

7

Unbinding the self

An eminent philosopher among my friends, who can dignify even your ugly furniture by lifting it into the serene light of science, has shown me this pregnant little fact. Your pierglass or extensive surface of polished steel made to be rubbed by a housemaid, will be minutely and multitudinously scratched in all directions; but place now against it a lighted candle as a centre of illumination, and lo! the scratches will seem to arrange themselves in a fine series of concentric circles round that little sun. It is demonstrable that the scratches are going everywhere impartially, and it is only your candle which produces the flattering illusion of a concentric arrangement, its light falling with an exclusive optical selection. These things are a parable. The scratches are events, and the candle is the egoism of any person . . .

George Eliot, *Middlemarch* .

7.1 Introduction

Many psychiatric disorders feature deleterious changes to self-related beliefs, and many treatments aim at ‘resetting’ these beliefs—at changing them for the better (Zhao et al. 2013, Moutoussis et al. 2014, Sui and Gu 2017). This is very difficult to do, however, when the beliefs themselves are heavily weighted, or assigned high prior probabilities, by the brain, and thus recalcitrant in the face of counter-evidence. The difficulty is exacerbated by the fact that high-level self-related beliefs typically exhibit the feature of phenomenal transparency introduced in Chapter 6: they do not manifest in the patient’s experience *as* beliefs, models, or representations, but simply as reality itself. Phenomenologically speaking, the interpersonal world simply *is* threatening, or the self simply *is* powerless: they are not (merely) *believed* to be so. Even when this transparency breaks to an extent, and these beliefs are recognised as such intellectually, they typically continue to exert representational dominance, filtering incoming information and determining what permeates our conscious reality-models.

By disrupting the neurocognitive substrates of the brain’s high-level beliefs, psychedelics can change this situation transiently but dramatically. They diminish the brain’s confidence in its deepest assumptions about reality, including about the self, and thereby weaken the influence of those beliefs on the contents of the phenomenal world. This has two important effects. First, subjects experience new phenomenal

worlds inconsistent with, and previously precluded by, those beliefs. Second, as a corollary, the beliefs themselves become phenomenally *opaque* as their representational nature becomes vividly apparent. Through radically altered forms of self-experience, subjects discover the contingency, mutability, and simulatory nature of their own sense of identity and habitual modes of attention. They learn directly that there are other ways of being, and other ways of seeing, because their ordinary ways of being and seeing result from a malleable modelling process.

In this chapter, I sketch a speculative account of psychedelic therapy along these lines. On a neural level, psychedelic therapy involves transient network disintegration and subsequent ‘re-wiring’. On a cognitive level, it involves decreasing the precision-weighting of high-level priors, especially concerning the self, facilitating revision of those priors. According to Letheby and Gerrans (2017), predictive self-models—priors concerning the self—perform a cognitive *binding* function: they integrate information across multiple modalities and levels of processing into a representation of a unified subject of experience. Hubs of the Salience Network (SN) integrate bodily and affective information into a model of a momentary subject of experience (i.e., the minimal self), while hubs of the Default Mode Network (DMN) situate this information in an autobiographical context, creating a story of a persisting protagonist (i.e., the narrative self). The ‘self-binding’ function of these systems parses the world into ‘me’ and ‘not-me’ and filters incoming information for self-relevance, attributing salience and allocating attention in accordance with the goals and interests of the hypothesised self. Hence, disruption to this function permits re-evaluation of basic assumptions concerning the properties and relations of the self. On a phenomenological level, this leads to changes in ‘existential feelings’ (Ratcliffe 2005): the pervasive but elusive background sense, rooted in bodily awareness, of one’s orientation to the world and the possibilities it affords. Many candidate psychological mechanisms of psychedelic therapy—including connectedness, acceptance, awe, emotional breakthrough, psychological insight, and enhanced meaning—can be subsumed by this general account of (i) disrupted self-binding leading to (ii) altered existential feelings. The common denominator is unbinding and revising predictive models of the experiencing subject and its relations to various aspects of the internal and external worlds.

The upshot is a two-factor theory of psychedelic therapy. The induction of neural, cognitive, and phenomenological plasticity, via multiple pathways, may be necessary but is not sufficient for long-term benefit. In line with Thomas Kuhn’s (1970) account of scientific paradigm shifts, it is not enough merely to undermine a dysfunctional self-model for a spell: for durable change to result, new forms of self-modelling must be discovered during the acute experience and consolidated during the subsequent period of integration.

I conclude the chapter with some brief remarks on philosophical questions about self and self-consciousness. The self-binding account might seem to license two controversial claims: that there can be states of phenomenal consciousness totally lacking self-consciousness, and that the self does not exist. I have defended both claims elsewhere (Letheby 2020, Letheby and Gerrans 2017), but neither follows

straightforwardly from the evidence, and the very real difficulties must be acknowledged. For present purposes, we can content ourselves with two weaker, but still extremely interesting, claims: (i) There can be states of consciousness lacking anything like the *ordinary* sense of self, and (ii) the *kind* of self that we ordinarily seem to be does not exist. Theoretically and existentially, this is plenty to be getting on with.

The structure of the chapter is as follows. The first three sections are devoted to outlining the predictive self-binding account of self-representation (sections 7.2, 7.3, and 7.4). The next three sections are devoted to showing how this account can explain important facts about psychedelic therapy: its beneficial effects on symptomatology and personality (sections 7.5 and 7.7), and some of the non-naturalistic metaphysical ideations that it can promote (section 7.6). One section is devoted to philosophical issues concerning self and self-consciousness (section 7.8), and a final section to summarising the discussion (section 7.9).

7.2 Predictive self-binding

In its quest to minimise prediction error, the brain builds complex hierarchical models of the world. These models must account for regularities at multiple scales and levels of abstraction. Like scientific theories, they need to latch onto the genuine, projectible patterns in past observations in order to generate accurate future predictions. And like successful scientific theories, they achieve this predictive feat by *going beyond the evidence*: by positing unobservable hidden causes whose existence would account for the relevant regularities (Yon et al. 2019).

One crucial inferential leap that the predictive brain makes is to the existence of a world of substantial objects. This leap is underdetermined by the evidence, as intractable philosophical disputes attest: our total sensory evidence is consistent with any number of metaphysical schemes, including bundle theories, varieties of nominalism, and processual and structuralist ontologies. But, whatever our reflective convictions, the sub-personal mechanisms that conjure our phenomenal worlds are congenital substantialists. Watching a gull glide over the ocean, we do not experience a free-floating mishmash of properties and parts—whiteness, wings, location, motion, and so forth—but a coherent, unified object that bears these properties and persists through time. This is remarkable given that these different properties are processed in distinct cortical regions. The question of how the brain integrates them into coherent object representations is known as the ‘binding problem’ and is a major unresolved issue in the philosophy and science of consciousness (Revonsuo 1999). The existence of binding mechanisms is demonstrated by their failure in pathological and experimental conditions—for example, a subject presented briefly with a blue circle and a red square might report perceiving a red circle and a blue square (Burwick 2014).

The predictive processing (PP) framework offers an elegant solution to the binding problem via its top-down inversion of the process of perception. PP holds that the process begins with the issuing of predictions by a pre-existing generative model. When

inputs are accurately predicted, the feedforward signal is ‘cancelled out’ and propagates no further up the hierarchy. The predictively successful model is reinforced. Only error signals resulting from unpredicted inputs demand a response.

The relevance to the binding problem is this: on PP, the brain does not start with separate property representations and face the problem of subsequently recombining them correctly (e.g., binding the whiteness to the gull and the blueness to the ocean, rather than vice versa). Rather, the process of perception begins with a rich and complex set of top-down expectations issuing from models that *already* populate the world with objects bearing properties, standing in relations, and persisting through changes. Such models are assigned high prior probabilities because of their past predictive success. Assuming such an ontology is an efficient strategy for prediction error minimisation. Regardless of the actual metaphysics of the world, postulating particulars is a parsimonious way to track regularities in complex and noisy streams of sensory input. In effect, a substantialist ontology is a data compression strategy. As Hohwy puts it:

The binding problem is ... dealt with by default: the system assumes bound attributes and then predicts them down through the cortical hierarchy ... This as it were turns the binding problem inside out because bound percepts are predicted and evidence sought in the input that will confirm the prediction.

(Hohwy 2013, p. 103).

To the extent that evidence from the input confirms the prediction—that the relevant properties do systematically co-occur—the model is reinforced. On this view, ‘binding is inference’ (Hohwy 2013, p. 103)—specifically, *abductive* inference, or inference to the best explanation. The analogy with scientific theorising is illustrative: just as the theoretician populates the unobservable world with entities bearing properties, and derives observational predictions from this ontology, so too does the predictive brain populate the ‘hidden’ extracranial world with substantial objects and thereby derive expectations of future sensory inputs.

A key difference, of course, is that the brain’s error-minimising posits become the very stuff of our phenomenal worlds—the tables, chairs, people, birds, and trees that we encounter in everyday experience. No wonder the Ship of Theseus still evokes puzzlement, after millennia: the substantialist ontology simulated by the predictive brain gives us the irresistible feeling that there is some underlying object that is *the* original ship, and a corresponding fact of the matter as to whether *that very object* remains after gradual, piecemeal, wholesale change (cf. Metzinger 2011). Of course, as Derek Parfit (1984) pointed out, this intuition is even more compelling, but every bit as questionable, when it comes to that axiologically central but ontologically elusive entity that we call ‘I’.

According to Letheby and Gerrans (2017), this symmetry is no accident, but stems from a deep symmetry between self-representation and object-representation in the predictive brain. Just as sub-personal mechanisms minimise prediction error by

modelling the existence of substantial entities in the extra-organismic world, so too do they minimise prediction error by modelling the existence of a substantial entity in the intra-organismic world: the self. As Hohwy and Michael (2017) point out, one of the major sources of inputs to the brain is the organism itself. Innumerable streams of interoceptive and proprioceptive input require prediction and regulation, and, at higher levels, abstract correspondences between internal and external variables demand explanation—as when thirst causes muscle contractions resulting in distinctive patterns of visual, proprioceptive, and gustatory feedback, or when the auditory perception of a threatening yell modulates levels of cortisol, adrenaline, and blood pressure. An efficient way to track these complex regularities is to attribute them to the actions and fluctuating fortunes of a single underlying entity—a simple and unitary particular that persists across changes in its properties.

On this view, predictive self-modelling is just another instance of a ubiquitous cognitive binding strategy. The brain needs to track stably co-instantiated collections of properties by integrating representations of those properties into unified object representations. On PP, this integration is achieved via the top-down postulation of particulars: substantial objects that bear properties and persist through time. Likewise, self-representation is a matter of integrating information across modalities, time-scales, and levels of processing into a unified representation of a simple and persisting entity—in this case, the subject of experience, the author of actions, the one to whom it all happens. Binding is a matter not just of integration, but of *segregation*: it is a process of parsing or segmenting the world into distinct objects and distinguishing them from each other by the attribution of properties. Binding makes boundaries. Likewise, *self-binding* (Sui and Humphreys 2015, Letheby and Gerrans 2017) is a matter of parsing and segmentation along the lines of a primordial division: subject and object, self and world, I and it. Metzinger speaks of the ‘phenomenal unit of identification’ (2013b, p. 5): the sub-section of the total phenomenal field that the brain represents or tags as being ‘me’, and which we therefore experience ourselves, quite simply, as *being*. In our terms, this is the sub-section of the phenomenal field whose contents are determined by representations bound to the self-model.

The basic idea of Letheby and Gerrans (2017) is that we can explain key features of self-awareness and its psychedelic disruption by integrating (i) the work of Sui and Humphreys (2015) on self-binding with (ii) the PP account of binding as top-down abductive inference. Sui and Humphreys review evidence that cognitive binding processes are enhanced at all levels, from basic perceptual processing through to higher cognition, for information deemed self-relevant. They suggest that this integrative advantage for self-relevant information is best explained by the existence of a ‘core self-representation’ implemented in cortical midline structures, including nodes of the DMN and SN, that is accessed rapidly in most cognitive tasks and unifies the otherwise disparate varieties of self-representation. Stimuli deemed as relevant to, belonging to, or constituting ‘me’ are integrated or bound into a model of a unified and persisting entity—a model that must represent not only the entity itself but also its various properties and relations. PP provides a cognitive account of how

this self-binding process is performed: the core self-representation posited by Sui and Humphreys is a hierarchical model of a unified entity underlying the flow of experience, and the top-down attribution of properties and relations to this entity binds self-relevant information across domains and levels of processing.

At the end of Chapter 6, I said that the predictive self-binding account has three distinct emphases, over and above the general principles embodied in the REBUS model (RElaxed Beliefs Under pSychedelics) of Carhart-Harris and Friston (2019):

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

In the present section, we have examined the second of these three features: the cognitive binding function of the self-model. We turn now to the first, which concerns the hierarchical structure of the self-model and its neural implementation.

7.3 Selfhood embodied and (temporally) extended

The view I have just outlined can readily explain the respective association of the SN and the DMN with minimal and narrative forms of self-awareness. As I have emphasised, the predictive brain needs to track the internal world, like the external world, at multiple spatiotemporal scales and levels of abstraction. One fundamental aspect of phenomenal selfhood is ‘spatiotemporal self-location’: the sense of occupying a particular point in space at a particular point in time. In ordinary waking consciousness, the relevant point in space is usually within the physical boundaries of the organism (Blanke and Metzinger 2009).

However, we do not merely experience ourselves as located *within* our bodies. We also have the *experience of body ownership* (EBO): the pre-reflective, unquestioned sense that ‘this is *my* body’. Like many aspects of self-awareness, the EBO can be introspectively elusive, but its existence is revealed by conditions in which it is absent or altered, such as pathologies or ingenious experimental manipulations. In depersonalisation disorder (DPD), patients sometimes report a global loss of the EBO (Sierra and David 2011). Meanwhile, in the rubber hand illusion (Costantini and Haggard 2007), a carefully calibrated combination of visual and tactile feedback tricks the predictive brain into inferring that a visually perceived rubber hand is part of the subject’s own body, resulting in a striking local alteration to this experience.

Anil Seth (2013) has argued from a PP standpoint that aspects of the minimal self, notably the EBO, depend on computational processes implemented in the anterior insular cortex (AIC), a core node of the SN. As we saw in Chapter 5, the AIC has been implicated in many different functions, including interoception, empathy, nociception (i.e., pain perception), and conscious emotional feelings. The idea of a link between interoception and emotion was given influential articulation in the nineteenth century by James and Lange, and has been revitalised in the context of modern cognitive neuroscience by the work of Antonio Damasio (1994, 1999). In philosophy, one recent exponent of an interoceptive theory of emotion is Jesse Prinz. Prinz (2004) argues that specific bodily states—‘somatic markers’, in Damasio’s parlance—have the evolved function of signalling the presence of *core relational themes* (Lazarus 1991). In other words, specific internal bodily states have been ‘set up [by evolution] to be set off by’ the perceived occurrence of adaptively relevant situation-types, such as the presence of danger, the suffering of an irrevocable loss, or the attainment of progress towards a valued goal.

One challenge for bodily theories of emotion is the lack of a simple one-to-one mapping between bodily states and emotions (or emotional feelings). For instance, increased arousal might be experienced as anxiety or elation, depending on the context (Schachter and Singer 1962). Therefore, some computational process beyond the mere representation of bodily state is required to generate the emotions we experience. Seth identifies this process with top-down probabilistic modelling in the AIC that explains interoceptive changes in terms of core relational themes. Evidence suggests that the posterior insular cortex (PIC) generates a ‘primary interoceptive image of homeostatic afferents’ (Craig 2002, p. 660)—in other words, a purely descriptive read-out of changes to the internal physiological milieu (Gerrans 2019). What transforms such interoceptive representations into the feelings that populate our emotional lives? Interpreting existing theoretical ideas in PP terms, Seth proposes that what does the trick is top-down abductive inference that models the hidden causes of these changing bodily states: namely, core relational themes. Mere elevated arousal, registered by the PIC, is transformed into anxiety or elation when it is represented by the AIC (in communication with multiple other systems) as caused by possible danger or present reward, respectively.

The phenomenology of DPD suggests a deep link between affect and the sense of self (cf. Damasio 1999) and this, too, is explicable on Seth’s hypothesis. If the AIC interprets interoceptive changes as signals of core *relational* themes, then the conclusion seems inescapable: relations require relata. If danger is present, it must be danger to someone; an opportunity must be an opportunity for someone; and so forth. DPD patients, who often display hypoactivity in the AIC, have not lost the ability to interpret bodily states as signals of importance or relevance simpliciter, but as signals of importance or relevance *to someone* (Gerrans 2015). This is why the EBO goes hand-in-hand with the minimal self—the bare feeling of being a subject or experiencing entity at a point in time. This also explains why such DPD patients may report feeling as though they have become unreal or ceased to exist.

As Klee put it in relation to ego dissolution under lysergic acid diethylamide (LSD): ‘Lacking a reliable inflow (or integration) of stimuli, particularly from his body, the subject has lost much of the basis for his self-percept’ (1963, p. 465). What DPD patients lack, on this view, is an integration of stimuli from their bodies with contextual information that reveals the significance of those stimuli for the fortunes of the organism. In healthy waking cognition, these representations are transformed into affects by top-down inferences that bind them to a representation of an entity to whom things matter (Gerrans 2015). This binding is evident phenomenologically: we do not experience a bodily feeling and a distinct thought that, for example, danger is present, but rather the feeling is simply (transparently) experienced as the very fact of danger’s being present.

On Seth’s view, then, emotional experience and the minimal self both depend importantly on top-down interoceptive inference in the AIC. He summarises the view as follows:

emotion and embodied [i.e., minimal] selfhood are grounded in active inference of those signals most likely to be ‘me’ across interoceptive and exteroceptive domains. In humans, self-related predictive [processing] simultaneously engages multiple levels of self-representation, including physiological homeostasis, physical bodily integrity, morphology and position, and—more speculatively—the metacognitive and narrative ‘I’.

(2013, p. 570).

This is essentially the view taken by Letheby and Gerrans (2017), with a couple of additional emphases. One of these is on the content of the model generated by such inferential processes, which we explored in section 7.2: it represents the self as a simple indivisible entity, or Cartesian substance. Another is on the hierarchical relations between embodied and superordinate, narrative or autobiographical levels of self-modelling, to which we now turn.

Seth’s mention of the ‘metacognitive and narrative “I”’ brings us to the role of the DMN in self-modelling. The need for this role stems from the fact that human life is not just a series of discrete, isolated, emotion-evoking encounters with the world. Such encounters form complex patterns and occur in spatiotemporally extended contexts. In many cases, especially when it comes to more abstract, evolutionarily recent (e.g., social), or long-term goals, the personal relevance of worldly encounters depends ineliminably on autobiographical information. Suppose that you end up in the elevator with someone who has your job application on her desk for review. This encounter will evoke all kinds of somatic, affective, cognitive, and attentional responses, due to the significance of this person to your long-term goals and interests, while another’s encounter with the same person may go completely unnoticed. It is only in your specific autobiographical context that this encounter takes on its distinctive and salient character, with myriad potential psychological and behavioural consequences.

On the PP scheme, higher levels of the brain's generative model track regularities at greater spatiotemporal scales, and self-modelling is no exception. As we have seen, a large body of evidence links activity and connectivity of DMN regions to self-referential tasks of a high-level, conceptual, or autobiographical nature (such as 'mental time travel'—the simulation of past and future events—and making explicit judgements about one's own personality), as well as to theory of mind (the attribution of mental states) and to mind-wandering in the task-free resting state (which often concerns self-referential, social, or autobiographical themes). This body of evidence can be explained parsimoniously by the hypothesis that a core function of the DMN is to encode the highest layers of a hierarchical predictive self-model, representing the fortunes, goals, and interests of the postulated self at greater levels of abstraction and larger spatiotemporal scales than those tracked by SN nodes.

This is the proposal of Letheby and Gerrans (2017): a hierarchical predictive model of the self is implemented in the cortical midline nodes of the SN and DMN, which represent synchronic/minimal/embodied and diachronic/narrative/autobiographical aspects of selfhood, respectively, and these distinct layers of self-modelling are integrated 'vertically' by attributing these diverse properties and relations to a single underlying entity. It is the coherent functioning and interplay of these networks in healthy, sober, waking cognition that underlie the ubiquitous but introspectively elusive sense of being a coherent persisting subject, distinct from all of its mental and physical states, but standing to them in relations of ownership, authorship, and so forth; the one who thinks the thoughts, feels the feelings, owns and inhabits the body; the one to whom it all happens, and to whom events in the world matter.

As we saw in Chapter 5, psychological insights induced by psychedelics often involve changes to the narrative self (cf. Amada et al. 2020). Aspects of narrative selfhood also illustrate the epistemic and psychological costs that predictive models can have. Deeply entrenched 'core beliefs' about the self, such as 'I am unworthy', are characteristic of depression and other pathologies, and have been theorised to play a role in their genesis and maintenance (Waller et al. 2001). Such beliefs have obvious emotional costs. PP theory illuminates their potential *epistemic* costs: if such beliefs are sufficiently highly weighted by the brain (like the assumption that human faces are convex), then conflicting evidence will be dismissed as noise.

Thus, a profoundly depressed person may simply fail to notice evidence of her worthiness. If her belief that she is unworthy is sufficiently highly weighted, then conflicting evidence will be deemed unreliable—which, on PP, means that it will go unattended. As such, it will be unable to force an update of the model. Even if the evidence is brought directly to the patient's attention, she may simply attempt to reinterpret it or explain it away as an anomaly, rather than revising her deep-seated belief that she is unworthy. Similarly, an addicted person may be unable vividly to imagine, or simulate, a non-addicted future for herself, because her identity as a (currently using) addict is bound so tightly to the self-model that it is effectively posited as an essential property of the self (Philip Gerrans, personal communication). Thus, it becomes neurocognitively impossible to represent that very entity without those properties.

As I have emphasised, predictive models have both costs and benefits. They are the only means we have for picking up on the real higher-order patterns that characterise the complex thinking, feeling, and behaviour of creatures such as ourselves. But, in their quest for computational efficiency, our predictive brains are not just substantialists but also essentialists: they reify mere patterns and tendencies into intrinsic properties. Oversimplifying the situation and failing to appreciate that multiple, seemingly dissonant patterns can co-exist, they condense us representationally to a fraction of our actual complexity, and thereby create the corresponding reality. Epistemologically, predictive models can function as echo chambers: cognitively, emotionally, and behaviourally, as self-fulfilling prophecies. Of course, transiently disrupting these models can make their very real benefits temporarily unavailable. People on heroic doses of psilocybin should not be driving cars, teaching schoolchildren, or conducting business meetings. But such disruption can also afford an opportunity to improve these models, minimising their psychosocial and epistemic costs by rendering them temporarily more flexible, plastic, and permeable by new information (Carhart-Harris and Friston 2019).

7.4 A centre of representational gravity

The first two emphases of the predictive self-binding account are (i) that the DMN and the SN encode distinct layers of a hierarchical self-model, and (ii) that this model binds or integrates multimodal stimuli by relating them to a postulated substantial entity. We have just examined those two emphases in reverse order: section 7.2 was devoted to the idea of self-binding, and section 7.3 to the hierarchical structure of the self-model.

The ideas I have just discussed concerning attention bring us to the third emphasis of the account. Dennett (1992) famously described the self as a ‘centre of narrative gravity.’ Letheby and Gerrans (2017) describe the predictive self-model as a centre of *representational* gravity, positing that it serves as an organising principle for cognitive processing: it regulates and constrains the allocation of attention, the attribution of salience, and the construction of mental representations at all levels—not just that of explicit narrative reflection.

The self-centred structure of our default phenomenal worlds is most obvious at the concrete sensorimotor level. The physical world is represented in perception as being apprehended from a specific spatiotemporal vantage point (Blanke and Metzinger 2009). The ordinary perceptual space that we inhabit from day to day is an egocentric space, in the sense of being organised relative to a centre where the self or ‘I’ is experienced as residing (Revonsuo 2006). However, our phenomenal worlds are also self-centred in more abstract senses, determined at higher levels of processing.

In particular, as I suggested in section 7.3, our sense that events in the world matter depends on inferences, computed in the AIC, which model interoceptive changes as signals of core relational themes. The AIC and the anterior cingulate cortex (ACC),

another major hub of the SN, then cooperate to initiate control signals appropriate to the content of the core relational theme—signals that allocate attention and processing resources to stimuli deemed relevant to the postulated goals and interests of the self (Menon and Uddin 2010). This underpins the phenomenal experience as of a world rich with meaning and significance, in which certain stimuli capture attention immediately and automatically, while others go unnoticed or ignored.

This process of salience attribution and attention allocation is informed by the hierarchical interactions between the SN and the DMN. The latter represents the properties, goals, and interests of the self at longer timescales and higher levels of abstraction than the former. One core function of the SN is regulating the contextually mandated switching between the DMN and anti-correlated ‘task positive’ networks (TPNs) in accordance with task demands (Sridharan et al. 2008). However, DMN nodes themselves have also been linked to attention regulation. The posterior cingulate cortex (PCC) has been called the ‘core node’ of the DMN (Davey et al. 2016) and has repeatedly been implicated in the psychedelic state (Carhart-Harris et al. 2012b, 2016c, Bouso et al. 2015, Smigielski et al. 2019b). According to one recent model, subregions of the PCC are involved in monitoring the balance (internal vs. external) and breadth (narrow vs. broad) of attention, by tuning the activity and connectivity profiles of multiple neurocognitive networks throughout the brain (Leech and Sharp 2014).

At lower levels of the processing hierarchy, the function of the amygdala deserves mention too. Findings from one cohort of patients have linked the therapeutic effects of psychedelics to changes in this region (Roseman et al. 2018b, Mertens et al. 2020), which is considered a subcortical node of the SN (Menon and Uddin 2010, Sui and Gu 2017) and was identified by Sui and Humphreys (2015) as part of the ‘core self’ network involved in self-binding. Multiple studies have found changes to amygdala activity and connectivity during the acute effects of psychedelics (e.g., Kraehenmann et al. 2015, Mueller et al. 2017, Grimm et al. 2018). Contrary to earlier descriptions as a ‘fear centre’, the amygdala has been described by one influential model as an evolved system for *relevance detection* (Sander et al. 2003) that plays a crucial role in orienting toward salient stimuli at the level of basic perceptual processing.

In a recent paper, Wanja Wiese (2019) has described the predictive self-model as a ‘hierarchical salience model’: a complex representation of the many classes, or dimensions, of stimuli that putatively matter to the organism, at multiple levels of abstraction. For Wiese, the phenomenal sense that the self is *substantial* arises from the fact that this hierarchical model unifies the many dimensions and levels of salience by attributing them all to an ‘abstract salience object’—an object that, while not concrete, is represented as strictly numerically identical over time.

It seems that there is a bidirectional, mutually reinforcing relationship between the contents of the self-model and the patterns of attention and salience that characterise default waking consciousness. Representing the self as a specific entity with a specific set of goals and interests supports characteristic patterns of attention allocation and salience attribution—and these patterns, in turn, provide evidence confirming the hypothesis: that ‘I’ am a specific entity with specific properties, including a distinctive

set of traits, goals, and interests. This illustrates the circular dynamic that can lead to the rigid entrenchment of dysfunctional self-models: the more heavily weighted are our beliefs about who we are, the less likely it is that we will think, feel, attend, or act in ways inconsistent with those beliefs, that generate high volumes of prediction error.

Thus, in ordinary waking life, our phenomenal worlds are not only organised around a central point at a basic sensorimotor level. Our attention is also captured selectively by stimuli that are represented as personally relevant and thus evoke feelings (whether subtle or intense) of salience or importance. When no stimulus in the immediate environment captures attention in this way, we often revert to the default mode of mind-wandering, whose topics typically concern self-referential themes (Smallwood and Schooler 2015). We travel mentally to the past and the future, spinning autobiographical narratives about where we have come from, where we are going, how we might seek or avoid desirable or undesirable stimuli (broadly construed), and how we might communicate these plans and motivations to conspecifics.

As Letheby and Gerrans put it:

Perceptual representations are organized in an egocentric space; interoceptive representations are interpreted as signals of adaptively relevant events or ‘core relational themes’ ... and [autobiographical] narratives are structured around the fortunes and prospects of a protagonist. At all levels, salience is attributed, attention directed, and information integrated in accordance with the [hypothesized] relevance of information to the organism’s goals.

(2017, p. 9).

If this is correct, then disruption to the coherent functioning of the relevant systems should lead to an ‘unconstrained’ mode of cognition (Carhart-Harris et al. 2012b, Swanson 2018) in which ‘consciousness itself, the basic sensation of being an ‘I,’ undergoes a remarkable change,’ and ‘the spotlight of consciousness [becomes] a floodlight which ... brings to light unsuspected details—details normally ignored because of their lack of significance’ (Watts 1964).¹ And on moderate-to-high doses of psychedelics, this is exactly what we find.

¹ What explains the difference between the beneficial disruptions to self-representation that often result from controlled psychedelic ingestion, and the detrimental disruptions that characterise pathologies such as DPD and schizophrenia? (I am indebted to Miri Albahari for pressing this question.) A complete answer is not yet available, but there are a few points to make (see also Dambrun and Ricard 2011). First, subjects in controlled settings receive psychedelics voluntarily, with preparation and foreknowledge. This decreases the probability of dysphoric responses. When psychedelics are given to unprepared or unwitting subjects, the results often *are* detrimental (Masters and Houston 1966). Second, Albahari herself has speculated that differences in attentional capacity might constitute an important difference between beneficial and pathological disruptions to self-consciousness (Albahari 2006, pp. 208–209). In this light, it is interesting to note that the psychedelic state is often described in terms of expanded or heightened attention. While *voluntary* attention regulation may be impaired, subjects describe a greater attentional capacity than in normal waking consciousness. Third, and finally, it may be important that some minimal form of self-consciousness seems to be preserved in most psychedelic experiences. Even if all forms of self-consciousness are sometimes lost, it is likely that this happens relatively briefly, with the important therapeutic processes occurring either side. The typical retention of the minimal self in beneficial psychedelic

7.5 The self unbound

The most obvious application of the predictive self-binding framework is in explaining the experience of ego dissolution. We have already seen, in Chapter 6, how the REBUS model might explain defining phenomenological features of the mystical-type experience, such as the transcendence of space and time and the noetic quality. However, when it comes to the cardinal feature of the felt loss of individual identity, and the concomitant experience of unity and interconnectedness, the predictive self-binding framework offers further, specific insight: if the DMN and SN perform their integrative functions by modelling the existence of a simple, indivisible entity, sharply bounded and distinct from the external world, then this would explain why disruption to these systems should lead to the feeling that the individual is disintegrating, dying, or merging with the cosmos (cf. Nencini and Grant 2010). This also amply explains why significant anxiety, disorientation, and confusion can result from high-dose psychedelic administration—especially when subjects are not adequately prepared or supported, but sometimes even when they are (Griffiths et al. 2008).

More specifically, the hierarchical structure of the self-model explains why the supposedly unitary phenomenon of ‘ego dissolution’ should actually have distinct phenomenological profiles (Millière et al. 2018) and neural correlates (Muthukumaraswamy et al. 2013, Carhart-Harris et al. 2014, Lebedev et al. 2015). The idea is that psychedelics disrupt different neurocognitive networks, to different extents, under different conditions—depending, no doubt, on such factors as substance, dose, set, and setting—and this leads to distinct types of ego dissolution experiences (cf. Millière 2020). Specifically, disruption to the DMN should be associated selectively with disruptions to the narrative self, while disruption to the SN should be associated with disruptions to the minimal self. This fits well with early clinical accounts that distinguished between the losses of ‘mental ego feeling’ and ‘bodily ego feeling’ under LSD (Savage 1955). As we have seen, some such accounts explained these losses in terms of disrupted mechanisms of expectation and integration (cf. prediction and binding): ‘Lacking a reliable inflow (or integration) of stimuli, particularly from his body, the subject has lost much of the basis for his self-percept’ (Klee 1963, p. 465). This is a testable prediction of the self-binding account: that narrative and minimal self-dissolution should be associated with DMN and SN disintegration, respectively (Letheby and Gerrans 2017; cf. Lebedev et al. 2015).

However, the experience of ego dissolution *per se*—the phenomenal feeling as of disintegrating or losing one’s identity—is not the most important factor in psychedelic therapy. What is more important is that the weakening of self-related beliefs creates an opportunity to revise these beliefs for the better. In successful psychedelic therapy, the

states may mark an important difference between these states and at least some pathological disruptions to self-awareness (Albahari, personal communication).

patient learns to see herself differently, which can encompass changes to many facets and dimensions of self-awareness. And this process can be facilitated by disruption to self-binding processes, even when they do not result in an experience that is naturally described as ‘ego dissolution’. As we saw in Chapter 5, recent quantitative research into therapeutic mechanisms highlights two psychological factors: (i) experiences of psychological insight, and (ii) increases in mindfulness-related capacities and psychological flexibility. I now consider these in turn to see what light the self-binding account can shed. (Please note that, in this chapter, I am pursuing a purely descriptive account of the mechanisms of psychedelic therapy. A normative epistemological evaluation of these mechanisms and their products will be the task of Chapter 8).

On the first count, regarding psychological insight, the story is fairly straightforward: by weakening the brain’s confidence in its fundamental hypotheses about who ‘I’ am—by relaxing the precision-weighting of self-related priors and unbinding the self-model—psychedelics induce different forms of self-modelling. Under the influence, patients can gain new perspectives on their lives, see things differently, and access information previously filtered out or ignored. They can thereby discover new and often healthier forms of self-modelling. This sort of process is described clearly by patients in clinical trials, as we saw in Chapter 3. Consider the example of a patient who received psilocybin-assisted therapy for alcoholism:

Lisa was a Latin-American female in her 50s with a family history of alcoholism, physical and emotional abuse, abandonment, and neglect. Her problematic drinking began around the age of 30 and resulted in social isolation, hangovers, strong feelings of guilt and shame, and severe self-critical thoughts ...

During the initial [psilocybin] session, Lisa ... examined the negative feelings that she harbored for herself and feelings of alienation from God ...

In the second session, Lisa received a higher dose of [psilocybin] and experienced an amplification of thought moving her into a confused and chaotic state. Underneath the chaotic thinking, she identified a deep well of overwhelming sadness. She was able to eventually surrender control over her thoughts and entered into a state of peacefulness, until her thoughts quieted completely. She heard her own inner voice rupturing the quiet, whispering into her ear: ‘I’m going to tell you a secret. It’s the worst-kept secret in the universe because everyone knows it but you. You are a perfect creation of the universe.’ At that moment she felt that everything in existence was unified and was made of love, though a part of her remained reluctant to fully believe this to be true. The voice repeatedly presented her with this reality, asking ‘do you believe this?’ over and over until each one of her objections had been addressed and dismissed. She examined herself and found that she finally did accept this to be true, which propelled her into a state of profound self-acceptance and wellbeing. She later said, ‘All there is is love, this is all that you are, this is all that matters.’

Following her [psilocybin] session, Lisa reported that her self-critical thoughts had dissolved and that alcohol had lost almost all of its appeal. She said that the

[psilocybin] sessions had illuminated how she had been unkind to her body and had been harming herself with alcohol. She noted her ability to manage stress and found that she was making time to care for herself through socialization, relaxation, and a resumed meditation practice. She reported improved concentration, a lack of negative self-talk, decreased anxiety, and a spacious quality of mind . . . At 54 weeks, Lisa reported a persisting reduction in alcohol consumption and alleviated anxiety.

(Bogenschutz et al. 2018, p. 5).

Clearly a key part of the therapeutic process for Lisa involved revising her self-conception on a deep level, abandoning the core beliefs underpinning her ‘severe self-critical thoughts’ and coming to see herself instead as a ‘perfect creation of the universe’. By her own account, this dissolution of self-critical thoughts led to reduced urges to drink and reduced actual drinking, as well as improved capacities for stress management and self-care—not to mention ‘a lack of negative self-talk, decreased anxiety, and a spacious quality of mind’. Her experience as described above corresponds to several items on the Psychological Insight Questionnaire (PIQ) of Davis et al. (2020), which we encountered in Chapter 5:

Realized ways my beliefs may be dysfunctional
 Discovered how aspects of my life are affecting my well-being
 Experienced validation of my life, character, values, or beliefs
 Realized the importance of my life
 Awareness of dysfunctional patterns in my actions, thoughts, and/or feelings
 Realized how critical or judgmental views I hold towards myself are dysfunctional

A different sort of experience is described by a patient who received psilocybin with psychological support for treatment-resistant depression:

[This patient] described how in adult life he ‘had always found it difficult to be emotional, found it uncomfortable to be around other people if they were emotional’ and saw ‘emotions as weakness’. As a child, his parents had told him that ‘boys don’t cry’, and as he matured he had learned to ‘put his feelings in a box because you can’t be upset, you’re a man’.

(Watts et al. 2017, p. 536).

Under psilocybin, this patient described reliving a pivotal event from his childhood that led him to internalise such attitudes:

[This patient] reported:

[I] became myself at age 7, after my [grandparent] had died. I totally was back there, so vivid, so real, I had the emotions that I would have felt at the

time: fearful, why did this happen, the naivety, the shock and confusion.
I was getting overly upset and my parents were saying ‘boys don’t cry’.

This helped him realize how his unhealthy emotional habits as an adult developed.

I saw that it’s not [a] weakness to be emotional, that’s an unhealthy attitude.
(Watts et al. 2017, p. 539)

At first glance, it might not seem obvious what this report has to do with revising the self-model. However, from his descriptions, it seems clear that this patient’s habits of emotional avoidance were intimately bound up with his conception of his own identity—first as a boy and later as a man. It seems plausible that his psychological tendency to avoid, ignore, or repress emotions, to put them ‘in a box’, stemmed in part from the deep-seated belief that ‘you can’t be upset, you’re a man’—a belief precisely about his identity and what it entailed.

Of course, the experience also involved autobiographical reappraisal: the patient vividly relived an early life event and gained a new understanding of its role in his psychological development. After the psilocybin session, he could see himself, for the first time, as someone who had developed dysfunctional emotional habits by internalising unhealthy attitudes in his childhood. This new self-conception was empowering because it suggested the possibility of alternative, healthier ways of relating to his emotions—some of which were experienced during the psilocybin session itself. This patient’s experience also corresponds to several PIQ items:

Realized how current feelings or perceptions are related to events from my past
Realized ways my beliefs may be dysfunctional
Gained a deeper understanding of events/memories from my past
Awareness of dysfunctional patterns in my actions, thoughts, and/or feelings
Realized how critical or judgmental views I hold towards myself are dysfunctional
Discovered I could explore uncomfortable or painful feelings I previously avoided
Discovered a clear pattern of avoidance in my life
Realized the nature and/or origins of my defenses or other coping strategies
Gained resolution or clarity about past traumas or hurtful events
Awareness of information that helped me understand my life

In considering these first two examples, I have focused largely on changes to the narrative self. The following report illustrates how changes to narrative, minimal, and bodily (Ho et al. 2020) aspects of self-representation can be intertwined in psychedelic therapy:

Victor, [who was] a 17-year-old high school student [at the time] when he was first diagnosed with leukemia, saw his body as riddled with cancer, and his narrative

suggested anger, grief, and an abiding rejection of his body. He described that during his psilocybin session, he experienced himself as a disembodied form and came to a point where he had to choose a body:

Until this point in the experience, I did not have a body. I was just this kind of soul, this entity ... I was shopping for a body, and the only body I could choose was my body. And this is meaningful because I had a lot of body issues associated with being sick with what chemo did to my body and how it changed. And so I was circling my body, and I saw everything that has happened to my body, all the food I have eaten, the drugs I have taken, the alcohol, the people I have had sex with, the chemo, the exercise, everything that has ever happened to my body. I took it in at once ... I kind of accepted my body for what it is, and I think up until that point I resisted that ... I saw this body for what it's worth. I picked it, it's mine ... It's more matter-of-fact—this is what it is. I think that acceptance has been liberating.

(Swift et al. 2017, p. 501).

In this experience, at least one aspect of the minimal self seems to be retained: the bare sense of being an experiencing subject, or 'I'. Nothing in Victor's report suggests total ego dissolution. However, other aspects of minimal selfhood, such as the EBO and spatiotemporal self-location, are disrupted in the experience of being a disembodied 'soul' or 'entity'. We can also see changes to bodily dimensions of the narrative self (dealing with 'bodily issues' and seeing 'everything that has happened to my body'), as well as the quintessential psychedelic shift from rejection or avoidance to acceptance—dramatised metaphorically through the vision of 'shopping for,' and choosing, a body.

Victor's vision of shopping for a body highlights another important point: often, the process of finding new and healthier forms of self-modelling involves an explicit sense of disidentifying with one's existing self-conception. This brings us to the second psychological correlate of psychedelic therapy: the increase in mindfulness-related capacities and psychological flexibility. What does the predictive self-binding account have to say about this?

The answer, in brief, is that when mental contents are unbound from the self-model they can attain a degree of *phenomenal opacity* that is closely related to mindfulness-related capacities and psychological flexibility. As we have seen, predictive models in general exhibit the property of phenomenal transparency: they are not experienced as models, but simply appear, phenomenologically, as reality itself. This applies at least equally to the self-model. The all-important substantial entity posited by the predictive brain does not appear to us as a model or a story. Rather, we experience it as nothing more or less than our own immediate, unproblematic existence. In Metzinger's words:

We do not experience the contents of our self-consciousness as the contents of a representational process ... but simply as *ourselves, living in the world right now* ...

the experiencing system, by necessity, becomes entangled in a naïve realism with regard to the contents of its own mental self-representation.

(Metzinger 2003, p. 331; *emphasis original*).

Indeed, Metzinger posits that transparency is a *defining* feature of a phenomenal self-model (PSM): to the extent that a cognitive system operates with an opaque model of its own functioning, it will not have the phenomenal experience of ‘being someone’, which can only result from the ‘naïve-realistic self-misunderstanding’ (Metzinger 2003, p. 333) involved in transparent self-modelling. This connects with an important observation from the phenomenology of mindfulness meditation: to see thoughts and feelings as *mere* thoughts and feelings is ipso facto to disidentify with them (Albahari 2006, pp. 63–64; cf. Wright 2017, p. 71, Sierra-Siebert and Jay 2020). To the extent that elements of the self-model become phenomenally opaque, they no longer form part of a self-model proper, because they are experienced as objects in awareness, rather than as the subject who is aware of those objects. The induction of phenomenal opacity into mental models, especially self-models, has been described as a central aim of traditional Buddhist meditation practices (Metzinger 2003, p. 566, Lutz et al. 2019).

My suggestion is that it can go both ways. On one hand, rendering thoughts and feelings phenomenally opaque by watching them calmly and non-reactively in meditation, seeing them as ‘*just* thoughts and feelings’, can lead to disidentification with them. On the other hand, disidentification with such mental contents, induced by psychedelic unbinding of the self-model, can confer an ability to watch them non-reactively. By reducing the brain’s confidence in its hypotheses about who, what, and where ‘I’ am, the phenomenal unit of identification can be expanded, contracted, shifted, or rendered ambiguous, such that it no longer determinately includes many of the thoughts, feelings, and perspectives with which we usually identify. These mental contents can then enter an open, spacious, non-goal-directed attention in the phenomenal guise of *mere* thoughts and feelings. No longer identified with so strongly, they evoke less defensiveness and reactivity. And if the control of attention is itself decoupled from self-representation, then the normal motivations for avoiding, suppressing, or rationalising uncomfortable thoughts and feelings may not wield their usual influence.

Through disidentification, we come to see these mental contents from the outside, as it were—as representations, rather than as the fixed essence of a substantial entity. By disidentifying with our self-representations, we see that they are mutable: they can change while we remain. The connections with psychological flexibility, the perspective-taking self, and decentring are obvious. As we have seen, some psychedelic subjects spontaneously distinguish between ‘I’ and ‘me’—the minimal experiencing subject, and the structures and roles of the narrative self, respectively—and describe an exhilarating sense of radical freedom that comes from a shift of felt identity away from the latter and towards the former. The therapeutic relevance of this experience is made clear by a patient who received psilocybin in a recent clinical trial:

In deep meditation ... there is a state which is not dissimilar to some aspects of being under the influence of psilocybin. Spaciousness, a realisation that much of what we think is concrete and can't be escaped from are actually just ways we think.
(Watts and Luoma 2020, p. 95).

Other patients describe similar experiences, both during and after a psilocybin experience:

[During the dose] It was like a holiday away from the prison of my brain, I was a ball of energy bouncing around the planet, I felt free, carefree, re-energised. [Patient 8]

[After the dose] It was like this great shroud had been lifted. [Patient 1]

You're not immersed in thought patterns, the concrete coat had come off. [Patient 4]

I felt spatial, not depressed. [Patient 6]

A feeling of space and openness. [Patient 17]

(Watts et al. 2017, p. 528).

In some cases, the entire narrative self seems to become opaque, causing the phenomenal unit of identification to contract to the minimal self—the bare feeling of subjectivity:

It became very important to distinguish between 'I' and 'Me', the latter being an object defined by patterns and structures and responsibilities—*all of which had vanished*—and the former being the subject experiencing and feeling. My normal life seemed to be all Me, all demands and responsibilities, a rushing burden which destroyed the pleasure and freedom of being 'I'.

(Durr 1970, p. 79; my emphasis).

For a few seconds, it was just like 'I'm me, and *there are no defining characteristics!*' ... that made me realise that I'm not a 'smoker'.

(Noorani et al. 2018, p. 759; my emphasis).

I am suggesting that such experiences result from the same basic mechanism as the observed increases in mindfulness-related capacities and psychological flexibility: the unbinding of mental contents (such as 'defining characteristics') from the self-model, diminishing the sense of fixed identity with them. What seems to persist for at least some time after the experience is a flattening of the 'prior landscape': various hypotheses about what kind of person 'I' am are put on a more level playing field, as the relative probabilities assigned to some are decreased, and to others increased. This

flattening of the prior landscape plausibly underlies the feelings of openness, spaciousness, and freedom—the sense of being unburdened—that characterises the short-to-medium-term afterglow (Carhart-Harris and Friston 2019).

One quite general effect of supervised psychedelic administration seems to be increased *plasticity*, or capacity for change (Carhart-Harris and Nutt 2017, Carhart-Harris 2019). The sense of freedom that comes from rendering the self-model opaque could be seen as a phenomenological manifestation of this effect. This effect manifests at other levels, too. As we have seen, psychedelics seem to stimulate the molecular mechanisms of neuroplasticity independently of their variable effects on cognition and consciousness (Ly et al. 2018). However, the flattening of the prior landscape I have just described amounts to the induction of a kind of cognitive plasticity: removing steep valleys or basins of attraction by decreasing the difference between the probabilities assigned to various hypotheses. Under these circumstances, the brain can slide more readily between world-models, trying a range of different predictive hypotheses on for size—a plausible cognitive interpretation of the fact that the psychedelic brain explores a wider range of global functional configurations than the sober waking brain (Tagliazucchi et al. 2014, Atasoy et al. 2017). Of course, the net result of this at the level of global neurodynamics is increased entropy (Carhart-Harris et al. 2014).

When it comes to the self-model specifically, by reducing the brain's confidence in its assumptions about who, what, and where 'I' am, the cognitive playing field is levelled with respect to one's own identity, allowing all sorts of hypotheses to be entertained that normally would be ruled out. This cognitive effect manifests phenomenologically, not just as ego dissolution but also as the experiences of disidentification, decentring, spaciousness, and freedom. We feel deeply and vividly that there are many ways we can be, many ways we can see ourselves, many ways we can relate to our lives and experience. Contrastively, we appreciate the extent to which we have been unknowingly identifying with a highly specific and rigid set of assumptions about who and what we are. Psychedelics induce neural, cognitive, and phenomenological forms of plasticity; the cognitive plasticity of the self-model and its phenomenological manifestations are of primary therapeutic importance.

7.6 Opacity and mystical epiphanies

Interestingly, the induction of opacity by psychedelics is not limited to the self-model. A widespread induction of phenomenal opacity might furnish a partial naturalistic explanation of an otherwise-mysterious fact: that psychedelic experiences often promote metaphysical views, such as idealism, which hold mind or consciousness to be ubiquitous or fundamental in the universe. Metzinger (2003, p. 250) has speculated that an experience of global opacity—the profound and existentially destabilising insight that one's entire *experienced* world is a purely mental construction—might be at the root of idealist philosophies. The central task of the present chapter is to explain psychedelics' therapeutic effects, not their tendency to promote non-naturalistic

aware that *what she ordinarily takes for* the external world is a state of her own body. But, absent further substantive assumptions, this is not the same as becoming aware that the *actual* external world is a state of one's own body. The former formulation extracts the accurate insight from psychedelic-induced phenomenal opacity without drawing unwarranted metaphysical conclusions.

If these speculations are on the right track, then this is what underlies many idealist-type metaphysical epiphanies induced by psychedelics: the philosophically illicit slide from (i) the realisation that what one has habitually taken for the mind-independent world is no such thing to (ii) the conclusion that there is no such thing at all. This is analogous to what Susan Haack (1999, pp. 196–197) calls the 'Passes-For Fallacy' — the argument, roughly, that:

1. Many things that have often *passed for* objective truth or honest unbiased inquiry have turned out to be no such thing
2. Therefore, there is no such thing as objective truth or honest unbiased inquiry

We will revisit inferences of this kind shortly, when we discuss the metaphysics of the self. Of course, Watts may simply be speaking loosely. But the ambiguity of his formulation shows how easily one might slide from one interpretation of this insight to another.

7.7 Psychedelic therapy: a two-factor theory

After a brief detour, we return to our scheduled programming. What sort of theory of psychedelic therapy falls out of the discussion of this chapter? On the view that I have outlined, psychedelics alter consciousness by changing the activity and connectivity profiles of pivotal hub regions at the top end of the brain's processing hierarchy. The disruption to synchronous neural oscillations in these regions amounts, cognitively, to a decrease in the precision-weighting of high-level priors, including self-related priors that integrate multimodal stimuli by attributing them to a single underlying entity. In philosophical terms, this basic cognitive effect of psychedelics is akin to striking at the heart of a Quinean web of belief. This informal observation goes some way to explaining why its downstream effects can be so variable: undermining these cognitively foundational representations, implemented in densely connected and functionally pivotal brain regions, can have unpredictable and far-reaching effects depending on the prior state of the system (Carhart-Harris and Friston 2019, p. 319).³ Under psychedelics, our self-conceptions and habits of attention are disrupted,

³ It also goes some way to explaining psychedelics' tendency to provoke explicitly philosophical reflection (i.e., reflection on fundamental questions), even in the philosophically uninitiated (Shanon 2001, 2002; cf. Langlitz 2016). I will return to this point in Chapter 9.

allowing us to experience alternative self-conceptions—in part, by attending to what we usually ignore.

In my view, this account supports a *two-factor theory* of psychedelic therapy. The first factor is the induction of plasticity at multiple levels: neural, cognitive, and phenomenological. Psychedelics induce molecular-level plasticity independently of their effects on cognition and consciousness (Ly et al. 2018), but also induce entropic brain activity by disintegrating cortical networks such as the DMN and SN. Cognitively, this network disintegration amounts to a form of plasticity because it flattens the prior landscape. At the phenomenological level, the unbinding of the self-model manifests in the increased opacity of the phenomenal self, especially its narrative layers. This leads to the feelings of freedom, spaciousness, and possibility that come from disidentifying with the contents of the narrative self, realising that there are other ways one can be, other ways one can see, and other ways one can parse experience. The sub-personal mechanisms that conjure our phenomenal worlds are not just substantialists but essentialists. Psychedelics can undermine both of these tendencies, but the weakening of tacit essentialism regarding the self is therapeutically central.

These feelings of freedom and possibility are examples of what Matthew Ratcliffe (2005, 2009) has called ‘existential feelings’: experiences of the body that also represent fundamental relations between self and world, and the space of possibilities afforded thereby. Many characteristic psychedelic experiences, including the sense of connectedness, fall into Ratcliffe’s category of existential feelings (Saarinen 2014). Indeed, dramatic change to existential feelings is probably one of the safest general characterisations of psychedelics’ variable phenomenological effects, especially at moderate-to-high doses. Drawing on Ratcliffe’s work, Juusi Saarinen notes that such dramatic changes can also cause existential feelings to move along the gradient from phenomenal transparency to opacity:

Although existential feelings generally remain in the background of experience ... they may become objects of attention and rational reflection ... Usually this happens when a notable change or shift in the existential feeling occurs, and the orientation one previously took for granted becomes conspicuous by its absence. In this process, the contingency of one’s earlier orientation is revealed ... Importantly, once an existential feeling becomes the object of attention, one’s reactions and attitudes toward it can reshape its structure.

(Saarinen 2014, p. 14).

The correspondence with descriptions of psychedelic therapy is obvious.

This first factor alone—the induction of plasticity—may be sufficient for the short-term ‘afterglow’ of several weeks, in which subjects report feeling lighter, freer, more energised and unburdened. (Incidentally, the afterglow seems to consist almost entirely of changes to existential feelings.) However, a second factor is required for truly durable change: the discovery of new forms of self-modelling during the experience,

and the consolidation of these during the subsequent period of integration. It's not just the plasticity—it's the way you use it.

On this view, patients and subjects do not merely discover, under psychedelics, that 'much of what we think is concrete and can't be escaped from are actually just ways we think'—although they certainly do discover this. They also discover *other ways of thinking* (and feeling, and attending, and so forth). They discover new, alternative forms of self-modelling, and the consolidation of these new forms is what the durable 'resetting' or 'rewiring' of high-level networks amounts to on a cognitive level. This is exemplified by the many varieties of psychological insight and emotional breakthrough that subjects report under psychedelics, all of which amount to variations on a central theme: reaching some new perspective on, belief about, or feeling toward some aspect of one's self and one's life. During the experience, subjects discover new ways of seeing, conceptualising, or relating to themselves and their lives. Subsequently, if all goes well, these new ways become consolidated as habitual modes of interpreting the world. The fact that LSD, psilocybin, and so forth act as *psychoplastogens*—agents that increase the plasticity of the mind—is part of what explains their therapeutic effects (Ly et al. 2018, Olson 2018). But the other part of the explanation is that they act as *psychedelics*—agents that reveal, or make manifest, various aspects of the mind, such as its potential for alternative self-conceptions and patterns of attention. Psychoplastogenic and (certain) psychedelic effects interact synergistically in psychedelic therapy.⁴

The fact that network resetting amounts, cognitively, to discovering and consolidating new forms of self-modelling is exemplified by reports of a kind that we have revisited repeatedly. These reports show how the phenomenal awareness of plasticity—the feeling of freedom, disidentification with the narrative self, and identification with a more minimal experiential self—can lead directly to the discovery and adoption of new, more beneficial self-conceptions:

There was a part there where I felt gone ... it wasn't like [the last psilocybin session] where I always remembered I was Adam. There was a part there where I was gone ... I remember telling myself you took this [the psilocybin] to try and help. And it is helping ... There's the old Adam, then the new Adam without alcohol.

(Nielson et al. 2018, p. 8).

⁴ The distinction between psychoplastogenic and psychedelic effects is similar to the distinction drawn by Andrew Jones (2018) between 'mind-modifying' and 'mind-manifesting' effects. However, Jones defines 'mind-manifesting' effects specifically in terms of the traditional idea that psychedelics facilitate access to previously *unconscious* mental contents. Here I remain neutral on whether psychedelics facilitate access to content that is unconscious in the classical psychoanalytical sense. My definition of psychedelic effects is broader: It encompasses any effect in which psychedelics make manifest an otherwise latent potential of the mind, such as its potential for alternative forms of self-modelling or patterns of attention.

It felt like I'd died as a smoker and was resurrected as a nonsmoker. Because it's my perception of myself, and that's how I felt. So I jumped up and I said 'I'm not a smoker anymore, it's all done.'

(Noorani et al. 2018, p. 759).

A report from a different patient in this second cohort illustrates how the process of revising one's self-conception can include metaphorical or imagistic aspects:

I used to hide sitting on the air conditioning unit on the side of my house, when I used to smoke. And so the image was me sitting there, smoking, all hunched over, stupid, smoking. And the vine just rising up and this purple flower face thing looking down at me like, 'how ridiculous!' And then I'm not really that [person], I'm really this vine, that's really me, and the Goddess within me ... how silly to think that [smoking] ... was going to do anything or solve anything. So it was really just that rising up feeling, and that powerful feeling, and it just filled me with such beauty and strength and life.

(Noorani et al. 2018, p. 759).

It is important to note that revisions to the narrative self need not only target one's fundamental attributes or roles (e.g., being 'a smoker' or 'not a smoker'). They can also involve, *inter alia*, increasing the overall coherence of one's autobiographical narratives by finding patterns, connections, and meaning within these narratives, bringing a new sense of order, unity, or satisfaction to the sense one has of one's life as a whole:

[One cancer patient] described how the psilocybin led him through a kind of life review where 'everything that had occurred to me since the day I was born until that very moment made sense.' He further explained that the psilocybin was:

... rolling back all the experiences of my life. Whether it's a sprained ankle, or a delicious meal, or my marriage, or the children, or the clothes I've bought all my life ... revealed in a more profound way ... So that everything had its place. I didn't think a lot about cancer, and in many ways, cancer wasn't in the room.

(Swift et al. 2017, p. 505).

This report clearly resonates with other descriptions, including by cancer patients (Gasser et al. 2015), of 'zooming out' and gaining *broader perspectives*. In such experiences, patients come to see their own lives and problems, including their illnesses, in a much larger context, leading to a diminished attentional fixation on prior topics of rumination. We will return to experiences of this kind momentarily.

Interestingly, in the qualitative study of psilocybin therapy for tobacco addiction conducted by Noorani et al. (2018), many patients described the transformation of their self-identity not in terms of learning something new, but in terms of

reconnecting with ‘core values’ that had been forgotten or neglected over the years. So perhaps, in some cases, subjects do not discover new forms of self-modelling so much as rediscover old, disused forms. This fits well with the contention of Nichols et al., in their presentation of the Reset Theory, that:

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 ... the brain is able to restore its network connectivity to a *predisease state*.
(2017, p. 215; my emphasis).

As we have seen, the self-model is not just a model of one’s body, thoughts, and biography, but also a model of what matters to one. This is why a rediscovery of, or reconnection with, ‘core values’ amounts to a revision of the self-model. As a hierarchical salience model (Wiese 2019), it represents the relative importance of various classes of stimuli to one’s postulated goals and interests, and uses this information to allocate attention and processing resources. This is why pharmacological disruption of the self-model can lead subjects to become aware of, or even fascinated by, stimuli that they would normally ignore—such as the pattern of veins on a leaf (Watts 1962), or the fabric of a pair of trousers (Huxley 1954).

This fact about self-modelling has an important corollary, foreshadowed above: that some of the ways in which self-models can become dysfunctional in pathology, and can be restored to a more beneficial form through psychedelic treatments, have to do with the *patterns of attention* that they underpin. Many reports from patients speak to this point. Patients who receive psychedelics often discover new ways of paying attention during the drug-induced experience—ways less filtered through the hypothesised cares and concerns of a solid, sharply bounded, isolated self:

It was less about my [terminal] illness. I was able to put it into perspective.... Not to see oneself with one’s sickness as centre. There are more important things in life ... The evolution of humankind for example ... Your Inner Ego gets diminished, I believe, and you are looking at the whole ... you are indeed starting to build relations with plants or with the entire living world around. You think less about yourself, you are thinking—across borders.

(Gasser et al. 2015, p. 26).

Moreover, they often make clear that these new patterns of attention outlast the acute experience, and sometimes describe making a deliberate effort, whether formal or informal, to consolidate these attentional changes:

I feel like a whole bunch of crap has been dumped off the surface. This stuff that made my world shut down so much and made me look at the ground and watch the clock numbers clicking by. There’s life and so many things going on, just watching that tree over there blowing in the breeze, seeing people in the street, and all the

different people in vehicles rushing by! I just feel good about being alive ... It's always there; we just don't notice, and *I'm trying to notice and not forget that I can see it at any time, I can hear it any time*. It's like waking up in the most profound way, that this is really what life is.

(Swift et al. 2017, p. 507; my emphasis).

I've been exploring whether I can bring back other sensations from [the psilocybin experience] ... I have been able to, and I've been doing a lot of meditating. I got into meditating afterward because it was like, 'I just don't want to lose this', ... I really felt like there was a real connection with Buddhism and meditation and the psilocybin experience for me. And I've been doing that everyday.

(Malone et al. 2018, p. 5).

Indeed, the predictive self-binding account of psychedelic therapy can subsume many psychological factors that have been proposed as potential therapeutic mechanisms. Shifts from feelings of disconnection to connection (Carhart-Harris et al 2018a) involve a change in the experienced boundaries or relations of the self. Shifts from attitudes of rejection or avoidance to acceptance (Watts et al. 2017) involve a change in how one experiences the relationship between the self and its thoughts, feelings, body, and world. Psychological insight (Davis et al. 2020) and emotional breakthrough (Roseman et al. 2019) are often focused on autobiographical themes, and mindfulness-related capacities (Madsen et al. 2020) centrally involve changes to the sense of self. Even the feeling of awe (Hendricks 2018) involves the experience of the 'small self': a sense of being small or insignificant relative to some vaster reality (Piff et al. 2015). Finally, changed perceptions of meaning (Hartogsohn 2018) are linked to self-representation via the functions of the SN (cf. Smigielski et al. 2020).

If these speculations are on the mark, then psychedelic therapy has a two-factor structure. The induction of plasticity at multiple levels—neural, cognitive, and phenomenological—is necessary, and may suffice for the relatively short-term afterglow. It also lays the foundation for truly durable change, because it includes the vivid experiential insight into the availability of other forms of self-modelling—other ways of seeing and being; other ways of paying attention and relating to one's world. This, I have suggested, is closely related to the induction of phenomenal opacity. But for truly durable change to occur, the induction of plasticity alone does not suffice. New, healthier forms of self-modelling must be discovered during the acute experience, and consolidated during the subsequent period of integration.

Earlier I invoked the popular analogy between the 'hypothesis-testing' predictive brain and a scientist formulating theories, generating predictions, and testing them against the evidence. In light of this analogy, it is interesting to note that psychedelic therapy is structurally analogous to a scientific paradigm shift as described by Thomas Kuhn (1970). The acute psychedelic experience amounts to a kind of neurocognitive crisis, featuring high levels of intractable anomalies (i.e., increased prediction error signalling) precipitated by disrupting the core tenets of the dominant paradigm.

(These are psychoplastogenic effects.) However, as Kuhn emphasised, crisis alone does not beget revolution. A new neurocognitive paradigm must be found that constitutes a viable replacement for the old. (This necessarily involves *psychedelic* effects.) Otherwise, once the afterglow wears off, business as usual, in the form of well-worn neural pathways and the cognitive habits that they underpin, will inexorably resume.

7.8 Self and self-consciousness

If the broad picture of psychedelic therapy and self-representation outlined here is correct, the philosophical implications are considerable. With respect to our main concerns in this book, the implications are twofold. First, the predictive self-binding account bolsters the claim that the epistemic risks of psychedelic therapy, given naturalism, are relatively small. Second, it provides a basis for the claim that the epistemic benefits of psychedelic therapy, given naturalism, are relatively large. I will argue for this second claim explicitly in Chapter 8. First, however, I should briefly mention some of the other apparent philosophical implications—specifically, for questions about self-consciousness and the metaphysics of the self.

7.8.1 Selfless consciousness

A controversial question in the philosophy of mind is whether there can be phenomenally conscious mental states that lack all forms of self-consciousness. Following Millière (2020), I will call these ‘totally selfless’ phenomenal states. (At this point, we are remaining neutral on whether any such things as selves actually exist. The question is simply whether there can be conscious experiences without the phenomenal feeling *as of* being or having a self.)

The claim that totally selfless phenomenal states are impossible has a long history. Recent empirically oriented debates have focused on putative counterexamples from pathological conditions such as akinetic mutism and epileptic automatism (Damasio 1999), as well as schizophrenic thought insertion and DPD (Graham and Stephens 1994, Billon and Kriegel 2015, Billon 2016). Other authors have examined mystical and meditative states (Forman 1998, Albahari 2006, Millière et al. 2018), and states of deep dreamless sleep in which phenomenal consciousness is allegedly retained (Thompson 2015, Windt et al. 2016). Of course, the induction of totally selfless phenomenal states has long been portrayed as an attainable goal by mystical traditions such as Buddhism and Advaita Vedanta (Albahari 2006, 2014, Metzinger 2003), and putative instances of such induction are now objects of empirical research (e.g., Winter et al. 2020). But the interpretation of relevant evidence is highly contentious.

Recently, several philosophers have turned to psychedelic research, in the hope that it might furnish clearer cases. Relatively few psychedelic experience reports describe, unambiguously, a *total* loss of self-consciousness, but some seem to. One of