Model hallucinations

Psychedelics have a remarkable capacity to violate our ideas about ourselves. Is that why they make people better?

Philip Gerrans & Chris Letheby

Psychedelic drugs are making a psychiatric comeback. After a lull of half a century, researchers are once again investigating the therapeutic benefits of psilocybin (‘magic mushrooms’) and LSD. It turns out that the hippies were on to something. There’s mounting evidence that psychedelic experiences can be genuinely transformative, especially for people suffering from intractable anxiety, depression and addiction. ‘It is simply unprecedented in psychiatry that a single dose of a medicine produces these kinds of dramatic and enduring results,’ Stephen Ross, the clinical director of the NYU Langone Center of Excellence on Addiction, told *Scientific American* in 2016.

Just what do these drugs do? Psychedelics reliably induce an altered state of consciousness known as 'ego dissolution'. The term was invented, well before the tools of contemporary neuroscience became available, to describe sensations of self-transcendence <https://aeon.co/essays/religion-has-no-monopoly-on-transcendent-experience> : a feeling in which the mind is put in touch more directly and intensely with the world, producing a profound sense of connection and boundlessness.

How does all this help those with long-term psychiatric disorders? The truth is that no one quite knows how psychedelic therapy works. Some point to a lack of knowledge about the brain, but this is a half-truth. We actually know quite a lot about the neurochemistry of psychedelics. These drugs bind to a specific type of serotonin receptor in the brain (the 5-HT2A receptor), which precipitates a complex cascade of electrochemical signalling. What we don't really understand, though, is the more complex relationship between the brain, the self and its world. Where does the subjective experience of being a person come from, and how is it related to the brute matter that we're made of?

It's here that we encounter a last frontier, metaphysically and medically. Some think the self is a real entity or phenomenon, implemented in neural processes, whose nature is gradually being revealed to us. Others say that cognitive science confirms the arguments of philosophers East and West that the self does not exist. The good news is that the mysteries of psychedelic therapy might be a hidden opportunity to finally start unravelling the controversy.

The nature of the self <https://aeon.co/essays/what-is-the-self-if-not-that-which-pays-attention> has been disputed for as long as people have reflected on their existence. Recent neuroscientific theories of selfhood are recognisably descended from venerable philosophical positions. For example, René Descartes argued <https://aeon.co/ideas/descartes-was-wrong-a-person-is-a-person-through-other-persons> that the self was an *immaterial* soul whose vicissitudes we encounter as thoughts and sensations. He thought the existence of this enduring self was the only certainty delivered by our (otherwise untrustworthy) experience.

Few neuroscientists still believe in an immaterial soul. Yet many follow Descartes in claiming that conscious experience involves awareness of a 'thinking thing': the self. There is an emerging consensus that such self-awareness is actually a form of *bodily* awareness, produced (at least in part) by *interoception*, our ability to monitor and detect autonomic and visceral processes. For example, the feeling of an elevated heart rate <https://aeon.co/ideas/the-brain-heart-dialogue-shows-racism-hijacks-perception> can provide information to the embodied organism that it is in a dangerous or difficult situation.

David Hume disagreed with Descartes. When he attended closely to his own subjectivity, he claimed to find not a self, but a mere stream of experiences. We incorrectly infer the existence of an underlying entity from this flow of experiential moments, Hume said. The modern version of this view is that we have perceptual, cognitive, sensory and, yes, bodily experiences – *but that is all.* There's an almost irresistible temptation to attribute all this to an underlying self. But this substantialist interpretation is a Cartesian mistake, according to Hume.

Certain modern philosophers, such as Thomas Metzinger, have endorsed versions of this 'no-self' view. They point to connections with non-Western traditions, such as the concept of *anatta* or no-self in Theravada Buddhism. Narrative theorists <https://aeon.co/essays/let-s-ditch-the-dangerous-idea-that-life-is-a-story> of the self adopt a similar interpretation. They argue that the mistake is to think that because we use 'I' to tell a story about experience, there must be a *real* 'I', distinct from and underlying the narrative we use to interpret and communicate the stream of experience.

Today there are neuroBuddhists, neuroCartesians and neuroHumeans all over the world, filling PowerPoint screens with images of fMRI scans supposedly congenial to their theory. Abnormal cognitive conditions, pathological or otherwise, serve as a crucial source of evidence in these debates, because they offer the chance to look at the self when it is not working 'properly'. Data floods in but consensus remains elusive. However, the emerging neuroscience of psychedelics may help resolve this impasse. For the first time ever, scientists are in a position to watch the sense of self disintegrate and reintegrate – reliably, repeatedly and safely, in the neuroimaging scanner.

Before we can properly explain the implications of this research, we need to bring in two important ideas from cognitive neuroscience. The first is the notion of *cognitive binding*. This refers to the integration of representational parts into representational wholes by the brain. If you're standing in the middle of the road with a bus coming towards you, the colour, shape and position of the bus are all being registered in different areas of your visual cortex. For your sake, your brain needs to 'bind' the right parts into the right wholes – and not, say, to combine the shape and location of the bus with the speed of the cyclist on the pavement. Fortunately, most of the time our brains manage to get it right (although experimental studies and pathologies show that they can get it wrong). But the question of *how* they do this – the so-called 'binding problem' – remains unresolved.

A possible solution comes from the *predictive processing* <https://aeon.co/essays/consciousness-is-not-a-thing-but-a-process-of-inference> theory of cognition, the second set of principles we need to introduce. The details of the framework are still hotly

debated, especially among its proponents. However, in broad outline, it views the brain as a prediction machine that models the causal structure of the world to anticipate future inputs. Any discrepancies between an expectation and an input take the form of an error signal that demands a response from the organism – either by updating the internal model, or acting to reduce the unpredicted input. Think of learning an instrument or parking a car in a tight spot: each involves a complicated series of adjustments as the brain registers discrepancies between the predicted and actual outcome of its instructions. If you can see that you have entered the car park at too sharp an angle, you might realise that the steering is more sensitive than you thought, and so rotate the wheel less next time. Humans build such accurate predictive models of their world that error signals are minimised, almost to the point of being eliminated.

The most successful perceptual models create a world and populate it with objects

The concept of a 'model' does a lot of explanatory work in predictive coding. By 'models', cognitive scientists mean mental representations that organise information and allow the brain to extract signals from noise. A classic example is the way in which we hear speech or music. The signal that reaches the ears is usually fuzzy and incomplete; a sound engineer looking at a computer display of the auditory data hitting our eardrums would see a mess that could take months of signal processing to decode. However, our brain can use its prior knowledge to produce coherent representations of words, sentences and tunes. We can hear our friends across a crowded room because we're capable of filtering and cleaning up the signal – because we have a lexicon of explanations ready to anticipate the streams of data with which we are confronted. What we ultimately experience, then, is the model that we've learned is the best fit for the information to hand, that best predicts and accounts for our perceptions before they happen.

One startling consequence of predictive coding is that perception becomes little more than a kind of controlled hallucination <https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one> . We do not experience the external world directly, but via our mind's best guess as to what is going on out there. What does this mean for ancient philosophical debates about objective reality – does the idea even make sense? The issues here are deep and complex, but suffice it to say that the predictive coding framework rests on the idea that there *is* some kind of world out there that our brains need to find a way to track. It is by approximating the structure of this reality (even if we can't apprehend its metaphysical truth or nature) that our predictive brains save us from getting run over.

Now, due to statistical regularities in the environment over time, the most predictively successful perceptual models turn out to be those that create a world and populate it with objects with particular properties, concrete and abstract, to be sensed and thought about. *This* is how our brains solve the binding problem. Past experience teaches us that certain combinations of features are more likely to co-occur than others – and this predicted coherence is increased by attributing these features to the same persisting object. So the reason that we see a bus moving towards us, rather than a mishmash of disjointed shapes and colours, is that the brain uses a model to assign such visual fluctuations to enduring things, and predicts the nature of experience as a result.

The 'bad' news is that your sense of self is nothing more than one of these rough-and-ready models. In other words, the self is a sort of meta-filter for the signals you get from the functioning of your whole organism. Our encounters with the world – actual, imagined or recalled – make us feel hot, cold, happy, sad, anxious or calm, and every gradation and combination of experience in between. Any time that the mind encounters such a flow of feelings and perceptions, it irresistibly attributes them to some underlying entity that accounts for what's going on. Just like the play of colours and shapes makes us see a bus careening towards us on the street, when happiness gives way to sadness, the mind infers that 'someone' (me) must have experienced a loss. The result is a model of a unified entity that allows us to act, think and interact – especially with other people – coherently and effectively. Self-modelling is simply an optimising strategy that allows us to bind together certain properties of the world so that they're easier to grasp. By striving to maximise predictive success, the mind irresistibly succumbs to the substantialist temptation.

This 'self-model' is complex and multi-layered. From what scientists have uncovered so far, it appears to be more like a hierarchy of models, in which each level deals with different aspects of organismic functioning. The lower levels track and maintain the integrity of bodily boundaries, and regulate homeostasis and sensory-motor encounters with the world. These feelings are then integrated with higher-level cognition that creates the sense of 'mineness' for episodes of thought, involving processes such as memory, inference and imagination. Finally, at the highest levels we can use the narrative 'I' to express the fact that experience is integrated and bound together across this hierarchy and through time. For example, in the moments preceding an important presentation, your heart might race or you might feel butterflies in your stomach. This creates a visceral awareness of the situation: danger is imminent! In turn, this feeling summons thoughts linking the current episode to past and future ones experienced by the 'same' entity: 'I'd better not stammer like I did last time; it's really important that I impress this crowd; will I ever be really good at this?' The flow of information up and down the hierarchy is unified by being attributed – or 'bound' – to a single, all-important entity called 'me'.

There's now considerable evidence about the patterns of brain activity that correspond to the hierarchical self-model. These *neural correlates* are implemented in certain brain circuits, in particular the *salience network* and the *default mode network*. The salience network allows us to feel the significance of bodily states triggered by worldly encounters. As we've discussed, organisms are constantly bombarded by information, only a fraction of which is ultimately relevant to their goals and interests. The salience network is what allows us to discern what matters and has meaning in its context. Meanwhile the default mode network underlies episodes of autobiographical thought such as memory, imagination, planning and decision-making. To simplify things a bit, we can say that the default mode network is frequently linked to the narrative self, while the salience network is associated with a more minimal, embodied self and its affective states.

How does this story explain the therapeutic effects of psychedelics? As we've seen, the self-model is an integrated bundle of predictions – and lots of these predictions, built up over a lifetime of experience, can make us deeply stressed and unhappy. A person with social anxiety expects and experiences the world to be hostile and uncontrollable because she feels vulnerable and unable to cope. The self-model that produces these feelings magnifies the adversity of her social world. Similarly, people with depression anticipate and recollect failure and unhappiness, and attribute it to their own inadequacy. Their self-model makes it hard to access positive experiences, and often feeds on itself in a negative downward spiral <https://aeon.co/essays/why-its-high-time-that-attitudes-to-addiction-changed> . Because our brains are endlessly trying to predict what's next and reduce the likelihood of error, it's no wonder that our expectations of ourselves tend to be self-fulfilling.

Theoretically we should be able to re-engineer the mechanisms of our self-model, and so change the way we organise and interpret our experience. The problem is that the self-model functions in a way that's quite similar to the lenses of our eyes. We see with them and through them, but it's almost impossible to see the lenses themselves, to really appreciate how they affect the signals that reach us, let alone take them off if they are unhelpful. In general, the mind presents us with the finished product in the form of images, not the modelling processes themselves. So too with the self: for better or worse, we feel like unified entities, not complicated and precarious hierarchical models that track and predict our organismic responses to what's happening.

That's a big part of why psychiatric disorders such as depression or anxiety are so hard to shake. It's almost impossible for the person to access an alternative way of being in the world. She might know intellectually that certain experiences are accessible, possible and beneficial, but she can't really identify with those alternative

selves. Her invisible self-model has been rigidly constructed to parse the world negatively, and to make her feel accordingly. Moreover, people often have a justified suspicion that engaging with different forms of therapy will change who they are in some fundamental way. They defend the familiar self even when it causes them distress.

Here's where psychedelics come in. These drugs put a spanner in the works of maladaptive self-models, because they affect the neural mechanisms that self-awareness springs from. At the point of ego dissolution, two things seem to happen. One, the integrity of the self-model degrades. And two, we no longer take it for granted that our experience must be interpreted by that model.

The first point simply means that the self drops away as the filter on the world. It becomes 'unbound' as the unit through which we understand our experience. This explains psychedelics users' reports of the loss of individuality, and their patterns of intense absorption in the world. The writer Aldous Huxley famously described his experience of taking mescaline like this in *The Doors of Perception* (1954): 'I was not looking now at an unusual flower arrangement. I was seeing what Adam had seen on the morning of his creation – the miracle, moment by moment, of naked existence.'

> The self *itself* does not exist as a persistent entity, but is a fundamental cognitive strategy

The second effect is more subtle. It concerns the way that psychedelics can enlighten us about the processes behind our own subjectivity. When the self falls apart and is subsequently rebuilt, the role of the self-model seems to become visible to its possessor. Yes, this offers a psychological reprieve – but more importantly, it draws attention to the difference between a world seen with and without the self. For an anxious or depressed person, psychedelics make it possible to appreciate the intermediate, representational role of the self-model. Ego dissolution offers vivid experiential proof, not only that things can be different, but that the self that conditions experience is just a heuristic, not an unchangeable, persisting thing.

So what do psychedelics reveal about the philosophical and neuroscientific controversies about the self? It seems clear to us that the self is not a mere narrative posit, as some theorists have suggested. It plays a crucial role in perceptual and emotional processing. But this does not mean, as others have claimed, that the self-model has the right attributes to qualify as a Cartesian self either. It might perform some of the right sorts of functions, but it is not the right kind of entity. The self-model plays an essential binding function in cognitive processing – but the self *itself* does not exist, at least not in the form of some persistent, substantial 'soul'. Better to

see it as a fundamental cognitive strategy, one which has developed over evolutionary time. As the science journalist James Kingsland puts it in *Siddhartha's Brain* (2016): 'It is difficult to escape the conclusion that we have evolved into an ape that takes things personally.'

That the self is a model, not a thing, doesn't mean it's completely fluid and arbitrary. Quite the opposite: it is constructed from birth over many decades. Particularly at lower levels, the cognitive processes that the self-model binds together – perception, interoception, basic regulatory mechanisms – are not especially flexible. That's why chaotic developmental environments are so damaging. Not only are they stressful in obvious ways, but in its formative years the mind has no stable patterns of experience on which to model a self.

So change can still be very hard. Imagine trying not to hear speech in your native language as meaningful: it's almost impossible. Better to learn another language, with all the effort that entails, rather than try to temporarily 'forget' your own. So too with the self. Psychedelics allow you briefly to hear your personal language of subjectivity as sound, not meaning. Whether you want to learn another language of selfhood is up to you.

aeon.co 08 August, 2017