

6

Resetting the brain

The Brain—is wider than the Sky—
For—put them side by side—
The one the other will contain
With ease—and You—beside –

The Brain is deeper than the sea—
For—hold them—Blue to Blue—
The one the other will absorb—
As Sponges—Buckets—do—

The Brain is just the weight of God—
For—Heft them—Pound for Pound—
And they will differ—if they do—
As Syllable from Sound—

Emily Dickinson, 'The brain'

6.1 Introduction

Psychedelic therapy involves changes to both neural and phenomenological markers of self-representation. Some have suggested that the therapeutic value of psychedelics lies in their ability to disintegrate rigid and dysfunctional patterns of connectivity in networks such as the Default Mode Network (DMN) and the Salience Network (SN). However, this *Reset Theory* of psychedelic therapy does not, on its own, illuminate the link between such neural changes and their characteristic phenomenological and behavioral correlates. In Chapter 5, I proposed that this link can be illuminated by neurocognitive theories that specify the information-processing functions of relevant neural systems. Therefore, in this chapter and the next, I will outline a neurocognitive theory of self-representation that connect these levels of explanation. This is the *predictive self-binding* account introduced by Letheby and Gerrans (2017), which promises to explain psychological facts about psychedelic therapy in terms of their neurocognitive underpinnings.

The predictive self-binding account invokes key principles from the predictive processing (PP) theory of cognition. According to PP, the brain is an inference engine that builds hierarchical models of the world and uses them to predict its future inputs. These models are updated on principles of error detection and correction.

Lower levels model more concrete perceptual features over smaller spatiotemporal scales, whereas higher levels model increasingly abstract features over larger scales. Each level attempts to predict the activity in the level below, and cognitive processing is driven by the imperative to minimise upward-flowing prediction error. On this view, conscious experience is a ‘controlled hallucination’—an internal virtual reality or world-simulation created by this modelling process. However, the mental models that constitute the furniture of our experiential worlds are *transparent*, meaning that they cannot easily be recognised *as* models. Phenomenologically, it is as though self, body, and world are simply, immediately, and unquestionably present (Metzinger 2003, 2014).

Carhart-Harris and Friston (2019) have proposed a PP-based account of psychedelic action that they call the REBUS model (‘RElaxed Beliefs Under pSychedelics’). According to the REBUS model, the most important and distinctive effects of psychedelics, especially at high doses, result from the weakening or relaxation of abstract, domain-general beliefs that reside toward the highest levels of the brain’s hierarchical predictive models. These fundamental beliefs, or high-level *priors*, are implemented in multimodal association areas such as the cortical midline hubs of the DMN and SN. In sober waking cognition, they structure and constrain the process of probabilistic inference that generates our phenomenal models of reality. Pharmacological disruption to these networks decreases the brain’s confidence in these foundational hypotheses, leading to an ‘unconstrained’ mode of cognition.

According to the REBUS model, the therapeutic effects of psychedelics stem from the fact that, by temporarily relaxing high-level beliefs, they afford an opportunity to revise deleterious beliefs that become rigidly entrenched in pathology. This raises the question: what specific beliefs are revised in the process of successful psychedelic therapy? The evidence surveyed in Chapter 5 suggests that beliefs about the self are the most plausible candidate. This is what the predictive self-binding account claims: that psychedelic therapy involves the temporary disintegration, and subsequent beneficial revision, of predictive models of the self. I begin the present chapter by sketching the Reset Theory, and then assemble ideas from PP and the REBUS model that can ground a neurocognitive interpretation thereof. In Chapter 7, I put these ideas to work in outlining the predictive self-binding account and showing how it can explain important facts about psychedelic therapy.

6.2 The Reset Theory

In order to understand how psychedelic therapy works, we need to know something about the cognitive (i.e., computational, or information processing) functions of the neural systems involved: the DMN and the SN. We saw in Chapter 5 that these networks have been linked to narrative and minimal/embodied forms of self-representation, respectively, and that abnormal activity and connectivity in them has

been linked to symptoms of multiple psychiatric pathologies (Broyd et al. 2009, Peters et al. 2016, Sha et al. 2019).

This last point provides the impetus for a class of theories based on a simple, common idea. We can call this idea the *Reset Theory* of psychedelic therapy. It says that (i) large-scale neural networks such as the DMN and SN become stuck in rigid and dysfunctional configurations in pathology, and (ii) psychedelics are therapeutic because they disintegrate these networks, creating an opportunity to ‘reset’ them to a healthier configuration. We encountered a sketch of the Reset Theory in section 5.4.1: Carhart-Harris et al. (2017) explained their findings of *increased* DMN integrity after psilocybin for treatment-resistant depression by suggesting something along these lines. This echoes the earlier suggestion by this research group, in the context of the Entropic Brain Hypothesis, that inducing a transient state of neural entropy or disorder can be therapeutic because it provides an opportunity to disengage from inflexible cognitive configurations characteristic of anxiety, depression, addiction, and obsessive-compulsive disorder (Carhart-Harris et al. 2014).

A version of the Reset Theory is formulated by Nichols et al. in terms of *hub dismantling*. As we have seen, the core nodes of the networks affected by psychedelics are densely connected ‘hub’ regions that facilitate communication and functional integration between multiple far-flung systems throughout the brain. Nichols et al. (2017) cite evidence of changes in the activity and connectivity profiles of these hub regions across various pathologies (pp. 214–215), and echo the suggestion of Stam (2014) that *hub failure* might be a unifying, transdiagnostic marker of psychiatric pathology. This could explain the remarkable apparent transdiagnostic efficacy of psychedelics, given that their network-level target seems to be precisely the connectivity profiles of these pivotal hub regions (Müller et al. 2018b, Carhart-Harris 2019). Perhaps by dismantling pathological patterns of hub connectivity, psychedelics provide an opportunity to reset these networks into a healthier state. Nichols et al. even suggest that the mystical-type experience might be a mere ‘behavioural marker’ for the real, underlying therapeutic mechanism of hub dismantling:

[There has been] debate among researchers studying psychedelic-assisted therapy as to whether the therapeutic improvements seen are due to some receptor-based change in neurochemistry, or whether the transcendent/mystical/religious experience is itself the change agent. Both of these points of view may miss the mark, however, as we will argue here that the mystical or transcendent experience may actually be a behavioral marker for the underlying therapeutic mechanism of psychedelics ... following psychedelic-induced disintegration within local networks, as well as increased global interconnectivity, connections responsible for psychiatric-disorder-associated hub failures are disrupted ... as the effect of the drug wears off, networks can reconnect in ‘healthy’ ways, in the absence of the pathological driving force(s) that originally led to hub failure and disease.

(Nichols et al. 2017, p. 215).

The suggestion that the mystical-type experience may be a therapeutically epiphenomenal ‘behavioural marker’ is ambiguous. It could mean that the non-naturalistic metaphysical ideations—the ‘mystical’ aspects of the experience, in an intuitive sense—are therapeutically unnecessary. I defended this view in Chapter 4. Or it could mean that the experience *in toto* is therapeutically epiphenomenal, because all the *real* causal action takes place at the level of hub dismantling and network rewiring. This second interpretation merits caution. Even if we grant that psychedelic therapy is, in some sense, a matter of ‘resetting the brain’ and rewiring dysfunctional neural networks, it does not follow that the experience is causally irrelevant. On a naturalist view of the type that I favour, conscious experiences are real, physical events that occur within the brain and participate in causal relations, just as much as action potentials and alpha oscillations. Neuroscience is a multilevel field, and there is no principled basis on which to single out one level of scale or description as the site of all the ‘real’ causal action (Craver 2007).

Indeed, if the ideas that I outlined in section 5.4 are on the right track, then we must go beyond *mere* talk of hub failure and network disintegration to arrive at a satisfactory integrative account of psychedelic therapy. On their own, correlations between hub failure and pathology, between hub dismantling and mystical experience, between network rewiring and symptom reduction, are mysterious. However, a neurocognitive theory that specified the computational or information-processing functions of these systems would allow us to understand why specific changes to their activity and connectivity profiles should have distinctive phenomenological and behavioural manifestations. Correlations between such neural and psychological variables can become explanatory if we can determine which cognitive functions are ordinarily performed by the regions and networks in question, and how the performance of these functions is affected by the relevant changes to these structures.

We cannot yet answer these questions definitively. However, enough evidence exists to make some informed conjectures. In collaborative work, Philip Gerrans and I have developed a neurocognitive account of self-representation in the DMN and SN, and its alteration by psychedelics (Letheby and Gerrans 2017). This account, while speculative, is consistent with the available neural and phenomenological evidence about psychedelic experience, and promises to shed light on its lasting psychological benefits. The account invokes key tenets of the PP theory of cognition; as such, a little background is in order.

6.3 Predictive processing

PP depicts the brain as an inference engine that builds hierarchical models of the world in order to predict its future inputs. Any discrepancy between predicted and actual input generates an error signal that must be cancelled, either by updating the model or by acting in the world to alter the source of the discrepant input. The overarching

imperative of the brain, on this view, is to minimise error signals by optimising its predictions of sensory inputs (Hohwy 2013, Clark 2016).

In pursuit of this aim, over a lifetime, the brain builds highly sophisticated internal models that replicate the causal structure of the extracranial world. Specifically, the predictive brain develops a *generative model*: a model¹ that recreates, or generates, a pattern of observed effects by representing the hidden causes that give rise to them (Clark 2016). In this case, the ‘observed’ effects are the brain’s sensory (including interoceptive) inputs: the hidden causes, which the brain must model, are simply the objects, properties, and events that populate the extracranial world, including the organism’s own body.

The generative model created by the brain is hierarchical, as mandated by the hierarchical causal structure of the actual world (Hohwy 2013). Lower levels model more concrete regularities at narrower spatiotemporal scales, such as basic perceptual features and objects. These are subsumed by more general or abstract regularities encoded at higher processing levels—from concrete events and situations, through (auto-) biographical narratives and social information, to unchanging fundamentals such as space, time, and self. If all goes well, the result will be a detailed and relatively accurate intracranial facsimile, encoded in patterns of synaptic connectivity, of the actual structure of the world (Gładziejewski 2016)—or, at least, of the relatively narrow portion of it that is ecologically relevant to human beings.

PP inverts traditional views of perception as a process of bottom-up, feedforward feature detection. On the received view, the process of perception begins with inputs to primary sensory areas specialised for the detection of elementary features. Feature representations generated in these areas are fed forward to areas specialised for the detection of increasingly complex and abstract patterns. As information propagates up the processing hierarchy, feature representations are integrated into percepts, which are multimodally integrated into object representations, situation models, and so on, in a linear accumulation of representational complexity (Clark 2016, p. 12).

In contrast, PP holds that perception begins from the top down, with the derivation of predictions from an existing generative model. Each level in the hierarchy generates predictions of activity in the layer below. As such, the entire model conspires to generate predictions of expected sensory inputs. (Like a Quinean web of belief, the hierarchical generative model makes direct contact with the empirical world only at the periphery.) Any successfully predicted signals are ‘cancelled out’ and travel no further up the hierarchy. Andy Clark (2016) likens the arrangement to a hierarchical corporation in which busy superiors instruct their juniors: ‘Only tell me the news!’ Only the unexpected commands valuable bandwidth. The brain endeavors to stay one step ahead of the world and stem the tide of prediction error so thoroughly that sensory surprises are almost eliminated. It is important to note that the process of error-driven

¹ For present purposes, the term ‘model’, in the context of PP, can be treated as synonymous with the term ‘mental representation’ as I defined it in section 5.4.

model updating happens rapidly and largely outside conscious awareness. In general, our experience is populated by the outputs of this modelling process, not by the process itself.

With its top-down, internalist vision of perception, PP has the startling consequence that conscious experience is little more than a ‘controlled hallucination.’ The contents of our phenomenal awareness are simply a subset of the contents of the generative model assigned the highest probability by the brain based on its total inputs to date. Conscious experience is sometimes described as ‘transparent’, meaning that it does not wear its representational or simulacry nature on its sleeve. In ordinary waking life, it does not seem to us as though our experience is comprised of generative models, virtual simulations, or intracranial mental constructions. Rather, it seems that the world itself in all its glory, with our own bodies and minds at the centre of everything, is simply, immediately, and unquestionably present. To say that consciousness is transparent, in this sense, is to say that we do not experience our mental models of the world *as* models: it is as though we look ‘through’ them and see only the world itself. This property of *phenomenal transparency* has been defined in many different ways; here, I adopt the following simple definition given by Thomas Metzinger:²

‘Transparency’ refers to a property of conscious representations, namely, that *they are not experienced as representations*. Therefore, the subject of experience feels as if [they are] in direct and immediate contact with [the] content [of the representations]. Transparent conscious representations create the phenomenology of naïve realism. An opaque phenomenal representation is one that is experienced as a representation, for example, in ... lucid dreams.

(Metzinger 2014, p. 123).

According to PP and related views, this transparency is profoundly misleading (Metzinger 2003, Revonsuo 2006). All the furniture of our waking experience—the people, animals, plants, tables, chairs, and our own bodies and selves—is as thoroughly virtual, internally constructed, and simulacry as the fantastic creations of nocturnal dreams and psychotic hallucinations. The main difference is that, in normal waking consciousness, the contents of the models that populate consciousness are strongly constrained (hence ‘*controlled* hallucination’) by complex streams of multi-modal sensory inputs from the actual extracranial world. (PP as a theory of cognition only makes sense on the assumption that there *is* an external world whose structure the brain must track in order to guide adaptive behaviour.)

² Here I remain neutral on other, related claims, such as Metzinger’s contention that the phenomenal feature of transparency results from the attentional or introspective unavailability of the earlier processing stages in which conscious representations are constructed. Such claims are, strictly speaking, distinct from the mere definition of transparent phenomenal representations as *phenomenal representations not experienced as representations*.

Thus, in the usual course of things, our models approximate the structure of this world well enough for practical purposes. The point, however, is that the experiential contents of veridical waking perception and the experiential contents of dreams and hallucinations are ontologically equivalent, and result from the same basic mechanism: the intracranial, inferential construction of a virtual, phenomenal model of reality. The core difference consists in modulations to the processing that determines the contents of those models—but in either case, an internal model is an internal model. As Metzinger puts it:

Phenomenal experience during the waking state is an online hallucination ... Phenomenal experience during the dream state, however, is just a complex offline hallucination ... the human brain [can be viewed] as a system that constantly simulates possible realities ... a fruitful way of looking at the human brain, therefore, is as a system which, even in ordinary waking states, constantly hallucinates at the world, as a system that constantly lets its internal autonomous simulational dynamics collide with the ongoing flow of sensory input, vigorously dreaming at the world and thereby generating the content of phenomenal experience.

(Metzinger 2003, pp. 51–52).

The transparency noted above arises partly from the fact that our conscious reality-models or world-simulations inherit a high degree of structure, coherence, and volition-independence from the actual world that they seek to emulate. As Antti Revonsuo (2006) puts it, our brains give us a thoroughly realistic and convincing ‘out-of-brain experience’: an experience as of simply being immediately present in a real, objective, extracranial world. The simulatory nature of our conscious reality is difficult to take seriously, other than as a merely intellectual proposition. Most of the time we cannot easily regard the constituents of our experiential worlds as virtual entities, or as products of a modelling process. Phenomenologically speaking, they simply appear as reality, and we automatically regard them as such: transparency in action.

Occasionally, however, this transparency is undermined by glitches in the neurocognitive Matrix: stubborn and startling perceptual illusions that expose the constructed and probabilistic nature of experience. One example is the Hollow Mask Illusion, in which a mask of a human face, presented in the concave orientation, is nonetheless perceived as convex (Hill and Bruce 1993). The assumption that human faces are always convex is assigned such a high probability by the brain that it simply overrides any apparent sensory evidence to the contrary. Conflicting inputs are dismissed as unreliable noise, and the perceptual model settled on by the brain represents the object in a manner that we know, intellectually, to be inaccurate. In such cases, it becomes obvious that the contents of experience are determined by a fallible sub-personal process of probabilistic inference. Our phenomenal representations become somewhat opaque: we can, unusually, experience them *as* representations—at least to an extent. (Perceptually, of course, the illusion persists, despite our metacognitive awareness of its true nature.)

The example of the Hollow Mask Illusion illustrates another central aspect of the PP scheme: the process of error-driven model updating requires the brain to estimate the relative *reliability* of both its prior expectations and its various input channels.³ In the Hollow Mask case, the prior assumption that faces are convex is assigned such a high probability that it outweighs sensory evidence to the contrary. Such evidence, no matter how persistent, is never deemed sufficiently trustworthy by the brain to override such a well-confirmed prior belief. Indeed, the sensory evidence is deemed untrustworthy precisely *because* it conflicts with such a deeply entrenched assumption. In such cases, sub-personal mechanisms in the predictive brain engage in something analogous to David Hume's refusal to countenance reports of miracles *qua* violations of natural law, on the grounds that our evidence for the inviolability of natural laws is sufficiently strong to outweigh any possible testimonial evidence for their violation.

The same principle can be seen at work in the phenomenon of binocular rivalry. When a different stimulus is presented to each half of a subject's visual field—say, a face to the left half, and a house to the right—we might expect the subject to perceive a mish-mash of a face and a house, reflecting the actual pattern of retinal stimulation. Instead, subjects alternate between seeing only a face and only a house, with brief periods of 'mixed percept' in between. The PP explanation is that the generative model maintained by the typical human brain harbors a deep-seated assumption that distinct objects never occupy the same spatial position. As such, a perceptual model positing only a house, or only a face, will always be deemed more probable than one positing a co-located house-face. But, whichever model is settled on, the conflicting input will generate high volumes of prediction error, forcing a revision. In other words, while the brain has 'decided' that there is only a house, the face-related input will generate an error signal, and vice versa: hence the alternation (Hohwy et al. 2008).

According to PP, then, evidence inconsistent with a sufficiently well-confirmed or deeply entrenched model will simply be filtered out by the brain, dismissed as noise. This leads to a crucial aspect of the PP story: as Andy Clark puts it, prior knowledge in the predictive brain is 'always both constraining and enabling' (2016, p. 288). Prior knowledge is enabling because, without predictive models, we could not perceive anything at all. Models are needed to extract the signal from the noise of our sparse and ambiguous sensory inputs, to impose coherence and structure on what would otherwise be a buzzing, blooming confusion. It is a familiar but remarkable fact that we can effortlessly recognise most instances of the letter 'A' as instances of the same category, or tokens of the same type, despite large variations in handwriting, lighting conditions, and so forth. Likewise, we can recognise the same person from multiple perspectives, in multiple modalities, across extensive changes in appearance and behaviour. All this is possible thanks to predictive models that carve away inessential detail to extract the essential, latching onto the higher-order invariant patterns that unify instances of a category.

³ Technically, these estimates of reliability are known as 'precision expectations', and are central to the PP account of attention (Hohwy 2013).

Nonetheless, extracting the essential requires dismissing and ignoring the (deemed-to-be) inessential, and this is where predictive models can carry epistemic costs as well as benefits. In the examples given above, unconscious prior beliefs that play vital roles in generating accurate reality-models can be seen, under ecologically anomalous conditions, to impede veridical perception. In these specific cases, the consequences are not too dire: the relevant situations are low-stakes, and we are able to deploy theoretical knowledge to generate accurate propositional judgements about the true states of affairs—albeit the process of perceptual inference itself remains unaffected. In other situations, however, the constraining influence of prior knowledge can cause more serious harms—psychological, social, and even epistemic. As such, weakening its constraining influence can have corresponding benefits.

6.4 Relaxed beliefs under psychedelics

The hierarchical structure of the generative model posited by PP mirrors known facts about the functional anatomy of the cortex. Lower-level regions nearer the periphery are specialised for more concrete and modality-specific processing, while higher-level supramodal areas perform more abstract, domain-general, and integrative functions.

The networks modulated in psychedelic therapy, such as the DMN and SN, lie towards the top end of this hierarchy. The cortical midline hubs of these networks are high-level multimodal ‘association’ areas rich in the 5-HT_{2A} receptors that are psychedelics’ primary molecular target (Carhart-Harris et al. 2014). According to Carhart-Harris and Friston (2019), while disruption at lower levels may play a role, it is disruption to these high-level networks that explains the most distinctive and important effects of psychedelics.

In a recent review, Carhart-Harris (2019) outlines a multilevel account of how psychedelics work in the brain: they bind to 5-HT_{2A} receptors, which are densely expressed in high-level cortical networks, especially on layer V pyramidal cells. By activating these receptors, they induce an irregular or asynchronous mode of glutamate release in the relevant cells, causing them to ‘spike’ (generate action potentials) more erratically. The net, macroscopic effect of this is to weaken the power of synchronous neuronal oscillations across multiple frequency bands, but especially in low frequencies such as the alpha rhythm. This broadband decrease in oscillatory power (Muthukumaraswamy et al. 2013) leads to three neural signatures of psychedelics. The first two are *disintegration* and *desegregation* of resting-state networks: reductions in functional connectivity within networks, and increases in functional connectivity between networks, blurring the boundaries between usually distinct systems. The third is *entropic brain activity*—an increase in the randomness, disorderliness, or unpredictability of brain activity⁴ (Carhart-Harris et al. 2014).

⁴ Interestingly, a recent study found that various meditation practices, but especially Buddhist Vipassana, also increase entropic brain activity (Vivot et al. 2020).

The obvious question is: why should this ‘cascade of neurobiological changes’ (Carhart-Harris et al. 2014, p. 16) manifest phenomenologically as experiences of ego dissolution, strange visions, intense emotions, and insights? According to the view outlined in section 5.4, it is because some of these neurobiological changes also constitute specific *neurocognitive* changes—i.e., changes to neurally implemented computational processes that generate and transform phenomenally conscious mental representations. So, the question becomes: which physiological processes are at issue, and what cognitive functions do they implement?

On the PP scheme, the networks targeted by psychedelics implement the highest levels of the brain’s generative model. This is to say that, among other things, they encode our most abstract, fundamental beliefs about self and world—our bedrock, largely unconscious, domain-general assumptions concerning space, time, and causality, the laws of logic, and the existence of the self. Philosophers have repeatedly noted the existence and significance of these foundational beliefs. Riding roughshod over important subtleties, we might liken them to Kantian categories, to Wittgensteinian ‘hinge’ or ‘framework’ propositions, to the constituents of a Kuhnian paradigm, or to the central nodes of a Quinean web of belief. In ordinary cognition, these ‘hyperpriors’ play a crucial role in constraining the brain’s hypothesis space, drastically limiting the kinds of world-models that it can generate, and thereby making the process of probabilistic inference computationally tractable (Swanson 2016).

According to the REBUS model (‘RElaxed Beliefs Under pSychedelics’) proposed by Carhart-Harris and Friston (2019), low-frequency neuronal oscillations in high-level cortical networks have the cognitive function of encoding high-level *priors* (i.e., prior beliefs) of this kind. As such, the basic cognitive effect of psychedelic-induced neuronal desynchronisation and network disintegration is to reduce the brain’s confidence in these fundamental beliefs—to ‘relax’ the beliefs by decreasing their *precision weighting*, in the jargon. Carhart-Harris and Friston suggest that ‘this straightforward model can account for the full breadth of subjective phenomena associated with the psychedelic experience’ (2019, p. 319). Given the picture of conscious experience outlined in section 6.3—that it is a controlled hallucination or virtual reality whose content is fixed by the brain’s probabilistic models—relaxation of high-level beliefs certainly offers compelling explanations of many psychedelic phenomena.

Nowhere is this clearer than when it comes to the mystical-type experience. The standard criteria for a ‘complete’ mystical-type experience read like a checklist of negated framework propositions. The *transcendence of time and space* is self-explanatory. *Paradoxicality* might result from weakening something functionally akin to the law of non-contradiction. The senses of *internal and external unity* plausibly result from a weakening of the assumption that the self exists as an entity distinct from the rest of the world (much more on this shortly).

Interestingly, the *noetic quality*—the feeling of certainty or direct knowledge—might also result from a weakening of the sense of separate selfhood. The idea here is that the sense of being an entity distinct from the object of knowledge is a prerequisite for conceiving of the possibility of error: with no sense of a distinct knowing subject,

there is no sense of a fallible epistemic connection between that subject and a distinct object of potential knowledge. Hence, doubt is inconceivable. This principle is illustrated clearly by Michael Pollan's report of his 5-MeO-DMT experience, which we encountered in Chapter 3:

I felt a tremendous rush of energy fill my head ... and then ... 'I' was no more, blasted to a confetti cloud by an explosive force. I could no longer locate [myself] in my head, because it had exploded that too, expanding to become all that there was. Whatever this was, *it was not a hallucination. A hallucination implies a reality and a point of reference and an entity to have it. None of those things remained.*

(Pollan 2018, pp. 276–277; my emphasis).⁵

From the standpoint of the REBUS model and the controlled hallucination view of phenomenal consciousness, the significance of this report is as follows: to experience a mental state *as* a hallucination requires the experience as of being an entity standing in (fallible) representational and epistemic relations of truth and error, accuracy and inaccuracy, to a world external to itself. This corresponds to what Metzinger (2003) calls the 'Phenomenal model of the intentionality relation' (PMIR). Pollan's inability to conceive of his experience as hallucinatory, and the noetic quality, can both be explained by the absence of the phenomenal self-model (PSM) and concomitant PMIR. It is because such experiences lack the seemingly ubiquitous, transparent sense of being an experiencing, knowing entity—a 'point of reference' on reality—that those who undergo them cannot readily conceptualise them as hallucinatory.

Alleged *ineffability* might not result directly from the relaxation of high-level priors, but simply from the fact that human languages have evolved to communicate representational content generated in the ordinary waking state, content consistent with the beliefs that psychedelics undermine. As such, no language has the words, or perhaps even the structure, for the sorts of mental contents inconsistent with those beliefs that emerge when their influence is diminished.⁶ The *sense of sacredness* and *deeply felt positive mood* are a little less straightforward, but it seems likely that a relevantly similar story could be told.

Thus, there is a noteworthy correspondence between defining phenomenological features of the mystical-type experience and (the negations of) the sorts of abstract beliefs that have been proposed independently as the bedrock of the standard-issue human reality-model. Moreover, increasing evidence suggests that psychedelics disrupt precisely the neural networks that encode those beliefs, according to the PP scheme. In light of these facts, the case for a naturalistic view of the psychedelic

⁵ Incidentally, when Pollan filled out the MEQ, he found that this experience scored above the threshold for a 'complete' mystical-type experience. However, this led him to conclude that the MEQ was a 'poor net' for capturing the experience (Pollan 2018, pp. 283–284).

⁶ This might also form part of the explanation of paradoxicality: Perhaps the inapplicability of ordinary language to the relevant kinds of representational contents results in an unavoidable sense of ambiguity regarding whether certain terms apply correctly or not.

experience looks quite persuasive. If key tenets of PP, such as the hierarchical structure of the brain's generative model and the controlled hallucination view of consciousness, were true, mystical-type phenomenology is *exactly the sort of thing* one should expect to result from the disintegration of high-level cortical networks. This can be construed as a naturalistic error theory of psychedelic mysticism: it explains why we should expect psychedelics to induce vivid experiences as of a cosmic consciousness, even if no such thing existed. In light of this, positing an actual cosmic consciousness to explain such experiences seems superfluous, and a naturalistic explanation more parsimonious. (We might wonder why such experiences so often seem to involve a cosmic *consciousness*, specifically, rather than simply a cosmic unity. I will return to this question in Chapter 7).

Meanwhile, when we shift our focus to the other varieties of psychedelic experience, we can see why high-level priors should constrain cognition, and relaxing them unconstrain it. The abstract beliefs disrupted by psychedelics constrain cognition in several ways. Most basically, as we have seen, they limit the brain's hypothesis space. While these foundational assumptions are deeply entrenched, the sub-personal process of probabilistic inference will never settle on a model of a spaceless, timeless, selfless, unitive world. It will never settle on a model of a world in which distinct objects stably occupy the same spatial position, or human faces are concave. And it will never settle on a model of a world in which walls 'breathe' or undulate, fractal patterns overlay the surfaces of objects, plant spirits communicate telepathically, humans shrink to the size of a thumbnail, or writhing multicoloured snakes emerge from the carpet. Hyperpriors constrain cognition by deeming many logically possible (and, indeed, logically impossible) worlds so improbable that they become cognitively and phenomenologically impossible. By diminishing the brain's confidence in its foundational axioms, psychedelics expand the phenomenological possibility space (cf. Metzinger 2003, 2011).

6.5 Resetting beliefs under psychedelics

According to the REBUS model, psychedelics alter consciousness by disintegrating high-level cortical networks, thereby decreasing the precision-weighting of domain-general high-level priors. This weakens the constraining influence of those priors on the process of probabilistic inference, thereby expanding the space of phenomenologically possible worlds. We have seen, in brief outline, how this model might explain some key phenomenological features of the psychedelic state. But what does it have to tell us about the therapeutic mechanisms of psychedelics?

A simple answer would be that the pathologies for which psychedelics show therapeutic promise are characterised by *deleterious high-level beliefs*, and undermining the brain's confidence in those beliefs, by pharmacologically disrupting their neural substrates, creates an opportunity to revise them for the better. This would amount to a neurocognitive version of the Reset Theory, which identifies the cognitive process

implemented by the relevant networks as the encoding of high-level beliefs, and the cognitive process implemented by the disintegration and ‘resetting’ of those networks as the relaxation and revision of those beliefs. This is precisely what Carhart-Harris and Friston propose:

With regard to their potential therapeutic use, we propose that psychedelics work to relax the precision weighting of pathologically overweighted priors underpinning various expressions of mental illness. We propose that this process entails an increased sensitization of high-level priors to bottom-up signaling ... and that this heightened sensitivity enables the potential revision and deweighting of overweighted priors.

(2019, pp. 316–317).

This raises the question: which specific priors are revised in the process of successful psychedelic therapy? It cannot be metaphysical beliefs about the ultimate nature of reality, or about the existence or non-existence of a divinity: as we saw in Chapter 4, successful psychedelic therapy can occur without any change in such beliefs. Moreover, several other classes of high-level beliefs disrupted by psychedelics seem like poor candidates. It is hardly plausible, for instance, that patients’ mental health is improved by adopting new beliefs about the nature of space and time, or by rejecting the law of non-contradiction.⁷

However, there is one class of beliefs that constitutes an eminently plausible candidate: beliefs about the self. The weakening or dissolution of ego boundaries is a central aspect of the psychedelic state that has repeatedly been singled out as distinctive and theoretically significant, and that unifies non-naturalistic cosmic consciousness experiences with more naturalistic ‘mystical-type’ experiences of connectedness, wonder, and awe. Moreover, as we saw in Chapter 5, both neuroscientific and psychological evidence implicates changes to self-representation in the mechanisms of psychedelic therapy.

Letheby and Gerrans (2017) propose a neurocognitive account of self-representation in the predictive brain, its disruption by psychedelics, and the potential therapeutic effects thereof. This *predictive self-binding* theory can be viewed as a special case of the REBUS model. Like the REBUS model, it holds that signature phenomenological effects of psychedelics result from the relaxation of high-level priors via the disintegration of high-level cortical networks that implement those priors. Moreover, it holds that psychedelic therapy works because this transient belief relaxation provides an opportunity to revise those beliefs for the better.

The further, distinctive emphases of the predictive self-binding account include:

- (i) the idea that the DMN and the SN encode distinct layers of a hierarchical predictive self-model;

⁷ I know of no evidence that dialetheists are, on average, more mentally healthy than monotheists.

- (ii) the idea that this model plays a cognitive *binding* function, integrating information across modalities and levels of processing, by positing the existence of a simple, indivisible, and enduring entity; and
- (iii) the idea that this model functions as a *centre of representational gravity*, governing and constraining the construction of our world-models by allocating salience, attention, and processing resources in accordance with the hypothesised goals and interests of the self.

Equipped with conceptual tools from our discussion of PP and the REBUS model, in Chapter 7, I will outline the predictive self-binding theory and its picture of how psychedelic therapy works.