology*.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... others (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Rogers, M. A., Elison, J. T., Blain, S. D., & DeYoung, C. G. (2022). A cybernetic theory of autism: Autism as a consequence of low trait plasticity. *Journal of Personality*.

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

Safron, A., & DeYoung, C. G. (2021). 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* (pp. 617–649). Elsevier.

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

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67.

Schmack, K., Schnack, A., Priller, J., & Sterzer, P. (2015). Perceptual instability in schizophrenia: Probing predictive coding accounts of delusions with ambiguous stimuli. *Schizophrenia Research: Cognition*, *2*(2), 72–77.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sessa, B. (2018). The 21st century psychedelic renaissance: heroic steps forward on the back of an elephant. *Psychopharmacology*, *235*(2), 551–560.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, *17*(11), 565–573.

Stanley, K. O., & Lehman, J. (2015). *Why greatness cannot be planned: The myth of the objective*. Springer.

Tagliazucchi, E., Roseman, L., Kaelen, M., Orban, C., Muthukumaraswamy, S. D., Murphy, K., ... others (2016). Increased global functional connectivity correlates with lsd-induced ego dissolution. *Current biology*, *26*(8), 1043–1050.

Vollenweider, F. X., & Geyer, M. A. (2001). A systems model of altered consciousness: integrating natural and drug-induced psychoses. *Brain research bulletin*, *56*(5), 495–507.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the new York Academy of Sciences*, *1464*(1), 242–268.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological bulletin*, *134*(2), 163.

Wimsatt, W. C., Brewer, M., & Collins, B. (1981). Robustness, reliability, and overdetermination. *Characterizing the Robustness of Science. Boston Studies in the Philosophy of Science*, *292*, 61–87.

Winterer, G., Musso, F., Beckmann, C., Mattay, V., Egan, M. F., Jones, D. W., ... Weinberger, D. R. (2006). Instability of prefrontal signal processing in schizophrenia. *American Journal of Psychiatry*, *163*(11), 1960–1968.

Yaden, D. B., Earp, D., Graziosi, M., Friedman-Wheeler, D., Luoma, J. B., & Johnson, M. W. (2022). Psychedelics and psychotherapy: cognitive-behavioral approaches as default. *Frontiers in psychology*, *13*, 1604.

Yaden, D. B., & Griffiths, R. R. (2020). The subjective effects of psychedelics are necessary for their enduring therapeutic effects. *ACS Pharmacology & Translational Science*, *4*(2), 568–572.

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series* (Vol. 1168, p. 022022).

Zeifman, R. J., Spriggs, M. J., Kettner, H., Lyons, T., Rosas, F., Mediano, P. A., ... Carhart-Harris, R. (2022). From relaxed beliefs under psychedelics (rebus) to revised beliefs after psychedelics (rebas): Preliminary development of the relaxed beliefs questionnaire (reb-q).

Zeifman, R. J., Wagner, A. C., Monson, C. M., & Carhart-Harris, R. L. (2023). How does psilocybin therapy work? an exploration of experiential avoidance as a putative mechanism of change. *Journal of Affective Disorders*.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. (2008). Maximum entropy inverse reinforcement learning. In *Aaai* (Vol. 8, pp. 1433–1438).