ALIUS BULLETIN

exploring the diversity of consciousness

Martin Fortier, Matthieu Koroma & Raphaël Millière (eds.)

Kieran Fox

Karl Friston

Thomas Metzinger

David Nichols

Kevin O'Regan

Matthew Ratcliffe

Ann Taves

n°2

FEBRUARY 2018

aliusresearch.org

Acknowledgments

The editors wish to express their gratitude to Cordelia Erickson-Davis, Brendan Fleig-Goldstein, Ella Letort, Olivia Marcus, and Sarvin Tafazoli for their valuable help in proofreading the interviews, and to all the contributors for accepting to participate to this issue.

Foreword

About ALIUS

ALIUS is an international and interdisciplinary research group dedicated to the investigation of all aspects of consciousness, with a specific focus on non-ordinary or understudied conscious states traditionally classified as *altered states of consciousness*.

In Latin, *alius* means "different". This lexical choice reflects the group's mission to study the diversity of consciousness in a systematic manner. ALIUS puts a particular stress on the need for a naturalistic approach to all aspects of consciousness, including states and experiences which have long been unduly associated to parapsychology and pseudoscientific hypotheses.

To this end, it fosters a unique interdisciplinary collaboration of researchers, involving neuroscientists, psychologists, philosophers of mind, psychiatrists and anthropologists, towards the development of a systematic and scientific model of consciousness supported by both theoretical work and experimental studies. This collaboration may take the form of joint articles, blog posts, editorial work on special issues, thematic workshops and international conferences.

Find out more about the group on the website: aliusresearch.org

About the Bulletin

The ALIUS Bulletin is an annual publication featuring in-depth interviews with prominent scholars working on consciousness and its altered states (ASCs). The goal of the Bulletin is to present a clear outline of current research on ASCs across a variety of disciplines, with an emphasis on empirical work. It also aims at dispelling the widespread stigma that still plagues the notion of ASC, while allowing a wider audience to discover rigorous scientific work on the topic presented by authors in their own words.

Table of Contents

Contributors

Kieran Fox completed his PhD in cognitive neuroscience at the University of British Columbia (Vancouver) in 2016. His research focused on investigating the neural correlates of meditation, metacognition, and mind-wandering using a mix of functional (fMRI) and morphometric (DTI) neuroimaging methods. He also spearheaded several statistical meta-analyses assessing the rigor and replicability of cognitive neuroimaging data. He joined the Laboratory of Behavioral and Cognitive Neuroscience at Stanford University (http://med.stanford.edu/parvizi-lab.html) in 2017 to pursue these lines of research using intracranial electrophysiological recording and stimulation in humans.

Karl J. Friston is a Professor in the Institute of Neurology at University College London, among many other appointments and honors. His contributions to science include fundamental advances in brain imaging research, mathematical biology, and precision psychology. He has played a critical role in the development of the far-ranging Free Energy Principle. More information about Karl Friston can be found at: http://www.fil.ion.ucl.ac.uk/~karl/

Thomas Metzinger is full professor and director of the theoretical philosophy group and the research group on neuroethics/neurophilosophy at the department of philosophy at the Johannes Gutenberg University of Mainz in Germany. He is also the director of the MIND Group, and co-editor of *Open Mind* and *Philosophy and Predictive Processing*, two online collections of original open access papers. His research focuses on philosophy of mind and philosophy of cognitive science, with a particular interest in phenomenal selfhood. Most recently, he has worked on mind wandering, mental autonomy and applied ethics (especially regarding neurotechnology, virtual reality and artificial intelligence).

David E. Nichols is currently an Adjunct Professor in the Eshelman School of Pharmacy at the University of North Carolina, Chapel Hill, NC. He has carried out teaching and research for more than 38 years prior to his retirement from Purdue University in 2012. Widely published in the scientific literature and internationally recognized for his research on centrally active drugs, he has studied all of the major classes of psychedelic agents. Among scientists, he is recognized as one of the foremost international experts on the medicinal chemistry of psychedelics.

J. Kevin O'Regan is ex-director of the "Laboratoire de Psychologie de la Perception" at the Université René Descartes, Paris 5 (CNRS). His current work, based on his book *Why Red Doesn't Sound Like a Bell: Explaining the Feel of Consciousness* (2011), involves exploring the empirical consequences of his sensorimotor approach to consciousness and "feel". He is particularly interested in the problem of the nature of phenomenal consciousness, which he addresses experimentally and theoretically

in relation to sensory substitution, color perception, infant development and artificial intelligence.

Matthew Ratcliffe did his BA in Philosophy and Psychology at Durham University, UK, between 1991 and 1994. He then studied for his MPhil and PhD degrees at the Department of History and Philosophy of Science, University of Cambridge. His PhD was awarded in October 1999. After being a Lecturer in Philosophy at University College Cork, Ireland, he returned to Durham as Lecturer in October 2002, where he was promoted to Senior Lecturer (2006), Reader (2007), and Professor (2009). In April 2015, he left Durham and joined the University of Vienna, Austria, as Professor for Theoretical Philosophy. After working on sense of reality (2008), depression (2015), he recently published a book on hallucinations (2017a) and is currently studying the nature of grief (2017b).

Ann Taves is a professor of religious studies at the University of California, Santa Barbara. She has published numerous studies on the American religious history and the history of Christianity in the modern era. She has crucially contributed to the interdisciplinary dialogue between religious studies, anthropology of religion and cognitive sciences. Her interdisciplinary endeavor resulted in a groundbreaking book, *Religious experience reconsidered* (2009).

Wandering along the spectrum of spontaneous thinking
Dreaming, meditation, mind-wandering, and well-being

An interview with
Kieran Fox

by Matthieu Koroma

Kieran Fox
kcrfox@stanford.edu
Department of Neurology and Neurological Sciences
Stanford University, USA

Matthieu Koroma
mkoroma@ens-paris-saclay.fr
LSCP, Département d'études cognitives
ENS/EHESS/PSL Research University, Paris, France

Citation: Fox, K. & Koroma, M. (2018). Wandering along the spectrum of spontaneous thinking: dreaming, meditation, mind-wandering, and well-being. An interview with Kieran Fox. *ALIUS Bulletin*, 2, 1-15.

What has triggered your scientific interest in the study of mind-wandering and other associated phenomena?

I really owe my interest in mind-wandering specifically and spontaneous thought more generally to my PhD supervisor, Kalina Christoff. When I started my PhD with Kalina, my interest was in studying the brain basis of meditation and neurofeedback. I thought of mind-wandering the way many other people did at the time: as an annoyance, something that gets in the way of meditation and distracts you from more valuable work. I already had a longstanding interest in sleep and dreaming, but at the time I didn't see all the parallels (both neural and psychological) between dreaming and mind-wandering; exploring those parallels was a big part of my graduate work (Domhoff & Fox, 2015; Fox & Christoff, 2014; Fox, Nijeboer, Solomonova, Domhoff, & Christoff, 2013). Talking with Kalina and reading her prior research showed me how interesting and important mind-wandering really is: how it's related to creativity, dreaming, personality, even mental illness. Our mutual interest in this topic recently led us to edit a book-length treatment of the subject which is due out soon (May 2018 : http://amzn.to/2rLQW9d).

Mind-wandering has been the focus of recent research with a diversity of approaches and definitions. It has been proposed that the notion of mind-wandering could be considered as an umbrella term for a range of phenomena with family-resemblances studied using different approaches, rather than being defined as a

cognitive phenomenon per se. What is the definition or approach of mind-wandering that you endorse in your work?

Our main effort at conceptualizing these states uses *spontaneous thought* as the umbrella term (Christoff, Irving, Fox, Spreng, & Andrews-Hanna, 2016). By "spontaneous" we mean cognition that is relatively unconstrained—either by bottom-up constraints like affective or perceptual salience, or top-down constraints like the executive control of attention. In this framework, mind-wandering falls somewhere in the center of a spectrum. On one end of this spectrum you have totally unconstrained thought; no one really knows what this would look like, but the closest we can imagine is acute psychosis—there is thought and perception and mental content, but it is totally disorganized and nothing is connected, no thought is really related to any other. At the other extreme, you have totally constrained thought, for instance if you were totally focused on writing a difficult exam. Cognition like dreaming, mind-wandering, and creative thinking instead fall in the middle—they are more spontaneous forms of thought than, say, writing a demanding exam, but they are still far more constrained and coherent than, say, psychosis. This is the current thinking of my colleagues and myself, but, of course, these states are difficult to study and define and we hope to make more progress in the future (see Figure 1, reproduced with permission from Christoff, Irving, Fox, Spreng, & Andrews-Hanna, 2016).

> "By 'spontaneous' we mean cognition that is relatively unconstrained—either by bottom-up constraints like affective or perceptual salience, or top-down constraints like the executive control of attention."

Is mind-wandering necessarily a conscious phenomenon or can there be a subconscious form of mind-wandering, with a drifting in the activation of different cognitive networks without being consciously accessed?

My own view, and I think this is amply supported by the empirical evidence, is that mind-wandering can definitely be unconscious—in fact, I think it's safe to say the majority of mind-wandering is below the level of full awareness. For example, in laboratory studies of mind-wandering, when randomly-timed "thought probes" catch people mind-wandering, we can also ask whether people were conscious or "meta-aware" of their mind-wandering. In these studies, meta-awareness of mind-wandering is only reported about half of the time (Christoff, Gordon, Smallwood, Smith, & Schooler, 2009; Fox & Christoff, 2015; Schooler et al., 2011). And this is even during studies where people know they are going to be asked about mind-wandering; even then, they only seem to notice about 50%. Assuming you are

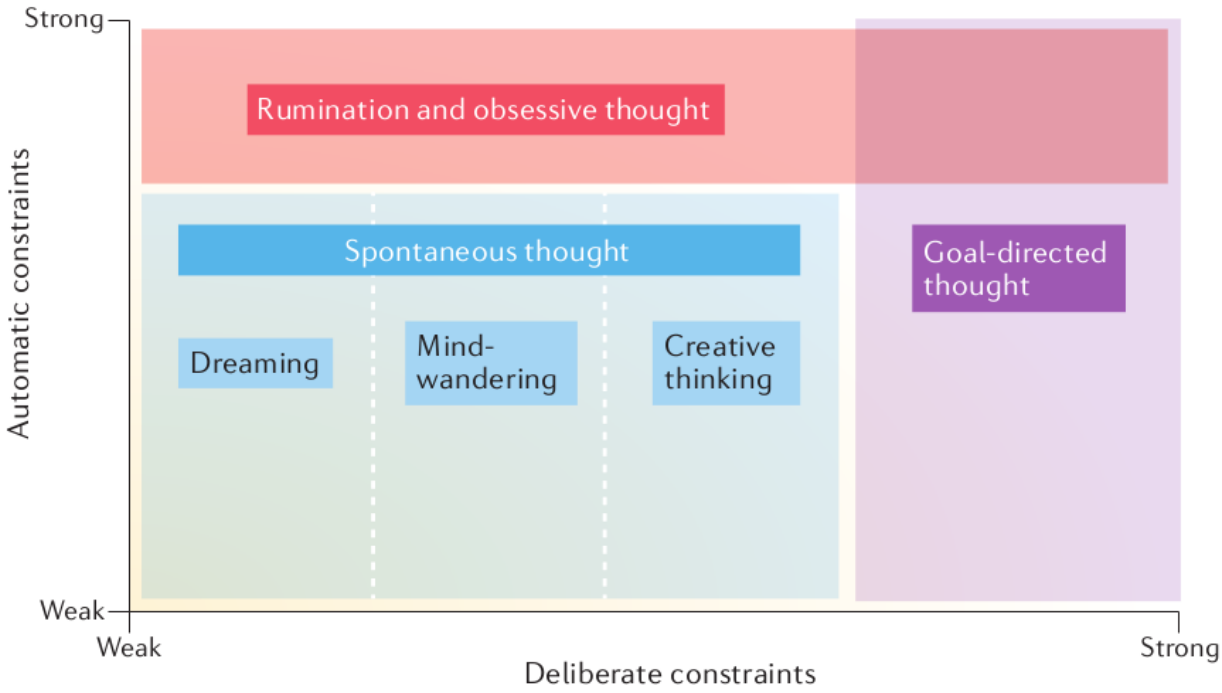


Figure 1 | **Conceptual space relating different types of thought.** Deliberate and automatic constraints serve to limit the contents of thought and how these contents change over time. Deliberate constraints are implemented through cognitive control, whereas automatic constraints can be considered as a family of mechanisms that operate outside of cognitive control, including sensory or affective salience. Generally speaking, deliberate constraints are minimal during dreaming, tend to increase somewhat during mind-wandering, increase further during creative thinking and are strongest during goal-directed thought[39]. There is a range of low-to-medium level of automatic constraints that can occur during dreaming, mind-wandering and creative thinking, but thought ceases to be spontaneous at the strongest levels of automatic constraint, such as during rumination or obsessive thought.

persuaded by the parallels we have drawn between mind-wandering when awake and dreaming while asleep (Domhoff & Fox, 2015; Fox et al., 2013), then dreaming offers an even more compelling case: if you come to a sleep laboratory and are awakened when your brain is in REM sleep, there is an extremely high chance (around 80% or higher) that you will report a detailed, immersive dream experience. We go through four to six REM periods each night, and there is certainly dreaming taking place in other sleep stages as well (Fox & Girn, in press; Nielsen, 2000). Even using a very conservative estimate, we are probably having 10 dream experiences each night at the very least, yet the average person remembers none of this whatsoever, and even people very interested in recording and analyzing their dreams find them difficult to recall. So I think the evidence that detailed and complex spontaneous thought processes can take place without any concurrent awareness or subsequent recall is overwhelming, and in fact this raises some very deep questions about why these phenomena take place at all—why are there rich psychological/subjective correlates of these brain processes that we so seldom notice and whose function (if any) is very difficult to determine?

> “I think the evidence that detailed and complex spontaneous thought processes can take place without any concurrent awareness or subsequent recall is overwhelming.”

If you consider mind-wandering as a conscious phenomenon, would you consider mind-wandering as a proper conscious state (in a sense that mind-wandering characterizes a way of being conscious (Bayne et al., 2016), like for example a meditative state)?

I don’t think of mind-wandering as a conscious state—I think of these processes as more or less ongoing, below the level of awareness, competing with other inputs and signals in the brain for our attention. We can tune in and pay attention to them, or not, and sometimes the thoughts will be strong enough or emotionally salient enough to grab our attention even when we don’t want them to. I think of the stream of inner thought in a way similar to other perceptual channels; for instance, you are constantly receiving a stream of auditory information, even when you’re asleep, but your brain is very good at blocking out probably 99% of this information as totally irrelevant, and you never become conscious of it. But this doesn’t mean that your ears are not receiving the sounds, that the sound is not being transduced and conducted along the auditory nerve, and processed at least at some low level in the brain. I suspect the brain is constantly generating thoughts, imagery, and so on at a “subthreshold” level as well, and noticing it is more a matter of this content catching

our attention and becoming illuminated by our conscious awareness than of entering a particular conscious state where mind-wandering then starts or is allowed to take place.

According to work from you and your colleagues (Fox et al, 2013, Domhoff & Fox, 2015, Christoff et al., 2016, Fox & Girn, in press, Domhoff, 2018), the difference between mind-wandering and dreaming can be seen as a matter of degree, dreams being a more intense form of mind-wandering. This intensification of internally-generated cognition can be temptingly explained by the specific situation of sleep, in which the brain is disconnected from its environment. As such, the disconnection of the dreamer can be seen as a condition for the occurrence and intensity of mind-wandering through the restriction on the relay of sensory information. Conversely, disconnection of the dreamer can be seen as a consequence of the increase in internal activity that competes for attentional resources with externally-oriented networks (Nir & Tononi, 2010). How is regulating the processing of external information occurring during episodes of spontaneous thinking?

I think there is evidence that both explanations are correct. On the one hand, there is a clear dampening and near-blockade of many sensory inputs during sleep, suggesting that internally-generated channels of information have less "competition" for access to conscious awareness. And on the other hand, although brain metabolism tends to decrease in the NREM sleep stages, PET research suggests that the brain's energy usage equals (Braun et al., 1997; Madsen et al., 1991; Maquet et al., 1990) or perhaps even exceeds (Buchsbaum et al., 1989; Heiss, Pawlik, Herholz, Wagner, & Wienhard, 1985) that of waking rest during REM sleep, when we know dreaming is most likely to occur. My colleagues and I have suggested that this has specific effects at both the psychological and neural level. In the brain, the increased activation appears to be preferentially localized to the default, memory, and visual networks (Fox et al., 2013), and we think this can help account for the psychological differences, namely an intensification of visual imagery, lengthy narratives, and so on. In waking spontaneous thought, there is a large body of work, especially from EEG, that mind-wandering is associated with decreased attention to the external environment (Kam, Dao, Stanciulescu, Tildesley, & Handy, 2013; Kam & Handy, 2013). So, to my mind the mechanisms appear quite similar: short, brief disconnections from external sensory inputs while awake can lead to correspondingly brief and relatively mild spontaneous thoughts; and similarly, much more drastic decoupling from the external world during sleep allows for much longer, more intense, and more immersive forms of spontaneous thought to take place.

In a recent article that you co-authored, mind-wandering is characterized within the spontaneous thinking framework as a thought process in which content is weakly constrained by automatic processes and task-related cognitive control (Christoff et

al, 2016). What, then, are the determinants of thought contents during mind-wandering?

That is a very complicated question and the short answer is that we still know very little about what drives and determines the content of mind-wandering. But so-called "thought sampling" or "experience sampling" studies have allowed us to draw some broad conclusions. For instance, we know about the modalities in which mind-wandering tends to occur: these thoughts are very likely to be visual (in the form of imagery), auditory (as in "talking to yourself" or imagining conversations with other people), and somatosensory (thoughts about how your body feels) (reviewed in Fox, Andrews-Hanna, & Christoff, 2016; Fox et al., 2013) These trends appear to hold across different populations in different countries, suggesting that this is a culture-independent neurophysiological process: human brains, in general, don't tend to think in smells or tastes; rather, thoughts take the form of visual imagery and imagined speech (most of the time, anyway). Another very robust finding is that mind-wandering-like thought contains a lot of emotional material; the majority of thoughts have some affective component, and on average they tend to be mildly positive. Contrary to widespread popular opinion, people overall are thinking about things they feel to be neutral or pleasant/positive (Fox et al., under review; Fox, Thompson, Andrews-Hanna, & Christoff, 2014). But probably the most important determinant is what we call "current concerns", meaning what the individual cares about most at any given time, be it major things like a job interview or a sick relative, or more trivial concerns like what to get at the grocery store for tonight's dinner. The evidence is overwhelming that people think about their own personal goals and concerns a huge amount of the time (Klinger, 2008, 2013; Klinger & Cox, 2004, 2011), which suggests that mind-wandering is not nearly as random or pointless as it is often made out to be. Instead, it seems as if there is a very clear functional role, that the brain is frequently working on processing and tackling those things that most concern us, even if we don't yet understand how this "work" is taking place. But I don't think it's a coincidence that insights into, and solutions to, our problems also usually come to us just as spontaneously and unexpectedly.

Mindfulness meditation is a practice based on the careful observation of the train of thoughts through the exercise of meta-awareness. Through adopting "an open, non-judgmental metacognitive stance" (Fox & Christoff, 2014), spontaneous mental activity can be accessed without trying to react to or control the content of the experience. Mindfulness meditation seems thus to offer a privileged situation for introspecting on mind-wandering. Could you elaborate on how such practice can offer insights into the nature and dynamics of spontaneous thinking?

A major issue that many of us worry about in mind-wandering research is, "Are we changing the content and dynamics of spontaneous thought by asking people to

observe and report on it?" The answer must almost certainly be "Yes." William James, in his *Principles of Psychology*, offered an apt simile for this problem, saying that trying to stop the mind at work to observe its functioning was like trying to stop a spinning top to more carefully investigate what its motion is like. By interrupting and analyzing the process, you alter it and lose something essential. This is where meditation broadly and mindfulness more specifically can potentially be helpful. Long-term mindfulness practitioners spend thousands of hours trying to observe their thoughts dispassionately, without reacting or altering things in any way. Whether they are successful or not is an open question and I think a very difficult one to answer, but there is some tentative evidence that long-term practitioners show more accurate and unbiased introspection (Fox et al., 2012; Sze et al., 2010). In a recent study I was involved in (Ellamil et al., 2016), we tried to harness these heightened introspective abilities in long-term practitioners to see if they could tell us about the exact moment when a spontaneous thought was arising in their minds. By using this timestamp as a marker, we could investigate what the brain was doing just *prior* to the conscious awareness of a thought, and try to infer how the brain was generating spontaneous thoughts in the first place. We found that the medial temporal lobe and default network regions were most prominent among the antecedent neural activations, suggesting that these areas play a key generative role. Other brain areas we know to be involved in spontaneous thought, such as prefrontal executive areas, came online a couple of seconds later, suggesting that they play a different role (perhaps in guiding or selecting how these thoughts are interpreted and responded to). This is just one example of how contemplative practitioners can help us answer subtle and tricky questions about spontaneous thought specifically, and human cognition more generally. I think this is a potentially very fruitful field that has only just begun to be explored, but neuroscientists have been calling for this kind of research for decades (Lutz & Thompson, 2003; Varela, 1996).

Since mind-wandering typically consists in a drift of task-focused cognition towards task-unrelated content, it is often accompanied by a drop in performance in the on-going task. Nevertheless, between 30% and 50% of waking thoughts are unrelated to ongoing activities (Klinger & Cox, 1987) and some evidence show that mind-wandering can benefit creativity (Baird et al., 2012). What could be the adaptive value of mind-wandering?

Although it's true that mind-wandering takes place during every conceivable activity (Killingsworth & Gilbert, 2010), the rates of mind-wandering are not identical across all activities. In a typical lab experiment, we deliberately give people a very boring task to do so that we induce high rates of mind-wandering and have some psychological content to study and analyze. So, when we find that people mind-wander a lot during a boring task, or during college lectures for example, we need to remember that in many cases the participant's interest in the task or the

lecture is minimal. For example, in the case of university lectures, perhaps listening to the lecture is relevant to some distant goal of doing well on a final exam and getting a high GPA years from now, but it's easy to see how such distant and relatively vague goals can lose out in the competition with more immediate concerns like interpersonal relationships and conflicts, what to eat, how to schedule your day, and so on. Although I'm not aware of any study that has directly addressed this question, I would say it's a safe bet that motivation and interest in a given task will be strongly inversely correlated with the amount of mind-wandering taking place. A good example is flow states, where people are fully engaged and at an optimal level of difficulty for their given skill level. In the flow state, people hardly report any thinking at all; in fact they often report that any sense of self whatsoever essentially disappears, and they are fully and completely immersed in the activity (Csikszentmihalyi, 2014). But of course much of everyday life isn't like this: we tend to have jobs employing us to do things we are good at, and most things we are good at are by definition overpracticed and hence not fully engaging. And then there are all the daily chores: cooking, cleaning, and so on. These things just need to be done, but they often don't require your full engagement; it's not surprising to me that the brain often disengages from the perceptual world during such activities and instead focuses on other concerns and emotions. I see this as a very clever and resourceful adaptation: if your immediate environment and activities don't require your full attention, your brain very quickly and naturally turns to the recent past or immediate future, or even more distant memories and hypothetical futures, and considers these instead. This seems like it would have clear benefits for everyday life. And if your current goals and concerns happen to be, say, an artistic or scientific problem you are working on, your brain will turn to these concerns—I see the relationship with creativity as just a special case of the more general phenomenon of an automatic, natural focus on current concerns whenever there are cognitive resources available. If you're an artist, your concern will be your next creation; if you're a new parent, your newborn child. In both cases, your mind will tend to focus on your central concerns whenever it has the chance.

> "I see the relationship with creativity as just a special case of the more general phenomenon of an automatic, natural focus on current concerns whenever there are cognitive resources available."

How is the study of mind-wandering informative regarding clinical conditions like ADHD or for mental well-being in general?

Well, if you accept the hypothesis I've advanced above, that spontaneous thought and mind-wandering are a natural, healthy, and probably useful function of the human mind, then you can then conceive numerous mental health conditions as

dysfunctions of this natural propensity toward spontaneous thinking detached from the here and now. Various clinical conditions will distort or exaggerate given aspects of spontaneous thinking, leading to pathological states and experiences that cause significant distress and life disruption for the individual. For instance, we often spontaneously think about the past, and there doesn't seem to be anything wrong with that; it's nice to remember pleasant experiences, and it's probably useful to recall bad experiences in order to process what can be learned from them and how they might be avoided in the future. But in, say, post-traumatic stress disorder, this tendency to spontaneously recall past experiences becomes extremely distressing and overwhelming. It becomes *intrusive*, in that the memories can't be ignored and instead completely dominate one's present experience and disrupt one's ability to carry on with normal life; and they become *repetitive*, focusing over and over again on the traumatic experience, rather than canvassing a wide range of times and topics in one's life, as happens with spontaneous memory recall in healthy people. We can look at depressive rumination in a similar way: rather than focusing on a particular traumatic memory, however, one engages repetitively in a series of negative thoughts about the self, one's past failings, and one's dismal prospects for the future. The normal tendency for the mind to spontaneously have thoughts with a wide range of emotional valence that errs on the positive side (Fox et al., under review; Fox et al., 2014) instead becomes intensely focused only on the negative side of things; the normal tendency to think about the self and its relationship to others becomes distorted to emphasize only the worst possible features and outcomes. On the other hand, some people can become "biased" toward emotionally positive and creative thoughts; for instance, there is evidence that practicing meditation and being mindful can nudge spontaneous thought in this direction (Brown & Ryan, 2003; Fox et al., under review; Frewen et al., 2008; Jazaieri et al., 2015). So on the one hand, I think we need to see overall thought tendencies and content as a very individual, almost trait-like quality, because people differ enormously in what they think about and these differences seem quite stable over time (Fox, 2016); but on the other hand, these baseline tendencies are clearly malleable: they can be skewed toward the negative by various mental health conditions, and perhaps they can be pushed in more positive directions by practices like mindfulness meditation. Of course all these factors are very relevant to general well-being. My colleagues and I (Andrews-Hanna et al., 2013; Christoff et al., 2016), as well as many other researchers, are continuing to work on these problems in an effort to better understand them, but this is a tough undertaking. We still understand very little about what spontaneous thought is, what causes mental illness, and what conduces to general well-being, so understanding how all these things are interrelated is necessarily a long-term project that is still in its infancy.

Because mind-wandering represents a spontaneous cognitive phenomenon unrelated to task-demands, it is challenging to probe and characterize experimentally. What has been the benefit of neuroimaging for the study of mind-wandering and how has it allowed us to derive specific hypotheses about what mind-wandering consists of?

Well, the clearest and most foundational finding is that mind-wandering is definitely tied to recruitment of the default network (Fox et al., 2015), which has helped us understand what is going on (psychologically and neurally) in the "resting" state, and forced neuroimagers to reconsider what state they use as a baseline and comparison state for other cognitive tasks. One of the most important insights from neuroimaging, in my view, is that spontaneous thought recruits "executive" brain regions (Christoff et al., 2009; Fox et al., 2015) that are clearly involved in top-down control of attention, metacognition, and so on, and are often thought of as sort of the opposite of what you would expect during mind-wandering. But this finding actually dovetails well with the fact that the wandering mind tends to focus on goals, concerns, future plans, and so on (as we discussed above). This is a nice example of the neuroimaging evidence supporting what has been found from first-person reports and introspective assessment; subjective reports have many limitations and potential biases, so it is encouraging when what people tell us about these private, unverifiable experiences is in fact supported by what's going on in their brains. Another important finding is the similarities between the neural correlates of dreaming and waking mind-wandering: again, first-person reports suggest a lot of similarities, but this doesn't prove that both processes are sharing neural mechanisms. But the neuroimaging evidence, and even evidence from brain lesion patients, strongly suggests that this is indeed the case (Domhoff, 2011; Domhoff & Fox, 2015; Fox et al., 2016; Fox et al., 2013). As we touched on earlier, such a "continuous" view of spontaneous thought has important implications for understanding different states of consciousness as well as the origins of mental health conditions that involve dysfunctions in spontaneous thinking.

So, neuroimaging has already contributed to our understanding in many ways even though there have been very few studies to date. But this is changing rapidly: spontaneous thought is becoming a popular and acceptable topic, and I expect our understanding to advance by leaps and bounds as more, and more sophisticated, neuroimaging studies of spontaneous thought are conducted. We are already starting to go beyond these broad correlations and correspondences and starting to test more specific hypotheses (Kucyi et al., 2013), to investigate the neural correlates of particular types of thought content (Gorgolewski et al., 2014), and to understand the neural basis of how emotions color spontaneous thoughts (Tusche et al., 2014).

Where do you see the study of mind-wandering evolving in the coming years?

A major, long-term project will be understanding its relationship(s) to mental illness and general well-being, as we already discussed. A strong personal interest of mine is investigating just how stable people's patterns of thought are, and how these relate to personality and creativity. In contrast, how malleable are spontaneous thought patterns? Can we steer people away from the negative biases that we see in mental illness, and instead nudge them toward positive, constructive, and creative patterns of thinking?

References

Andrews-Hanna, J. R., Kaiser, R. H., Turner, A. E. J., Reineberg, A., Godinez, D., Dimidjian, S., & Banich, M. (2013). A penny for your thoughts: Dimensions of thought content and relationships with individual differences in emotional well-being. *Frontiers in Perception Science*.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: mind wandering facilitates creative incubation. *Psychological Science*, 23(10), 1117-1122.

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are there levels of consciousness?. *Trends in cognitive sciences*, 20(6), 405-413.

Braun, A. R., Balkin, T., Wesensten, N., Carson, R., Varga, M., Baldwin, P., . . . Herscovitch, P. (1997). Regional cerebral blood flow throughout the sleep-wake cycle. *Brain, 120*(7), 1173-1197.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology, 84*(4), 822.

Buchsbaum, M. S., Gillin, J. C., Wu, J., Hazlett, E., Sicotte, N., Dupont, R. M., & Bunney, W. E. (1989). Regional cerebral glucose metabolic rate in human sleep assessed by positron emission tomography. *Life sciences, 45*(15), 1349-1356.

Christoff, K., Gordon, A. M.,Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106*(21), 8719-8724.

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews Neuroscience, 17*, 718-731.

Csikszentmihalyi, M. (2014). Toward a psychology of optimal experience *Flow and the foundations of positive psychology* (pp. 209-226): Springer.

Domhoff, G. W. (2011). The neural substrate for dreaming: is it a subsystem of the default network? *Conscious Cogn, 20*(4), 1163-1174. doi: 10.1016/j.concog.2011.03.001

Domhoff, G. W. (2017). *The emergence of dreaming: Mind-wandering, embodied simulation, and the default network*. Oxford University Press.

Domhoff, G. W., & Fox, K. C. R. (2015). Dreaming and the default network: A review, synthesis, and counterintuitive research proposal. *Conscious Cognition, 33*, 342-353.

Ellamil, M., Fox, K. C. R., Dixon, M. L., Pritchard, S., Todd, R. M., Thompson, E., & Christoff, K. (2016). Dynamics of neural recruitment surrounding the spontaneous

arising of thoughts in experienced mindfulness practitioners. *Neuroimage, 136*, 186-196.

Fox, K. C. R. (2016). *Functional neuroanatomy of self-generated thought: investigating general brain recruitment, specific neural correlates, and neural origins using functional magnetic resonance imaging and diffusion tensor imaging.* (PhD), University of British Columbia, Vancouver, Canada.

Fox, K. C. R., Andrews-Hanna, J. R., C., M., Dixon, M. L., Markovic, J., Thompson, E., & Christoff, K. (under review). Affective neuroscience of undirected thought. *Annals of the New York Academy of Sciences, Special Issue: The Year in Cognitive Neuroscience.*

Fox, K. C. R., Andrews-Hanna, J. R., & Christoff, K. (2016). The neurobiology of self-generated thought from cells to systems: Integrating evidence from lesion studies, human intracranial electrophysiology, neurochemistry, and neuroendocrinology. *Neuroscience, 335*, 134-150.

Fox, K. C. R., & Christoff, K. (2014). Metacognitive Facilitation of Spontaneous Thought Processes: When Metacognition Helps the Wandering Mind Find Its Way. *The Cognitive Neuroscience of Metacognition* (pp. 293-319): Springer.

Fox, K. C. R., & Christoff, K. (2015). Transcranial direct current stimulation to lateral prefrontal cortex could increase meta-awareness of mind wandering. *Proceedings of the National Academy of Sciences, 112*(19), E2414

Fox, K. C. R., & Girn, M. (in press). Neural correlates of self-generated imagery and cognition throughout the sleep cycle. In K. C. R. Fox & K. Christoff (Eds.), *The Oxford Handbook of Spontaneous Thought*. New York: Oxford University Press.

Fox, K. C. R., Nijeboer, S., Solomonova, E., Domhoff, G. W., & Christoff, K. (2013). Dreaming as mind wandering: Evidence from functional neuroimaging and first-person content reports. *Front Hum Neurosci, 7*, 412.

Fox, K. C. R., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christoff, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage, 111*, 611-621.

Fox, K. C. R., Thompson, E., Andrews-Hanna, J. R., & Christoff, K. (2014). Is thinking really aversive? A commentary on Wilson et al.'s "Just think: The challenges of the disengaged mind". *Frontiers in Psychology: Cognition.*

Fox, K. C. R., Zakarauskas, P., Dixon, M. L., Ellamil, M., Thompson, E., & Christoff, K. (2012). Meditation experience predicts introspective accuracy. *PLoS ONE, 7*(9), e45370.

Frewen, P. A., Evans, E. M., Maraj, N., Dozois, D. J., & Partridge, K. (2008). Letting go: Mindfulness and negative automatic thinking. *Cognitive therapy and research, 32*(6),

758-774.

Gorgolewski, K. J., Lurie, D., Urchs, S., Kipping, J. A., Craddock, R. C., Milham, M. P., . . . Smallwood, J. (2014). A correspondence between individual differences in the brain's intrinsic functional architecture and the content and form of self-generated thoughts. *PLoS ONE, 9*(5), e97176.

Heiss, W.-D., Pawlik, G., Herholz, K., Wagner, R., & Wienhard, K. (1985). Regional cerebral glucose metabolism in man during wakefulness, sleep, and dreaming. *Brain Research, 327*(1), 362-366.

Jazaieri, H., Lee, I. A., McGonigal, K., Jinpa, T., Doty, J. R., Gross, J. J., & Goldin, P. R. (2015). A wandering mind is a less caring mind: Daily experience sampling during compassion meditation training. *The Journal of Positive Psychology*(ahead-of-print), 1-14.

Kam, J. W., Dao, E., Stanciulescu, M., Tildesley, H., & Handy, T. C. (2013). Mind wandering and the adaptive control of attentional resources. *Journal of Cognitive Neuroscience, 25*(6), 952-960.

Kam, J. W., & Handy, T. C. (2013). The neurocognitive consequences of the wandering mind: a mechanistic account of sensory-motor decoupling. *Frontiers in Psychology, 4*.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science, 330*, 932.

Klinger, E. (2008). Daydreaming and Fantasizing: Thought Flow and Motivation. In K. D. Markman, W. M. P. Klein & J. A. Suhr (Eds.), *Handbook of Imagination and Mental Simulation* (pp. 225-239). New York: Psychology Press.

Klinger, E. (2013). Goal Commitments and the content of thoughts and dreams: basic principles. *Frontiers in Psychology, 4*.

Klinger, E. & Cox, W. M. (1987). Dimensions of thought flow in everyday life. *Imagination, Cognition and Personality, 7*: 105—128

Klinger, E., & Cox, W. M. (2004). Motivation and the theory of current concerns. *Handbook of motivational counseling: Concepts, approaches, and assessment*, 3-27.

Klinger, E., & Cox, W. M. (2011). Motivation and the goal theory of current concerns. *Handbook of motivational counseling: Goal-based approaches to assessment and intervention with addiction and other problems*, 1-47.

Kucyi, A., Salomons, T. V., & Davis, K. D. (2013). Mind wandering away from pain dynamically engages antinociceptive and default mode brain networks. *Proceedings of the National Academy of Sciences, 110*(46), 18692-18697.

Lutz, A., & Thompson, E. (2003). Neurophenomenology integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of consciousness studies*, *10*(9-10), 31-52.

Madsen, P., Schmidt, J., Wildschiodtz, G., Friberg, L., Holm, S., Vorstrup, S., & Lassen, N. (1991). Cerebral O2 metabolism and cerebral blood flow in humans during deep and rapid-eye-movement sleep. *Journal of Applied Physiology*, *70*(6), 2597-2601.

Maquet, P., Dive, D., Salmon, E., Sadzot, B., Franco, G., Poirrier, R., . . . Franck, G. (1990). Cerebral glucose utilization during sleep-wake cycle in man determined by positron emission tomography and [18 F] 2-fluoro-2-deoxy-d-glucose method. *Brain Research*, *513*(1), 136-143.

Nielsen, T. A. (2000). A review of mentation in REM and NREM sleep:"covert" REM sleep as a possible reconciliation of two opposing models. *Behavioral and Brain Sciences*, *23*(06), 851-866.

Nir, Y., & Tononi, G. (2010). Dreaming and the brain: from phenomenology to neurophysiology. *Trends in cognitive sciences*, *14*(2), 88-100.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in cognitive sciences*, *15*(7), 319-326.

Sze, J. A., Gyurak, A., Yuan, J. W., & Levenson, R. W. (2010). Coherence between emotional experience and physiology: Does body awareness training have an impact? *Emotion*, *10*(6), 803-814.

Tusche, A., Smallwood, J., Bernhardt, B. C., & Singer, T. (2014). Classifying the wandering mind: Revealing the affective content of thoughts during task-free rest periods. *Neuroimage*, *97*, 107-116.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of consciousness studies*, *3*(4), 330-349.

Walker, M. P. (2009). The role of sleep in cognition and emotion. *Annals of the New York Academy of Sciences*, *1156*(1), 168-197.

Walker, M. P., Liston, C., Hobson, J. A., & Stickgold, R. (2002). Cognitive flexibility across the sleep-wake cycle: REM-sleep enhancement of anagram problem solving. *Cognitive Brain Research*, *14*(3), 317-324.

Of woodlice and men
A Bayesian account of cognition, life and consciousness

An interview with
Karl Friston

By Martin Fortier & Daniel A. Friedman

Citation: Friston, K., Fortier, M. & Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bulletin*, 2, 17-43.

Karl Friston
k.friston@ucl.ac.uk
Wellcome Centre for Human Neuroimaging
University College London, UK

Martin Fortier
martin.fortier@ens.fr
Institut Jean Nicod
ENS/EHESS, Paris, France

Daniel A. Friedman
dfri@stanford.edu
Department of Biology
Stanford University, USA

You are well known for being the founder of the free energy principle, a wide-ranging theoretical framework aiming to unify the psychological, neural and biological nature of living beings (Friston, 2010, 2013; Ramstead, Badcock, & Friston, 2017). When did you first come up with the idea of the free energy principle? How did this first insight gradually develop to the point of being such a groundbreaking framework?

I first came up with a prototypical free energy principle when I was eight years old, in what I have previously called a "Gerald Durrell" moment (Friston, 2012). I was in the garden, during a gloriously hot 1960s British summer, preoccupied with the antics of some woodlice (small armadillo like bugs—see *Figure 1*) who were frantically scurrying around trying to find some shade.

After half an hour of observation and innocent (childlike) contemplation, I realized their "scurrying" had no purpose or intent: they were simply moving faster in the sun—and slower in the shade. The simplicity of this explanation—for what one could artfully call *biotic self-organization*—appealed to me then and appeals to me now. It is exactly the same principle that underwrites the ensemble density dynamics of the free energy principle—and all its corollaries.

The beautiful simplicity (or nihilistic tautology) of this sort of explanation for life—and creatures like us—crystallized in my teens (for an autobiographical account, see my personal supplementary material in this issue). My father thought it would be a good idea for me to read *Space, Time and Gravitation* by Sir Arthur Eddington (Eddington, 2014). Like my father, I took it to be a compelling essay on the structure

of space-time, in which dynamics and motion are just shapes. The implication was that a sufficient explanation—for nearly everything we see around us—lies in the structured dynamics of their behavior, which is just the "shape of things" in space and time. On this view, the self-organized world "just is" its shape.

Over the subsequent 20 years, I learned enough mathematics to think about these shapes in terms of density dynamics; namely, the evolution of probability density distributions over ensembles of states (e.g., swarms of woodlice). Happily, people had been using exactly this sort of framework both to model the world and analyze their data. I came to know this as *ensemble learning* and, in particular, *variational Bayes*. This is how the free energy principle developed into the current framework. In brief, I was very lucky to meet the right people—and work in an era—when these ideas were "in the air".

You may get a sense that the explanations on offer under this framework are rather deflationary; something that Andy Clark refers to as a (Quinean) desert landscape (Clark, 2013a). Personally, I am drawn to that parsimony—always trying to chase those early "aha moments" when I was a young boy—when insight meant that something that looked very complicated was, in fact, very simple. Although, at its heart, the free energy principle is the ultimate deflationary (possibly tautological) account, one can spin-off a number of interesting corollaries, which I am sure you will press me on.

Figure 1: A woodlouse (Oniscidea)
(original image: http://bit.ly/2nhmDT8)

The Bayesian brain hypothesis (e.g., Knill & Richards, 1996), predictive coding (e.g., Clark, 2013a) and the free energy principle (e.g., Friston, 2010) are often equated with one another. You have yourself suggested that these three frameworks are "variations" of the same basic mechanisms (Friston, 2010; Friston, Kilner, & Harrison, 2006).

To be clear, what we call the Bayesian brain hypothesis is the idea that the brain performs inference according to Bayes' theorem, integrating new information in light of existing models of the world. A perceptual or cognitive state can be modeled as being a *posterior probability*, $P(H|D)$, where P stands for "probability", H "hypothesized causes" and D "observed or available data". The posterior probability is the product of the *likelihood*, $P(D|H)$, and the *prior probability*, $P(H)$. In other words, the probability of the model H being true is the likelihood of the model H given the observation D, multiplied by the likelihood of model H relative to other models under consideration.

To make these equations a bit more concrete, let us take the following example: the brain receives scarce data (D) from the retina and has to form a model (H) of how the world has caused this pattern on the retina. In Bayesian terms, the problem to be solved is the following: $P(H|D)$.

The Bayesian brain hypothesis implies that the posterior probability at time 1 (t_1) provides the prior probability at time 2 (t_2):

t_1: $P(H|D)_1 \propto P(D_1|H) \bullet P(H)$
t_2: $P(H|D)_2 \propto P(D_2|H) \bullet P(H|D)_1$

...
t_n: $P(H|D)_n \propto P(D_n|H) \bullet P(H|D)_{n-1}$

This is known as *Bayesian belief updating* and is the underlying principle behind all forms of evidence accumulation such as Bayesian (Kalman) filtering, predictive coding, and other principled schemes for data assimilation.

Considering now the hierarchical structure of the brain, the Bayesian framework implies that the posterior probability of level 1 (l_1) of the cortical hierarchy provides the content of D at level 2 (l_2):

$P(H_1, H_{2,...,} H_n|D) \propto P(D|H_1) \bullet P(H_1|H_2) \ldots P(H_n)$
l_1: $P(H|D)_1 \propto P(D|H_1) \bullet P(H_1|H_2)$
l_2: $P(H|D)_2 \propto P(H|D)_1 \bullet P(H_2|H_3)$

...
l_n: $P(H|D)_n \propto P(H|D)_{n-1} \bullet P(H_n)$

Note, in this construction, the most general hypotheses are divided into a nested

hierarchy of spatially-realized hypotheses, whereas in Bayesian belief updating there is a temporal re-evaluation of one hypothesis. The Bayesian brain model suggests that both these spatial and temporal processes are co-occurring in the animal brain.

With these technical details in mind, we can now define what the free energy principle and the predictive coding framework add to the Bayesian brain hypothesis (cf. Aitchison & Lengyel, 2017). The free energy principle states that the brain aims at reducing *surprise*, where this surprise (or *surprisal*) is quantified as accuracy (expected log likelihood) minus complexity (informational divergence between the posterior probability and prior probability). This complexity is also known as *Bayesian surprise (or salience)*, and represents the extent to which the new data is "surprising" to the prior model. The predictive coding framework depicts the brain as making predictions based on prior hypotheses and then updating these hypotheses by taking into account the difference between predictions and recent data (rather than data as a whole).

Although these three frameworks share many commonalities, they also have striking differences. For example, within the Bayesian framework, all data are taken into account in the likelihood to compute the posterior probability. This is quite different from what happens within the predictive coding framework where only the data which were inconsistent with the prior hypothesis are sent up in the hierarchy in order to update the model of the world. Predictive coding also differs from the Bayesian framework as it implies that prediction comes first and the correction of predictions by data comes at a separate time. By contrast, in Bayesian models the prior probability and the likelihood are computed at the same time to obtain the posterior probability. The free energy principle seems to differ from both the Bayesian brain and predictive coding models as it regards the reduction of informational entropy between hypotheses and sensory data rather than maximization of hypothesis likelihood given sensory data. If the brain is Bayesian, then perceptual and cognitive states are the product of the likelihood and the prior probability, but this is not to say that the difference between the prior probability and the posterior probability tends to be reduced over time. The latter claim is an additional requirement that proponents of the Bayesian brain or predictive coding may not need to make.

Do you agree with this characterization of the Bayesian brain hypothesis, of predictive coding, and of the free energy principle? If so, how do you conceive of the relation between the free energy principle and predictive coding? In your view, does free energy endorse the two central tenets of predictive coding, that predictive top-down processing has a primacy over corrective bottom-up processing and that not all sensory data are sent up into the hierarchy, but only those that were not predicted by top-down processing?

Conversely, what do you consider lacking from the Bayesian brain and predictive

coding models as long as they do not focus on entropy reduction, as the free energy principle? In other words, what is explained by the entropic reduction within the free energy principle that is not explained by any model parameters in the other frameworks?

Do I agree with this characterization the Bayesian brain hypothesis? Yes, I do—with a couple of caveats. I think it is useful to make a fundamental distinction at this point—that we can appeal to later. The distinction is between a *state* and *process* theory; i.e., the difference between a normative *principle* that things may or may not conform to, and a *process theory* or hypothesis about how that principle is realized. Under this distinction, the free energy *principle* stands in stark distinction to things like predictive coding and the Bayesian brain *hypothesis*. This is because the free energy principle is what it is—a principle. Like Hamilton's Principle of Stationary Action, it cannot be falsified. It cannot be disproven. In fact, there's not much you can do with it, unless you ask whether measurable systems conform to the principle. On the other hand, hypotheses that the brain performs some form of Bayesian inference or predictive coding are what they are—hypotheses. These hypotheses may or may not be supported by empirical evidence.

On this view, the relation between the free energy principle and predictive coding is the relationship between a principle and a process theory. Crucially, there are lots of process theories that conform to the free energy principle. Predictive coding is arguably the predominant process theory in cognitive neuroscience; however, there are other contenders (based on discrete as opposed to continuous state space models). These would include things like *belief propagation* and *variational message passing*. These schemes or processes serve as plausible metaphors for neuronal message passing that may or may not have the look and feel of predictive coding. It is important to note that there have been other process theories that have not fared so well in light of empirical evidence; for example, probabilistic population codes and attempts to understand ensemble dynamics in terms of sampling from the posterior; e.g., Gibbs sampling and particle filtering (Beck et al., 2008; Lee & Mumford, 2003).

In short, predictive coding is one of many ways of minimizing variational free energy. It is formally equivalent to Bayesian filtering; e.g., Kalman filtering in engineering (Rao & Ballard, 1999). One aspect of these Bayesian filtering schemes—that speaks to a possible confusion in your question—is that the "predictive" bit of predictive coding is not about anticipation or the future. It is more simply generating predictions of "what is happening now", under my current beliefs or expectations about how my sensations are caused. I am trying to emphasize that there is no alternation between prediction and subsequent correction; everything happens seamlessly over time—with continuous self-adjusting, self-organizing

dynamics which try to keep your expectations flowing in exactly the right direction. If you get this right, it will look as if you are predicting things. In other words, if you can predict the motion of something now, you know where it will be after a short period of time.

The Bayesian brain hypothesis *per se* does not trouble itself to commit to a particular process theory; other than requiring the implicit beliefs to conform to Bayes rule. The Bayesian brain hypothesis is a corollary of the free energy principle and is realized through processes like predictive coding or abductive inference under prior beliefs. However, the Bayesian brain is not the free energy principle, because both the Bayesian brain hypothesis and predictive coding are incomplete theories of how we infer states of affairs.

> “It is this enactive, embodied, extended, embedded, and encultured aspect that is lacking from the Bayesian brain and predictive coding theories; precisely because they do not consider entropy reduction.”

This missing bit is the *enactive* compass of the free energy principle. In other words, the free energy principle is not just about making the best (Bayesian) sense of sensory impressions of what’s “out there”. It tries to understand how we sample the world and author our own sensations. Again, we come back to the woodlice and their scurrying—and an attempt to understand the imperatives behind this apparently purposeful sampling of the world. It is this enactive, embodied, extended, embedded, and encultured aspect that is lacking from the Bayesian brain and predictive coding theories; precisely because they do not consider entropy reduction.

So why have we introduced notions like *entropy production* and *entropic reduction*? Well, entropy is just a measure of the “shape of things”. In this instance the “things” in question are the ensemble densities above (i.e., the relative probabilities of states of affairs). Interesting shapes (i.e., those characteristic of self-organizing systems like you and me) have a low entropy because our sensory states are concentrated in small regions of state space, with large regimes that are sparsely occupied (Schrödinger, 1944). This is exactly the same as the (non-equilibrium steady-state) distribution of woodlice in the shade. Crucially, in the absence of any movement, a low entropy “shaped” probability distribution would simply not exist (Friston, 2013). In other words, had my woodlice just been basking in the sun—making exquisite Bayesian inferences about their inexorable desiccation—there would have been no self-organization (and nothing of note to witness). In short, the free energy principle

fully endorses the Bayesian brain hypothesis—but that's not the story. The only way you can change "the shape of things"—i.e., bound entropy production—is to act on the world. This is what distinguishes the free energy principle from predictive processing. In fact, we have now taken to referring to the free energy principle as *active inference*, which seems closer to the mark and slightly less pretentious for non-mathematicians.

David Marr (1982) famously proposed to distinguish between three levels of analysis: the *computational level* is concerned with identifying the general problem to be solved; the *algorithmic level* is concerned with specifying the rules and representations which can solve the problem; finally, the *implementational level* is concerned with the physical implementation of the algorithmic blueprint. When you speak of the Bayesian brain, of predictive coding, and of the free-energy principle, do you hold these frameworks to accurately describe how the mind/brain works at a computational, algorithmic and/or implementational level?

These three frameworks are often criticized for not being falsifiable and for being exceedingly speculative—especially when they are endorsed at an implementational level. How would you reply to these objections? What evidence do you think we have for each of these frameworks and at each of Marr's levels?

I think the free energy principle ticks all David Marr's boxes. The *computational* level is the normative principle; namely *what* is optimized. For the free energy principle, this is variational free energy, expected surprise, or uncertainty.

The *algorithmic* level depends upon which process theory you want to put forward as a hypothesis. I mentioned a few above; namely, predictive coding, Bayesian filtering, belief propagation, and variational message passing, particle filtering, and so on. The *implementational* level corresponds to a biophysical process theory. This usually entails identifying the biological substrates that perform one of the above algorithmic process theories. In the systems neurosciences, at the moment, the most popular seems to be predictive coding in canonical microcircuits (Bastos et al., 2012; Mumford, 1992; Shipp, 2016). I am continually impressed by how much this particular process theory explains; in terms of neuroanatomy and neurophysiology—at nearly any level you care to specify.

> "I think the free energy principle ticks all David Marr's boxes."

In short, I do "hold that these frameworks accurately describe how the brain and mind works" at all three levels. I have yet to see any empirical evidence that would

seriously question predictive coding as an algorithmic and implementational explanation of early sensory processing. A whole range of predictions and empirical facts can be explained or predicted under this particular process theory. Furthermore, there are many predictions that have yet to be confirmed. One of my favorites is from Stewart Shipp: the prediction—from the computational level—is that there are no principal cells (thought to encode expectations and errors) that pass messages (via axonal bifurcations) up and down cortical hierarchies at the same time.

As opposed to listing all the evidence for predictive coding—in terms of computational architectures and canonical microcircuits—I will amuse myself by deconstructing your question. I would assert that the notion that a "framework" can have the attribute "falsifiable" is a category error. The only thing that can be falsified is a null "hypothesis". In other words, the only way you can falsify something is to reject the null hypothesis in favor of an alternative hypothesis. The notion of falsifiability is thus a very weak notion. It is weak on several fronts. First, and my favorite, is that the hypothesis that "a hypothesis is falsifiable" is itself not falsifiable. This usually keeps people quiet when they ask me whether the free energy principle is falsifiable.

On a more serious note, falsifiable hypotheses are a hangover from classical inference. The better way to frame evidence-based selection of hypotheses is in terms of how much empirical evidence is accrued by competing hypotheses. In this light, you have to ask yourself what are the alternative hypotheses on offer? If one subscribes to the free energy principle there are a number on the table; however, at this stage, there is no serious alternative to predictive coding. One might imagine, in a few years time, contending schemes will be proposed. At that point, we can then evaluate the evidence for competing hypotheses or process theories and proceed in a righteous and Popperian fashion.

Within the predictive framework, cognitive processes and consciousness are conceived as being the result of a computational trade-off between top-down processing (predictions based on the model of the world) and bottom-up processing (prediction errors based on gathered data). Along with other authors, you have emphasized the hierarchical nature of these processes. However, the interaction between the different levels of the hierarchy remains understudied. One important question is that of knowing whether the laws at work at one level of the hierarchy also apply at other levels of the hierarchy.

Some recent studies suggest that there may be crucial differences between these distinct levels. For example, Andrey Chetverikov (2014; 2016) has recently explored the conscious and affective manifestations of prediction errors. A great deal of the ongoing research on the feelings of fluency and disfluency (Unkelback &

Greifeneder, 2013) can be interpreted as exploring the conscious output of subpersonal accurate predictions and subpersonal prediction errors. Fluency refers to the ease of processing information. This ease is experienced every time predictions prove right. On the other hand, disfluency refers to the sense of effort and unease with which information is being processed. Disfluency seems to be typically experienced when predictions prove inaccurate and when prediction errors are being subsequently triggered.

Rephrased at the conscious and affective level, the free energy principle would thus imply that living organisms aim at minimizing disfluency (prediction errors) and maximizing fluency (accurate predictions). Now, this is precisely what some psychologists have disputed. According to Chetverikov, at the experiential level human beings aim at finding a sweet spot between fluency and disfluency rather than minimizing disfluency. For example, it has been shown (Chetverikov & Filippova, 2014) that people's pleasure is maximized not when they are presented with an image easy to process (i.e., a very clear and simple image) nor when they are presented with an image particularly difficult to process (i.e., a fuzzy or very complex image) but when they are presented with an image initially difficult to process and subsequently easy to process as the trick contained in the image is being figured out (i.e., typically, gestalt images that require some effort to be elucidated). To summarize, affective valence seems to be best described as an inverted U shape: fluency is boring (and therefore unpleasant), disfluency is too much effort (and therefore unpleasant), while the right combination of some disfluency and some fluency is a (pleasant) sweet spot that people seem to be seeking in their everyday life.

At the experiential level, this implies that humans are not driven simply by minimization of entropy (i.e., minimization of disfluency) but by the optimal blending of entropy and negentropy (i.e., of disfluency and fluency). Chetverikov and Kristjánsson (2016, pp. 2–3) further remark that this proposal provides us a new solution to the so-called "dark room problem": people do not seek dark rooms—i.e., perfectly fluent environments—because these are too boring; what they rather seek are sweet spots characterized by some fluency (certainty and familiarity) and some disfluency (uncertainty and unfamiliarity).

Do you think that different laws may apply at different levels—e.g., reduction of entropy at the subpersonal levels and a balanced equilibrium between fluency (negentropy) and disfluency (entropy) at the conscious level? Alternatively, do you think that the kind of finding put forward by Chetverikov and colleagues can easily be accommodated by the free energy principle and that minimization of entropy effectively obtains at every level of the hierarchy?

I do not think that "different laws may apply at different levels". I see a singular and simple explanation for all the apparent dialectics above: they are all explained by minimization of expected free energy, expected surprise or uncertainty. I feel slightly

puritanical when deflating some of the (magical) thinking about inverted U curves and "sweet spots". However, things are just simpler than that: there is only one sweet spot; namely, the free energy minimum at the bottom of a U-shaped free energy function.

If you subscribe to the premise that that creatures like you and me act to minimize their expected free energy, then we act to reduce expected surprise or, more simply, *resolve uncertainty*. So what's the first thing that we would do on entering a dark room—we would turn on the lights. Why? Because this action has epistemic affordance; in other words, it resolves uncertainty (expected free energy). This simple argument generalizes to our inferences about (hidden or latent) states of the world—and the contingencies that underwrite those states of affairs.

> "I do not think that 'different laws may apply at different levels'. I see a singular and simple explanation for all the apparent dialectics above: they are all explained by minimization of expected free energy, expected surprise or uncertainty."

This means that any opportunity to resolve uncertainty itself now becomes attractive (literally, in the mathematical sense of a random dynamical attractor) (Friston, 2013). In short, as nicely articulated by (Schmidhuber, 2010), the opportunity to answer "what would happen if I did that" is one of the most important resolvers of uncertainty. Formally, the resolution of uncertainty (aka intrinsic motivation, intrinsic value, epistemic value, the value of information, Bayesian surprise, etc. (Friston et al., 2017)) corresponds to *salience*. Note that in active inference, salience becomes an attribute of an action or policy in relation to the lived world. The mathematical homologue for contingencies (technically, the parameters of a generative model) corresponds to *novelty*. In other words, if there is an action that can reduce uncertainty about the consequences of a particular behavior, it is more likely to be expressed.

Given these imperatives, then the two ends of the inverted U become two extrema on different dimensions. In a world full of novelty and opportunity, we know immediately there is an opportunity to resolve reducible uncertainty and will immediately embark on joyful exploration—joyful because it reduces uncertainty or expected free energy (Joffily & Coricelli, 2013). Conversely, in a completely unpredictable world (i.e., a world with no precise sensory evidence, such as a dark room) there is no opportunity and all uncertainty is irreducible—a joyless world. Boredom is simply the product of explorative behavior; emptying a world of its epistemic value—a barren world in which all epistemic affordance has been

exhausted through information seeking, free energy minimizing action.

Note that I slipped in the word "joyful" above. This brings something interesting to the table; namely, the affective valence of shifts in uncertainty—and how they are evaluated by our brains (please see discussion of precision later). I think most people now regard emotion as associated with the opportunity for (or actual) reduction of uncertainty (or accompanying changes in precision). The implicit selfhood of an emotion is usually tied in to (free energy minimizing) interoceptive inference—and autonomic reflexes. This would take us into another fascinating area about minimal selfhood and embodiment—of the sort that Anil Seth and colleagues would speak to (Seth, 2013).

In short, we expect to be surprised in a world that is predictably unpredictable—and this is the very stuff of free energy minimization.

The previous question naturally leads us to explore the link between computational processes and phenomenological contents. Some authors (Fletcher & Frith, 2009; Ratcliffe, 2013) investigating the mechanisms of schizophrenia within the predictive framework have proposed that the feeling of strangeness that schizophrenics sometimes report could be explained by the abnormally high number of prediction errors triggered in schizophrenics' brains. However, many of the prediction errors described by neurocomputational models of schizophrenia are presumably strictly subpersonal. It thus seems disputable to claim that prediction errors so easily translate into some phenomenological sense of strangeness. Many prediction errors can obviously take place out of the field of consciousness.

What is your take on this question of the mapping of subpersonal processes and phenomenology within the predictive coding framework? Methodologically speaking, how can we decide whether a prediction error will be expressed—and experienced—at the phenomenological level—through a feeling of disfluency or strangeness—or not?

Again, I am forced into the deflationary corner. The explanation for how we decide whether a prediction error will be expressed—and experienced—is simple; particularly in the context of predictive coding. The degree to which a prediction error will be expressed (and experienced) depends upon its precision. This means we also have to predict the precision of prediction errors. This is how we decide whether the prediction error will be expressed. This means that the generative models entailed by cortical and subcortical hierarchies are in the difficult game of predicting not just the *content* of the sensorium but also its *context* in terms of second order statistics; i.e., the precision or confidence that should be afforded prediction errors. There is a large literature on this; ranging from psychological and neurophysiological accounts of attention, through to detailed discussions of sensory

attenuation in terms of attenuating the precision of sensory prediction errors (Clark, 2013b). The common theme here is a focus on how we predict and model precision or uncertainty—and what can go wrong when the underlying neuromodulatory mechanisms are compromised (e.g., Palmer, Seth, & Hohwy, 2015).

This account makes a lot of sense from the point of view of an engineer. Precision is just the Kalman gain; namely, the weight ascribed to prediction errors during online data assimilation or evidence accumulation. Physiologically, it corresponds to the excitability or postsynaptic gain of neuronal populations encoding prediction errors. Psychologically, it is thought to be the predictive coding homologue of attention (Feldman & Friston, 2010). This is potentially important, because it places attention in very close relation to the experience of prediction errors. I notice that you ask about the "phenomenological level". The inferential or sentient phenomenology is straightforward. In terms of a more phenomenological and quantitative experience, I think the story still holds. In other words, some form of attention is necessary to underwrite the access of ascending prediction errors to deeper levels of processing; such that they can revise our beliefs and expectations about states of the world. The key role of precision will figure prominently below; particularly in relation to psychopathology and psychosis.

> "The degree to which a prediction error will be expressed (and experienced) depends upon its precision."

More broadly, this raises the question as to how the mind/brain should be parsed. Psychologists have long considered that two levels were sufficient (e.g., Evans, 2003). More recently, however, some psychologists have advanced that the ontology of the mind/brain should be somewhat ramified (e.g., Shea & Frith, 2016). What do you think is the most parsimonious number of levels that should be distinguished in order to properly model the mind/brain?

When Chris Frith and I are asked this question (which we often are), we answer six. The answer is six. We say this without smiling and wait patiently for the answer to settle in. We may be joking—or we may not. Some of the more principled reasons for assuming that there are six levels to the mind and brain include the following. First, neuroanatomy suggests that there are probably about six levels to the brain's hierarchy. This fits comfortably with the observation that as one moves higher or deeper into the hierarchy, the beliefs entailed by expectations pertain to constructs of greater temporal extent. In turn, this suggests that we are privy to about six orders of magnitude of temporal scale. For example, if the lower bound on predictive coding at the implementational level is about 25 ms (a duty cycle of fast gamma synchronization) then one might imagine the following hierarchy or Kabalistic

taxonomy:

> **Peripheral reflexes**: enacted over a timescale of about 64 ms.
>
> **Transcortical reflexes** (and related phenomena like saccadic eye movements): unfolding on a timescale of the perceptual moment (about 128 ms).
>
> **Percepts** (possibly associated with qualitative experience): unfolding in lower levels of the cortical and subcortical hierarchy – subtending the cognitive moment (about 256 ms).
>
> **Concepts**: corresponding to amodal or domain general expectations – that generate predictions in multiple domain-specific or modality-specific subordinate hierarchical levels. The timescale here now enters the range of 512 ms to seconds; of the sort associated with delay period activity in the prefrontal cortex and elsewhere.
>
> **Narratives**: expectations at levels of the generative model that contextualize sequences of concepts and may unfold over minutes.
>
> **Self-awareness**: appealing to high order constructs that embody a degree of self-modeling by contextualizing lower levels, such as the minimal selfhood necessary for embodied narratives and interactions with the world (including our body that lasts for years).

Note that the timescales here pertain to the things (content items) that are represented not the duration of representations. In other words, we may all have thought "we would live forever"—for a few seconds. It would be interesting to go through and substantiate this partition in terms of the time constants of the underlying neurophysiological processes (Smith, Gosselin, & Schyns, 2006); ranging from fast synchronized neuronal dynamics, through population dynamics, through short-term plasticity and after-hyperpolarization effects, through long-term plasticity right the way through to neuroendocrinology and epigenetic processes (e.g., DNA methylation).

A more philosophical perspective on the above speaks to the notion of self-modeling in a Thomas Metzinger sense (Metzinger, 2003). In other words, by the very construction of hierarchal generative models (implicit in hierarchal predictive coding), there is a statistical separation (known formally as a Markov blanket – see also: http://bit.ly/2BzMxWv) between levels (Clark, 2017). In turn, this means that each level of the hierarchy is in essence trying to perform predictive coding on the basis of evidence from subordinate levels. This separation destroys any phenomenal transparency and lends a form of separation or decomposition that may be consistent with self-inference, the emergence of selfhood, agency, and self-modeling.

If, as the free energy principle states, living organisms aim at minimizing entropy, how should we explain and understand altered states of consciousness involving an abnormally high entropy (Carhart-Harris et al., 2014; Schartner, Carhart-Harris,

Barrett, Seth, & Muthukumaraswamy, 2017), or, on the other hand, an abnormally low entropy (Burioka et al., 2005; Schartner et al., 2015)? Are low entropy altered states more optimal than others? If so, would not this lead us to redefine the criteria of normality and abnormality? Indeed, everyday states of consciousness—which are characterized by some average entropy—would appear to be less optimal than non-ordinary states characterized by low entropy. However, such a claim would be somewhat paradoxical as low entropy states seem closer to death than life!

This question is easy to deal with. As noted above, there is only one imperative; namely, to give existential shape to the way we are. Mathematically, this entails a minimization of entropy (or at least a bound on entropy production). The only interesting states are low entropy states. The only interesting processes are those that bound an increase in entropy. Having said this, the way that we decrease (sensory) entropy can have the look and feel of sensation seeking; through novelty and the resolution of uncertainty (Friston et al., 2017). The apparent paradox here is dissolved by noting that, mathematically, uncertainty is expected free energy. Expected free energy bounds expected surprise and expected surprise is entropy.

Low entropy states are not closer to death. Death is characterized by dissipation, decay and dispersion. It is the ultimate high entropy state—literally, the edge of our existential world, when we are gently absorbed back into the universe.

Some authors have suggested that the predictive coding framework dissolves the classical dichotomy between cognition and perception (Fletcher & Frith, 2009; Lupyan, 2015). Since both perceptual and cognitive states are the results of a trade-off between top-down processing (which can be assimilated to cognition), and bottom-up processing (which can be assimilated to perception), any mental state would consist of the blending of both perceptual and cognitive ingredients. In the same vein, some (Fletcher & Frith, 2009; Hohwy, 2004) have maintained that the predictive coding framework undermines two-factors theories of psychopathologies according to which delusion results from both an abnormal experience and an abnormal cognitive appraisal of this experience (e.g., Davies, Coltheart, Langdon, & Breen, 2001). Such conclusions may appear as a bit hasty, though (see Macpherson, 2017). That exteroception, interoception, proprioception and cognition can all be modeled in terms of a trade-off between top-down predictions and bottom-up prediction errors does not mean that the boundaries between them should be blurred, or that it would be pointless to try to isolate one from the other. As Anil Seth and yourself have proposed (e.g., Hobson & Friston, 2014; Seth, 2015), at a relatively low level, each of these modalities remain largely encapsulated and it is only at the highest levels that intermodal information is integrated. According to this view, for instance, a specialized circuit of predictions and prediction errors would underlie exteroception and another specialized circuit would underlie interoception. It would only be at a relatively high level that the distinction between the two would not be relevant anymore.

What is your take on this issue: do you consider that the predictive framework undermines classical dichotomies between perception and cognition or experience and interpretation, or that it is perfectly compatible with such dichotomies?

Yes, I think this is nicely put. I think that predictive coding undermines these classical dichotomies yet, at the same time, is perfectly compatible with them. As noted above: perception and cognition can be associated with sentient (free energy minimizing) neuronal dynamics, in our hierarchal generative models. On this view, cognition is the process of inference, whereby empirical priors contextualize and predict perceptual content and—at a phenomenal level—possibly qualitative experience. There is nothing magical about this. You entertain hierarchically separable beliefs whenever you perform an analysis of variance that includes both within and between subject effects. In other words, it is perfectly possible to have "beliefs" or expectations about treatment effects in groups and, at the same time, report within subject effects. Both effects depend upon each other, are internally consistent and yet pertain to different levels of description.

Philosophers interested in predictive coding and in the free energy principle have extensively discussed the philosophical implications of these two frameworks. On the one hand, embodied and direct realist philosophers have emphasized the importance of action within the predictive framework. Active inference seems to provide a way of coping with and predicting the world that vindicates philosophies of embodiment and non-representational engagement in the world (Clark, 2017; Downey, 2017; Gallagher & Allen, 2016; Kirchhoff, 2016). On the other hand, representationalist philosophers have insisted that the very structure of Bayesian modeling rules out direct realism—for the brain has only direct access to partial data caused by the world, and no direct access to the world itself, hence the necessity to build a model of the world. By the same token, the very structure of predictive coding modeling rules out direct realism—for the brain has only direct access to bottom-up prediction error data inconsistent with top-down predictions, and not direct access to the world itself, hence the necessity to build a model of world on which future predictions will be based (Hohwy, 2016, 2017). Moreover, the formalization of the free-energy principle in terms of a Markov blanket where the boundary between internal nodes (or states) and external ones plays a key role seems to vindicate the representationalist view. As well as the idea that the structure of internal states (i.e., of the brain) mirrors and recapitulates the causal structure of the world.

In some of your papers (e.g., Allen & Friston, 2016) you endorse the embodied and anti-representationalist view; but in other papers (e.g., Hobson & Friston, 2014), you unequivocally champion the representationalist view. What is your actual position on this heated philosophical issue?

This is an excellent question. My position on this philosophical issue is context

sensitive: I basically agree with the person that I am talking to. In other words, I am quite happy to bat for both sides in the "representation wars" (Williams, 2017). I find these wars most interesting—in terms of the personalities involved, but also from a mathematical perspective.

Exactly the same sort of dialectic emerges in the free energy formulation. In other words, one could take the skeptical position that our Markov blankets provide an evidentiary boundary that separates everything we are and do from stuff "out there" that may or may not exist (Fabry, 2017; Hohwy, 2016). However, for this Markov blanket (evidentiary boundary) to exist there has to be a partition of states into self (internal states) and unself (external states). This forces one into the uncomfortable position that in order for the Markov blanket to exist there must be states "out there". In other words, a radically skeptical free energy minimizing agent only exists in virtue of a mathematical construct that appeals to philosophical realism.

"I am quite happy to bat for both sides in the 'representation wars'. I find these wars most interesting—in terms of the personalities involved, but also from a mathematical perspective."

My favorite way of eluding this dialectic is to either treat the Markov blanket as something that you hide under to preserve a skeptical position (Hohwy, 2016). Alternatively, the Markov blanket can be regarded as an existential interface that keeps as glued to stuff "out there" (Clark, 2017; Hoffman, Singh, & Prakash, 2015). I have wondered whether active inference would dissolve the representation argument. In the sense that a "representation" has semiotic or structural connotations, then I think, again, you can play both sides. Clearly, a posterior belief about the causes of my sensations is, in some sense, representing or "standing in" for a hypothesis that explains my sensorium. On the other hand, the desert landscape perspective of ensemble dynamics does not call on any representations—it is just in the game of minimizing free energy by destroying free energy gradients (i.e., prediction errors).

One twist to this argument is the fact that the most interesting "shapes of things" are actually generated by the phenotype or agent herself. In other words, when one puts action or movement into the mix, prior beliefs about how I will behave structure the world in a way that does not require a generative process (out there, beyond the Markov blanket) to be isomorphic with the generative model (on the inside). This begs the question: can one represent something that does not exist—before one has authored it?

Many proposals have been made to model the neurocomputational mechanisms of several neuropsychiatric illnesses. The case of psychosis is particularly suggestive. Strikingly enough, the consensus is far from being established as to what these key mechanisms are. Some authors conceive of psychosis as first and foremost resulting from an anomaly of bottom-up processing—of an unusually high triggering of prediction errors mainly due to excessive dopaminergic activity (Smith, Li, Becker, & Kapur, 2006). Conversely, it has been proposed that the anomaly at work in psychosis would lie in top-down rather bottom-up processing: delusions or hallucinations would be caused by overactive priors (Powers, Mathys, & Corlett, 2017; Teufel et al., 2015). Combining the two former views, it has also been advanced that psychosis actually results from a bidirectional anomaly: both bottom-up prediction errors and top-down predictions are at work, because, it is suggested, of the unusual activity of AMPA and NMDA receptors respectively (Corlett, Honey, Krystal, & Fletcher, 2011). A fourth Bayesian model pinpoints the precision—inverse variance—ascribed to bottom-up and top-down processing rather than the content of these processes themselves (Fletcher & Frith, 2009; Friston, Brown, Siemerkus, & Stephan, 2016). According to this model, psychosis is essentially an anomaly concerning synaptic gain: precision weighting—and contextualization—of a given signal. Neuromodulators are thus identified as being crucially involved in psychosis.

With your *disconnection hypothesis* of schizophrenia, you seem to have a preference for the last model of psychosis: the precision anomaly model. Is that the case?

If so, how do you think process-based models and precision-based models can straightforwardly be distinguished from one another? Indeed, these two neurocomputational accounts do not differ from one another in how they regard the output of the mechanisms of psychosis. For example, saying that psychotic patients ascribe an abnormally high precision to prediction errors is equivalent to saying that their prediction error system is overactive. The two accounts differ only as regards their etiological story: in the precision-based account, higher bottom-up processing is mediated by precision weighting, whereas in the process-based account, higher bottom-up processing is malfunctioning. Is there more to the difference between precision-based and process-based models than the etiological story? In other terms, do these two accounts of neuropsychiatric illnesses have also distinct implications at the end of the causal chain?

I am starting to bore myself with the preamble about deflationary answers. However, here it is: there is no distinction between *process*-based and *precision*-based models of psychopathology. If one subscribes to the free energy principle, then you are implicitly subscribing to approximate Bayesian inference. Technically, this rests upon something called a mean field assumption. In turn, this means that the (approximate) Bayesian beliefs about anything depend upon beliefs about everything else. This holds for beliefs about process or content and beliefs about

precision or context.

The implication is you cannot break any sentient or inferential machinery without breaking both process and precision-based inference. Put more simply—in context of predicting process and precision—if you cannot measure something when performing a statistical analysis, you cannot estimate the standard error (i.e., the inverse standard precision). Conversely, if you can't estimate the standard error you can never make an inference. I think this little metaphor is useful because it speaks to false inference as the common denominator behind all current theories of psychopathology and pathophysiology.

> “There is no distinction between *process*-based and *precision*-based models of psychopathology. […] you cannot break any sentient or inferential machinery without breaking both process and precision-based inference.”

False inference here means exactly what it sounds like; namely, type I and type II errors associated with false positives and false negatives. These provide a compelling metaphor for the positive and negative symptoms of many neuropsychiatric disorders. For example, delusions and hallucinations can be regarded as positive symptoms, while things like a resistance to illusions and psychomotor poverty play the role of false negatives (Friston, Brown, Siemerkus, & Stephan, 2016). The question then reduces to what sorts of pathophysiology could result in false inference.

All the available evidence points to a failure of subjective or predicted precision; ranging from psychopharmacology, psychophysics, clinical phenomenology, synaptic neurophysiology, and so on. In short, I do not think there is a canonical distinction between process theories of false inference that can be divided into process-based and precision-based. The more prescient distinction is between the processes that underwrite active inference. I do not know of anybody working in this field who would not, at the end of the day, agree that aberrant precision is the most likely explanation.

If we understand it correctly, the disconnection hypothesis that you embrace states both that schizophrenia is caused by a dysfunction of precision weighting of neuromodulation, and that this dysfunction is mainly mediated by anomalies of the glutamatergic system (especially of NMDA receptors). This might appear a bit surprising: indeed, many researchers seem inclined to think that bottom-up and top-down processes are underlain by glutamatergic and GABAergic activity whereas precision weighting is underlain by the neuromodulatory activity of acetylcholine,

norepinephrine, serotonin and dopamine (e.g., Yu & Dayan, 2005). Do you consider that the difference between process-based and precision-based models can be neurochemically boiled down to a difference between neurotransmission proper and neuromodulation? If so, why does the disconnection hypothesis identify glutamate anomalies as centrally mediating abnormal precision weighting?

These are interesting questions—especially from the perspective of computational psychiatry. In short, my take on these issues is that the computational failure is in terms of precision control or, more generally, the encoding of uncertainty in generative models of the world. In predictive coding, this translates into an abnormal excitability, sensitivity or postsynaptic gain of neuronal populations encoding prediction error. The implication of this aspect of the process theory is that any pathophysiology that affects excitation-inhibition balance or postsynaptic gain becomes etiologically relevant in terms of pathophysiology. These factors range from classical modulatory neurotransmitters, such as dopamine and serotonin, through to ensemble (neuronal) dynamics and the synchronous gain associated with fast neuronal oscillations. This can be characterized in terms of intrinsic connectivity changes or measures of excitation-inhibition balance. The reason that we have focused on NMDA receptors is that they may play a profound role in reporting and structuring the coupling between fast spiking inhibitory interneurons and pyramidal cells—thought to report prediction errors. Generally speaking, it is these fast inhibitory dynamics that set the overall excitability of pyramidal cells and thereby, operationally, encode precision. Crucially, there is abundant evidence to implicate modulatory neurotransmitters—via their effects on NMDA receptor function—in the control of inhibitory dynamics. In short, my suspicion is that all of these phenomena (glutamate neurotransmission, inhibitory neurotransmission, synchronous gain and classical neuromodulators) all have a deeply enmeshed role in the control of precision and the attention paid to—or attenuation of—sensory evidence for our internal models of the world.

Regardless of the distinctions among the free energy, predictive coding, and Bayesian brain frameworks, all these theories agree that anticipation is crucial for skilled action in the world (Bruineberg, Kiverstein, & Rietveld, 2016). One might hypothesize that any sufficiently complex life form would have to anticipate internal and external stimuli, since an improved ability to maintain a physiologically-rewarding state amidst uncertainty is adaptive for all organisms. Indeed, you have advocated for the "predictive processing" framework to include plants (Calvo & Friston, 2017), and others have explored predictive cognition in single-celled life forms (Lyon, 2015) and even ecosystems (Rosen & Kineman, 2005). Though bacteria, plants, animals, and ecosystems certainly use diverse mechanisms to implement predictive models of their environment, it is also true that algorithmically similar processes exist across these systems. On a related note, in a recent *Aeon* article (Friston, 2017) you argued that consciousness in general is a process rather

than a thing. You claimed that beyond simple self-organization (as in a virus), our self-ness is granted by our "temporal thickness", or skill at minimizing surprises in the distant future. For example, saving money while working so that one can have a more comfortable retirement.

So, if our consciousness hinges on the generation or maintenance of accurate long-term models of the world, to what extent do other Free Energy-minimizing systems have genuine introspective capacity or consciousness? For example, if a computer program were able to make "thick temporal models" of its own existence, would it qualify as a "self"? If an ant colony is able to "store its provisions in summer and gather its food at harvest", does this not count as "temporal thickness"? Do super-national predictive organizations such as the UN represent the emergence of a new level of consciousness? How would free energy delineate the arrival of self-awareness in digital and/or decentralized multilevel systems?

The straightforward answer to your question is that—in my world—consciousness is a process and it is the process of inference. Therefore any system that minimizes variational free energy is conscious to a greater or lesser extent (Hobson & Friston, 2014), in virtue of maximizing Bayesian model evidence (the complement of surprise or free energy). In short, self-organization through a process of minimizing self-information is, mathematically, self-evidencing (Hohwy, 2016). Self-evidencing is just active inference and therefore must entail a rudimentary form of consciousness.

Your question is more searching. I take it as asking what is the difference between self-evidencing systems that are aware of themselves—or at least have a minimal selfhood—and those systems (perhaps like an ant colony) that do not. As you rightly note, we have made this distinction on the basis of the counterfactual breath and temporal depth of generative models. In other words, if we are talking about systems that act to minimize free energy, and those systems have been selected (by the process of free energy minimization at an evolutionary timescale) to possess prior beliefs they will minimize free energy, then they must have generative models that include the future. In other words, they must have predictions about the consequences of their action.

The time horizon or depth of these models may be very short or very long. Usually, the deeper the model, the greater the number of policies that can be entertained—and the greater the counterfactual breadth or richness (Seth, 2014). Put another way, counterfactual breath scores the latitude an agent has to select among viable policies that she expects to resolve uncertainty (i.e., reduce the expected surprise of being hungry or ignored). This means that to answer your question about the ant colony one would need to know whether it had (i.e., if it *entailed*) a generative model of counterfactual outcomes. In short, did it make a decision to select one policy (store its provisions) over another (gather its food) in the summer. If one could find

evidence for the encoding of the sufficient statistics of these counterfactual beliefs, in any (biophysical) aspect of the colony, then one would ascribe it a minimal selfhood. In other words, if there was evidence for the capacity to choose (technically, perform Bayesian model selection), then the system would be equipped with a sufficiently rich generative model to qualify as a "self".

This does not necessarily mean that such systems would be aware of themselves. Self-awareness requires something else; namely, a generative model that allows for a distinction between self and other. This may sound like an obvious assertion; however, it becomes quite fundamental in terms of theory of mind, action observation, and the role of things like mirror neurons. In short, the only universes in which I would need to contextualize my predictions—by calling upon inferences about agency—are universes in which the things I see are caused by "creatures like me". In this, and only in this setting, does there become a need to discriminate between self-made acts and the actions one observes others making. Given that most of us populate such worlds, inference about agency and concomitant self-awareness would be an emergent property. In this sense, unless the ant colony spends much of his time engaging with other ant colonies, I suspect the ant colony would not be self-aware. Another example might be a worm. The only worms—on this argument—that can be self-conscious are those whose "soil" comprises a writhing mass of other worms. If one now applies this treatment to computers that are globally connected in our modern world, I am not sure that what the answer would be. I hope that I am around long enough to see what transpires—to resolve my uncertainty.

References

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. Current Opinion in Neurobiology, 46(Supplement C), 219–227. https://doi.org/10.1016/j.conb.2017.08.010

Allen, M., & Friston, K. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 1–24. https://doi.org/10.1007/s11229-016-1288-5

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The Depressed Brain: An Evolutionary Systems Theory. *Trends in Cognitive Sciences*, *21*(3), 182–194. https://doi.org/10.1016/j.tics.2017.01.005

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695-711.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., . . . Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron, 60*(6), 1142-1152. doi: 10.1016/j.neuron.2008.09.021

Burioka, N., Miyata, M., Cornélissen, G., Halberg, F., Takeshima, T., Kaplan, D. T., … Shimizu, E. (2005). Approximate Entropy in the Electroencephalogram During Wake and Sleep. *Clinical EEG and Neuroscience : Official Journal of the EEG and Clinical Neuroscience Society (ENCS)*, *36*(1), 21–24.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28. https://doi.org/10.1007/s11229-016-1239-1

Calvo, P., & Friston, K. (2017). Predicting green: really radical (plant) predictive processing. *Journal of the Royal Society, Interface / the Royal Society*, *14*(131). https://doi.org/10.1098/rsif.2017.0096

Carhart-Harris, R., Leech, R., Hellyer, P., Shanahan, M., Feilding, A., Tagliazucchi, E., … Nutt, D. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, *8*.

Chetverikov, A. (2014). Warmth of familiarity and chill of error: Affective consequences of recognition decisions. *Cognition and Emotion*, *28*(3), 385–415. https://doi.org/10.1080/02699931.2013.833085

Chetverikov, A., & Filippova, M. (2014). How to tell a wife from a hat: affective feedback in perceptual categorization. *Acta Psychologica*, *151*, 206–213. https://doi.org/10.1016/j.actpsy.2014.06.012

Chetverikov, A., & Kristjánsson, Á. (2016). On the joys of perceiving: Affect as feedback

for perceptual predictions. *Acta Psychologica*, *169*, 1–10. https://doi.org/10.1016/j.actpsy.2016.05.005

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–253.

Clark, A. (2013b). The many faces of precision. *Front Psychol.*, *4*, 270.

Clark, A. (2017). How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

Corlett, P., Honey, G., Krystal, J., & Fletcher, P. (2011). Glutamatergic model psychoses: Prediction error, learning, and inference. *Neuropsychopharmacology*, *36*(1), 294–315.

Davies, M., Coltheart, M., Langdon, R., & Breen, N. (2001). Monothematic Delusions: Towards a Two-Factor Account. *Philosophy, Psychiatry, & Psychology*, *8*(2), 133–158. https://doi.org/10.1353/ppp.2001.0007

Downey, A. (2017). Radical Sensorimotor Enactivism & Predictive Processing: Providing a Conceptual Framework for the Scientific Study of Conscious Perception. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

Eddington, A. S. (2014). *Space, time, and gravitation : an outline of the general relativity theory*. [Rockville, Maryland]: Wildside Press.

Evans, J. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Fabry, R. E. (2017). Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology*, *30*(4), 395-414. doi: 10.1080/09515089.2016.1272674

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.

Fletcher, P., & Frith, C. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58.

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

Friston, K. (2012). The history of the future of the Bayesian brain. *Neuroimage*, *62*(2), 1230-1233. doi: 10.1016/j.neuroimage.2011.10.004

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475.

https://doi.org/10.1098/rsif.2013.0475

Friston, K. (2017, May 18). Consciousness is not a thing, but a process of inference. *Aeon Essays*. Retrieved October 19, 2017, from https://aeon.co/essays/consciousness-is-not-a-thing-but-a-process-of-inference

Friston, K., Brown, H. R., Siemerkus, J., & Stephan, K. E. (2016). The dysconnection hypothesis (2016). *Schizophrenia Research*, *176*(2-3), 83–94. https://doi.org/10.1016/j.schres.2016.07.014

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology - Paris*, *100*(1-3), 70–87.

Friston, K., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active Inference, Curiosity and Insight. *Neural Comput*, 1-51. doi: 10.1162/neco_a_00999

Gallagher, S., & Allen, M. (2016). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, 1–22. https://doi.org/10.1007/s11229-016-1269-8

Hobson, A., & Friston, K. (2014). Consciousness, Dreams, and Inference: The Cartesian Theatre Revisited. *Journal of Consciousness Studies*, *21*(1-2), 6–32.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The Interface Theory of Perception. *Psychonomic Bulletin & Review, 22*(6), 1480-1506. doi: 10.3758/s13423-015-0890-8

Hohwy, J. (2004). Top-Down and Bottom-Up in Delusion Formation. *Philosophy, Psychiatry, & Psychology*, *11*(1), 65–70. https://doi.org/10.1353/ppp.2004.0043

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062

Hohwy, J. (2017). How to Entrain Your Evil Demon. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comput Biol, 9*(6), e1003094. doi: 10.1371/journal.pcbi.1003094

Kirchhoff, M. D. (2016). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, 1–22. https://doi.org/10.1007/s11229-016-1100-6

Knill, D., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis, 20*(7), 1434-1448.

Lupyan, G. (2015). Cognitive Penetrability of Perception in the Age of Prediction: Predictive Systems are Penetrable Systems. *Review of Philosophy and Psychology*, *6*(4),

547–569. https://doi.org/10.1007/s13164-015-0253-4

Lyon, P. (2015). The cognitive cell: bacterial behavior reconsidered. *Frontiers in Microbiology*, *6*, 264. https://doi.org/10.3389/fmicb.2015.00264

Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*, *47*(Supplement C), 6–16. https://doi.org/10.1016/j.concog.2016.04.001

Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Metzinger, T. (2003). *Being No One. The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Mumford, D. (1992). On the computational architecture of the neocortex. II. *Biol. Cybern.*, *66*, 241-251.

Palmer, C. J., Seth, A. K., & Hohwy, J. (2015). The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Conscious Cogn*. doi: 10.1016/j.concog.2015.04.007

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600. https://doi.org/10.1126/science.aan3458

Ramstead, M., Badcock, P., & Friston, K. (2017). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2017.09.001

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci.*, *2*(1), 79-87.

Ratcliffe, M. (2013). Delusional atmosphere and the sense of unreality. In G. Stanghellini & T. Fuchs (Eds.), *One Century of Karl Jaspers' General Psychopathology* (pp. 229–244). Oxford: Oxford University Press.

Rosen, J., & Kineman, J. J. (2005). Anticipatory systems and time: a new look at Rosennean complexity. *Systems Research: The Official Journal of the International Federation for Systems Research*, *22*(5), 399–412. https://doi.org/10.1002/sres.715

Schartner, M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K., & Muthukumaraswamy, S. D. (2017). Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Scientific Reports*, *7*, srep46421. https://doi.org/10.1038/srep46421

Schartner, M., Seth, A., Noirhomme, Q., Boly, M., Bruno, M.-A., Laureys, S., & Barrett, A.

(2015). Complexity of Multi-Dimensional Spontaneous EEG Decreases during Propofol Induced General Anaesthesia. *PLOS ONE*, *10*(8), e0133532. https://doi.org/10.1371/journal.pone.0133532

Schmidhuber, J. (2010). Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010). *Ieee Transactions on Autonomous Mental Development, 2*(3), 230-247. doi: 10.1109/tamd.2010.2056368

Schrödinger, E. (1944). What Is Life? : The Physical Aspect of the Living Cell (pp. 1-32). Dublin: Trinity College, Dublin.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn Sci, 17*(11), 565-573. doi: 10.1016/j.tics.2013.09.007

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn Neurosci, 5*(2), 97-118. doi: 10.1080/17588928.2013.877880

Seth, A. (2015). The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies. In T. Metzinger & J. Windt (Eds.), *Open MIND* (pp. 1–24). Frankfurt am Main: MIND Group.

Shea, N., & Frith, C. (2016). Dual-process theories and consciousness: The case for "Type Zero" cognition. *Neuroscience of Consciousness*, *2016*(1), 1–10. https://doi.org/10.1093/nc/niw005

Shipp, S. (2016). Neural Elements for Predictive Coding. *Front Psychol, 7*, 1792. doi: 10.3389/fpsyg.2016.01792

Smith, M. L., Gosselin, F., & Schyns, P. G. (2006). Perceptual moments of conscious visual experience inferred from oscillatory brain activity. *Proceedings of the National Academy of Sciences, 103*(14), 5626-5631. doi: 10.1073/pnas.0508972103

Smith, A., Li, M., Becker, S., & Kapur, S. (2006). Dopamine, prediction error and associative learning: A model-based account. *Network: Computation in Neural Systems*, *17*(1), 61–84. https://doi.org/10.1080/09548980500361624

Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P., … Fletcher, P. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences*, *112*(43), 13401–13406. https://doi.org/10.1073/pnas.1503916112

Unkelback, C., & Greifeneder, R. (Eds.). (2013). *The Experience of Thinking: How the Fluency of Mental Processes Influences Cognition and Behaviour*. London/New York: Psychology Press.

Williams, D. (2017). Predictive Processing and the Representation Wars. *Minds and*

Machines. doi: 10.1007/s11023-017-9441-6

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Am I autistic?

An intellectual autobiography

Karl Friston

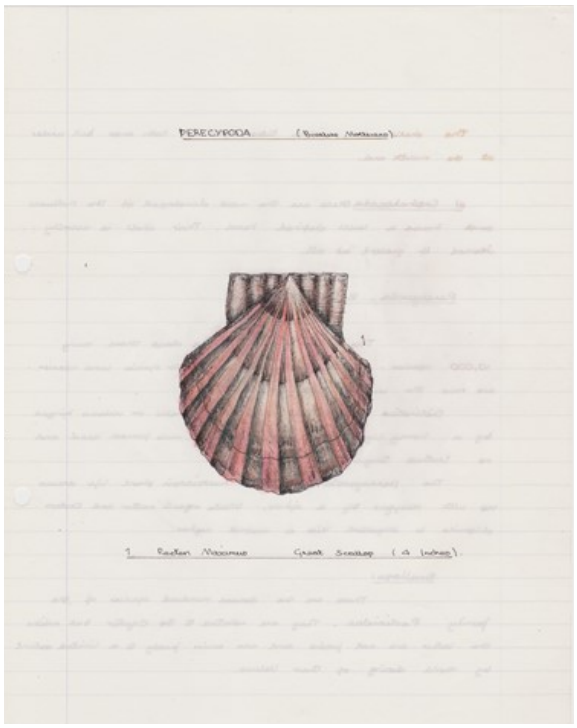
Citation: Friston, K. (2018). Am I autistic? An intellectual autobiography. *ALIUS Bulletin*, 2, 45-52.

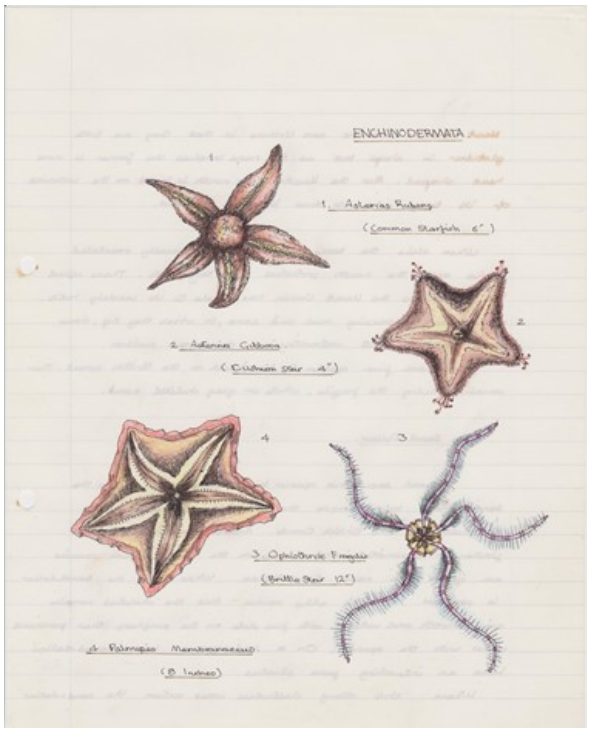
Karl Friston
k.friston@ucl.ac.uk
Wellcome Centre for Human Neuroimaging
University College London, UK

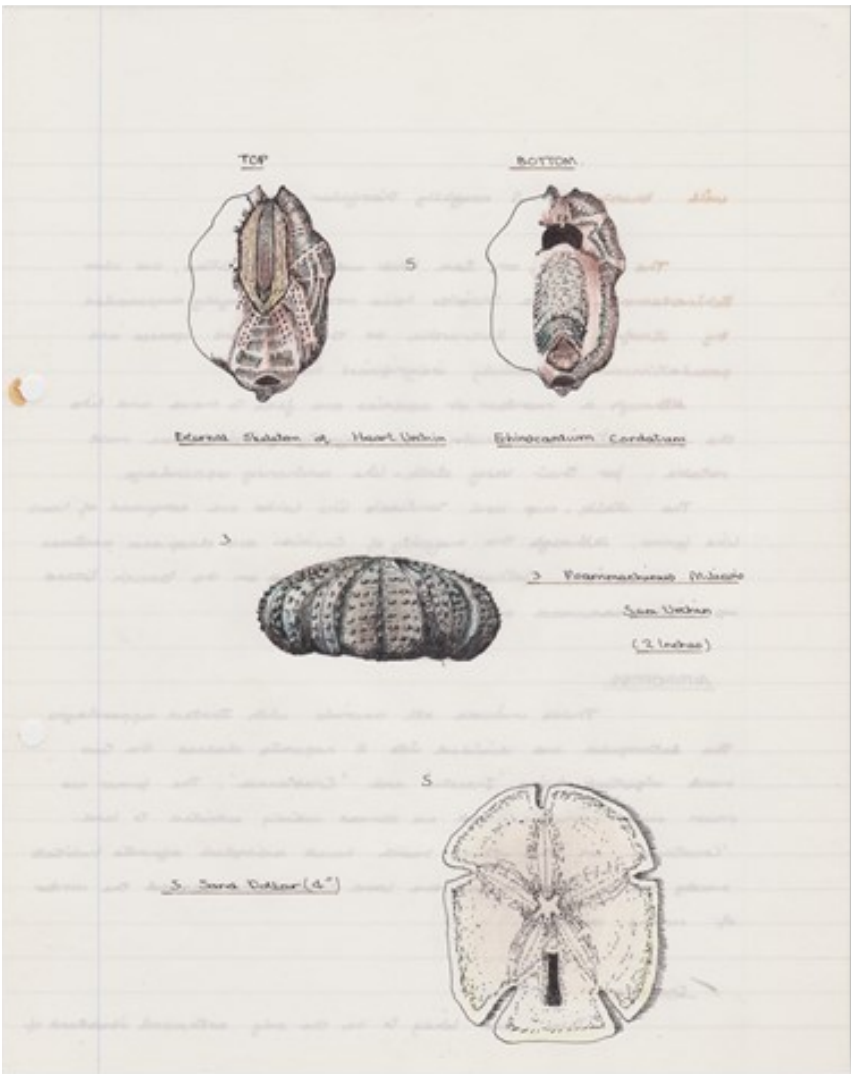
What follows are some personal notes that were inspired by answering the first question for *ALIUS Bulletin*. In looking back at my life, I can see some distinctly autistic traits in my childhood—and indeed current ways of engaging with the world. For example, I religiously avoid mobile phones and do not Skype. In fact, I find any disruption to my weekly routine rather nerve wracking. Unhappily, this means travelling to international conferences can be unsettling—where I spend most of the time avoiding other human beings; especially in the morning. Curiously, I feel most at home with myself when lecturing "onstage"—close to lots of people who are, at the same time, comfortably distant.

When reflecting on my early (academic) experiences, similar themes come to mind. I was obsessed with the natural world and would commit to recording it in a somewhat obsessional fashion: see, for example, the illustrations of aquatic flora and fauna that decorate these notes. I must have spent hours on these for a school project—at the expense of actually learning what I should have been learning.

I am not pretending that I was autistic; however, I remember being assessed by educational psychologists on several occasions. The first (at age 5) was a mildly traumatic experience that was meant to resolve a confusing relationship with my teacher. This educational intervention led to my withdrawal from the state education system and I was sent to a private school run by Catholic nuns (where I flourished). The second was more amusing: I remember being asked whether I thought the puppets in Thunderbirds ever got hungry. I recall thinking at that time "what on earth does a psychologist expect me to say?" After several levels of recursive sophistication, I opted for "yes". The third encounter with a psychologist followed a science project when I was 10 years of age. I had designed a self-righting robot—involving mercury levels and feedback actuators that would enable a little robot table to traverse uneven surfaces (a useful endeavour that set me in good stead to understand the notion of feedback, optimal control, and in later life, cybernetics). The psychologist wanted to know how I came up with the idea. I somehow knew she


PERECYPODA
1


ENCHINODERMATA
1
2
3
4

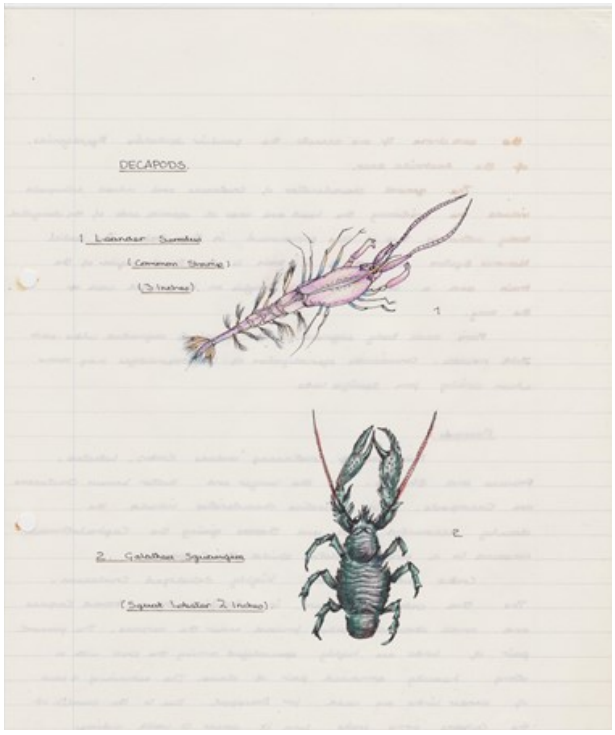

TOP
BOTTOM
5
3
5

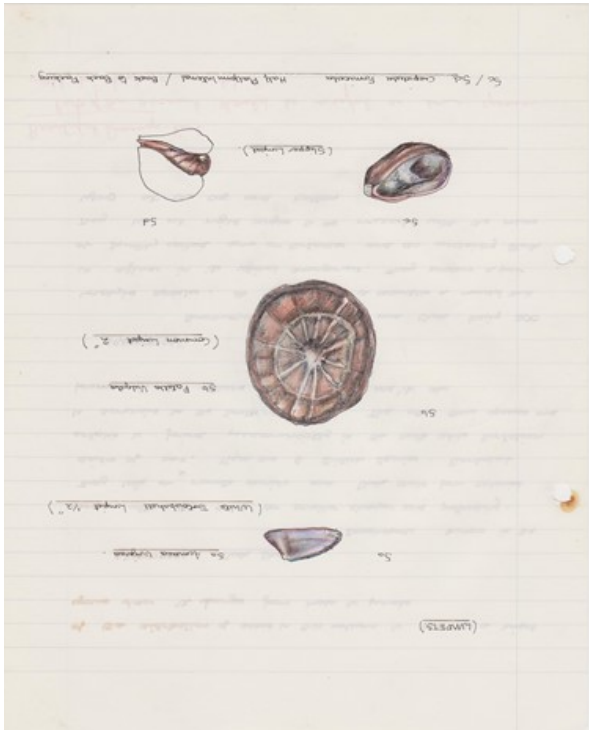
was more interested in me than robots *per se*. Shortly afterwards, something happened, which I want to write down before I forget it: on walking to primary school every day I had to pass the grammar school students waiting for the school bus. I recall thinking: "one day, I want to be in that queue" and then, more poignantly, "I must remind myself about this moment when I am grown—so I do not forget the insight, ambition and sophistication of childhood".

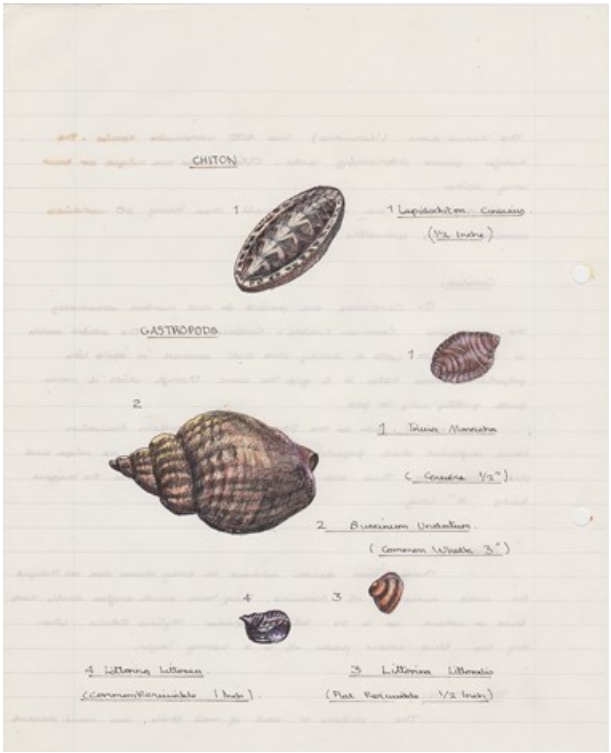
Throughout my education, my primary sources of self-esteem were largely drilling down into an area or problem in a somewhat perfectionist fashion and deriving a delightful sense of comfort in isolating the problem domain—that felt like my little world. My teachers seemed to know this and used to play games with me. One of these games (of which I was most proud) was to see if I could derive answers to A-level questions in mathematics—that were more parsimonious than the worked answers supplied by the examination boards. I recall being obsessed by mathematical issues and enormously pleased by their resolution. One of my favorite achievements was being able to derive Schrödinger wave equation from scratch. My reward was to take the physics class while my physics teacher (Ged Proctor) amused himself in the stockroom (I don't know how, because his primary passion was Morris dancing).

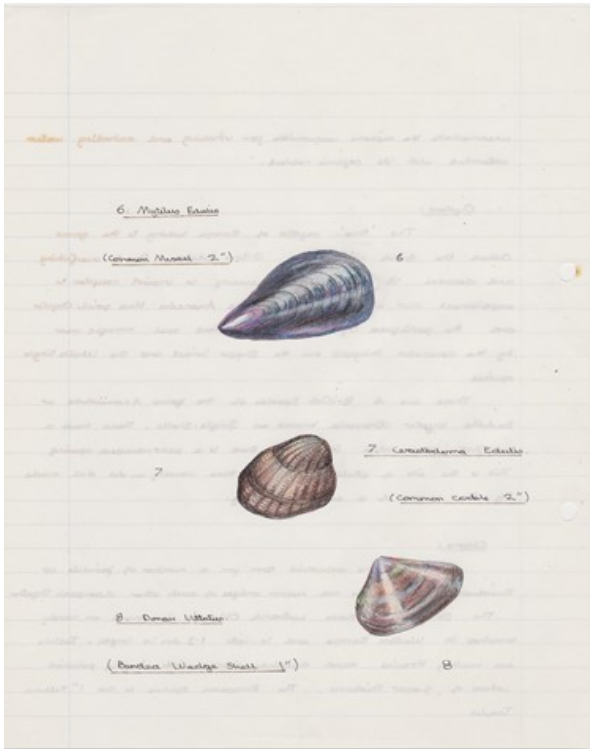
I do not think I was really autistic; however, any unusual traits may have been compounded by my early schooling: my father was a civil (bridge) engineer during an active period of motorway construction in the United Kingdom. This meant that we had to move around the country. By the age of 10, I had attended six schools, and had come to realize that the only constants in my life were my family—and the things inside my head.

This background sets the scene for a series of vignettes that, in retrospect, trace a clear path to the current formulations of the free energy principle. The woodlouse example (see my interview in this issue of *ALIUS Bulletin*) was formative in terms of subsequent exposure to evolutionary thinking. The apparent emergence of purpose from purposeless but "shaped" dynamics made it easy for me to understand natural selection; however, there was a more subtle aspect to the insight that speaks to second order selection. In other words, irrespective of the implicit gradient descent in any optimization process (e.g., natural selection) a simpler mechanism can be in play —without any gradient destroying dynamics. This is simply the fast evacuation of volatile, high-energy regimes of phase-space. In evolutionary thinking, this has often been exemplified in terms of selection for selectability (e.g., the increased mutation rate of *Drosophila* unexposed to a volatile temperature environment). Mathematically, this underwrites generic optimization schemes such as stochastic dynamic optimization. In later life, I often thought about trying to develop this idea in terms of meta-selection—and even ended up using it in the context of active


DECAPODS
1 Leander Serratus
(Common Prawn)
(3 inches)
1
2 Galathea Squamifera
(Squat Lobster 2 inches)
2




CHITON
1
1 Lepidochiton Cinereus
(½ inch)
GASTROPODS
1
2
1 Tritia Monodonta
(Common ½")
2 Buccinum Undatum
(Common Whelk 3")
4
3
4 Littorina Littorea
(Common/Periwinkle 1 inch)
3 Littorina Littoralis
(Flat Periwinkle ½ inch)


6 Mytilus Edulis
(Common Mussel 2")
6
7
7 Cerastoderma Edule
(Common cockle 2")
8 Donax Vittatus
(Banded Wedge Shell 1")
8

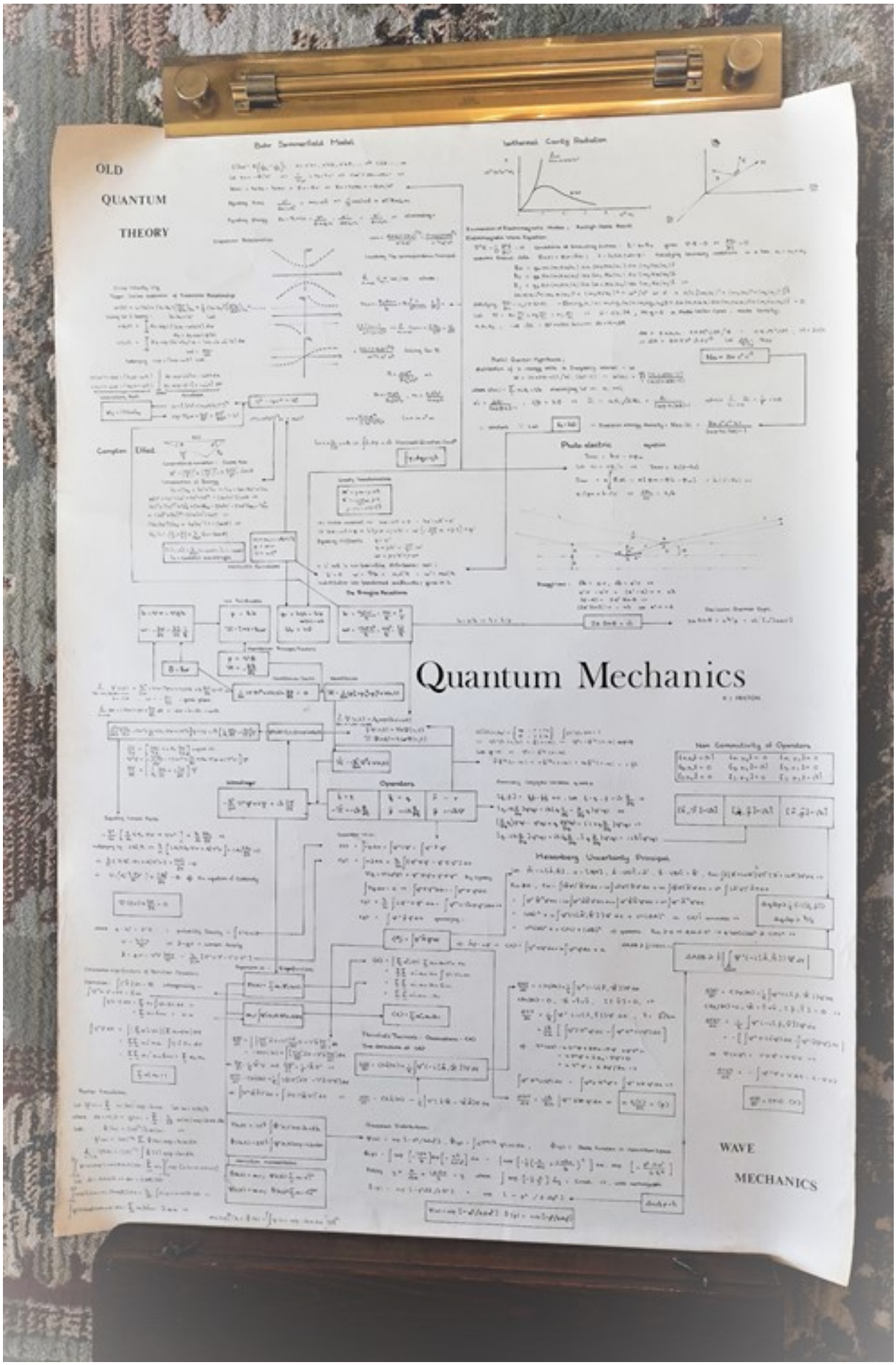
inference; in the form of autovitiation (the destruction of unlikely fixed points by simply moving faster in regimes of high surprise or low probability density).

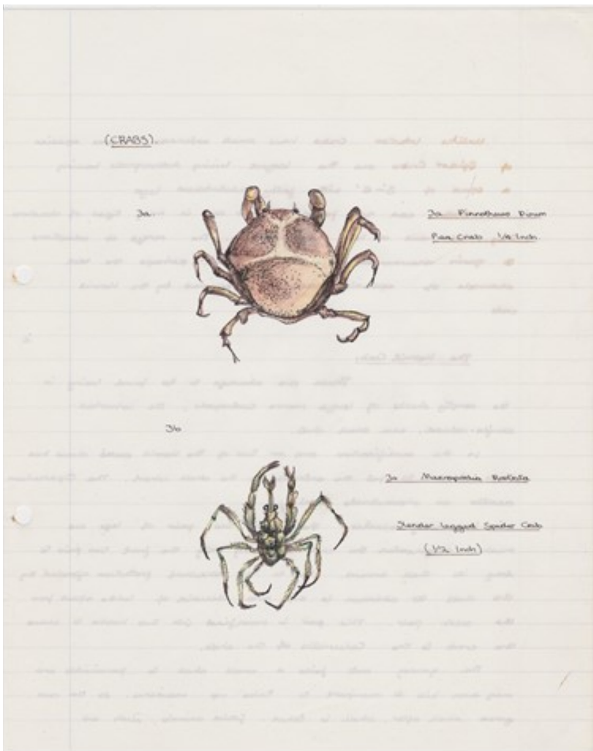
The translation of this sort of thinking into physics started, for me, in my late teens, when I became preoccupied with holistic explanations based on minimal assumptions. I recall spending hours thinking furiously in my bedroom—overlooking cherry blossoms in the front garden: I was convinced that there should be a singular explanation for the shape of things, just starting from the premise that something existed. My best conceptualization of this was some abstract point in an abstract space that, in later life, transpired to be a point attractor in a phase-space. This style of thinking made it easy to understand dynamical systems theory in terms of attracting sets—and the distinction between different forms of attracting manifolds.

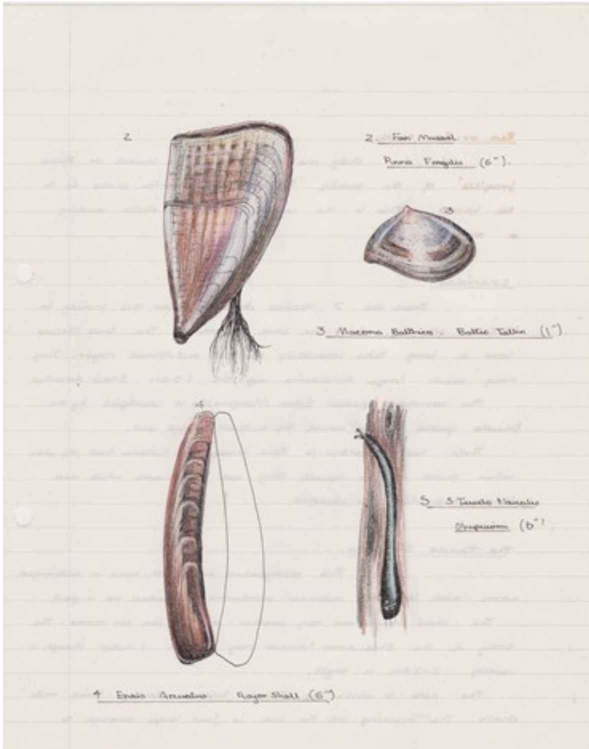
The obsession to put things together came to a practical head in the summer holidays after a year of Medical Science Tripos studies at Cambridge University. I had, with deliberate intent, chosen psychology and physics to pursue for the subsequent years of my undergraduate study. This meant that I had to catch up with the other physics students. I spent an absorbing holiday—to the exclusion of everything else—trying to get all of physics onto one page. I failed—but was able to condense quantum theory into one page (see Figure on the next page). I think that this was symptomatic of an obsessional drive to integration and simplification. Although I forgot nearly everything I had learned during this period, it meant I was not intimidated when taking up these themes in later life—largely by foraging in Wikipedia.

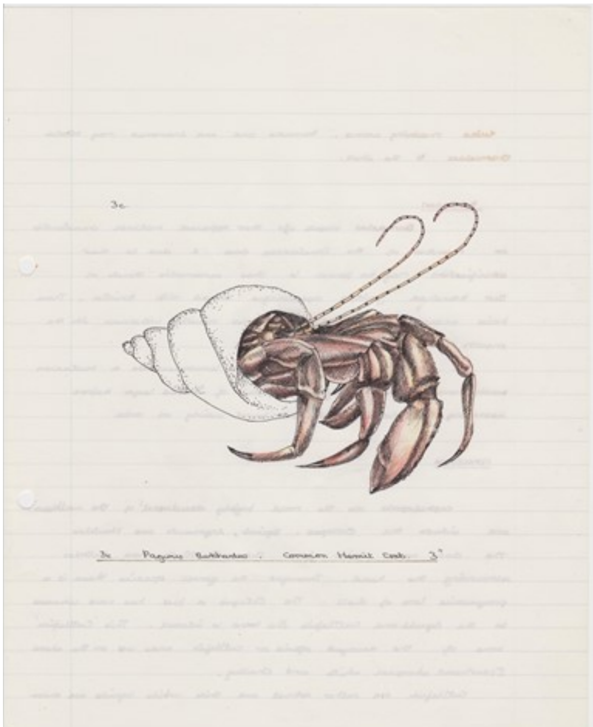
Another memorable episode of intense thinking occurred when on a Christmas break from University, thinking earnestly in the early hours over a nourishing coal fire in the family living room. The conclusion of this contemplation was that all interesting things have to occupy a compact domain of phase-space and must therefore possess an attracting set. The key insight here was that the only invariance that lent "shape to things" entailed correlations. I nurtured this idea for several years (during which I qualified as a doctor and started psychiatric training). I found a peaceful distraction from my job in musing on these issues, while working in a therapeutic community of chronic schizophrenics in an old-style Victorian asylum.

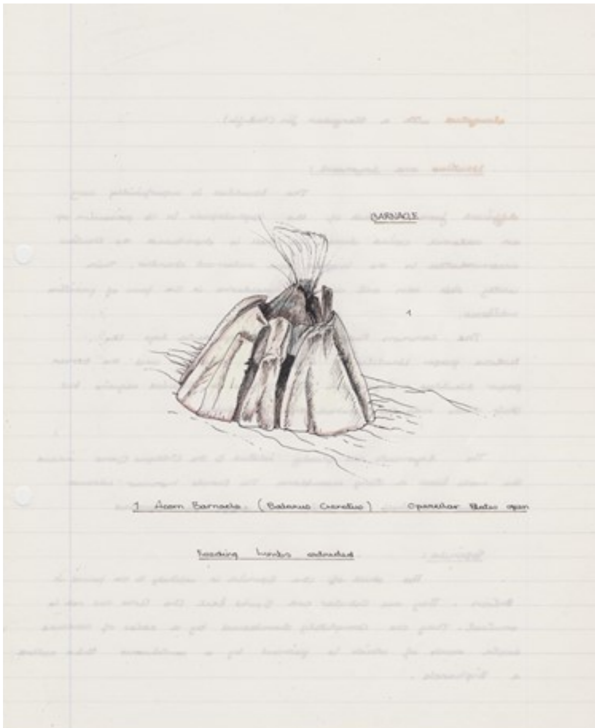
I had, at this point, concluded that statistical invariance (i.e., correlations) had to be transcribed into the physics of our brains—in order for them to possess an attracting set. I found this idea so compelling that I spent an entire Saturday at Blackwell's bookshop in Oxford (where I was training), scouring medical and mathematics books for related ideas (this was before the World Wide Web and Wikipedia). After about three hours searching, I found references to the writings of Hebb and


OLD
QUANTUM
THEORY
Quantum Mechanics
WAVE
MECHANICS


(CRABS).
3a
Pea Crab 1/4 Inch.
3b
Slender legged Spider Crab
(1/2 Inch)


2
2 Fan Mussel
Pinna Fragilis (6").
3
3 Macoma Balthica, Baltic Tellin (1").
4
4 Ensis Arcuatus Razor Shell (6")
5
Shipworm (6")


3c
3c Pagurus Bernhardus Common Hermit Crab 3"


BARNACLE
1 Acorn Barnacle (Balanus Crenatus) Opercular Plates open
Feeding limbs extended

surprised myself with an ambivalent reaction: intense pleasure that the idea was valid and intense displeasure I had wasted several years on something that was already known. I remember trying to work out how old Hebb was—and whether I could have ever met him. From that point on, I waited patiently until I could get into research proper, at around the age of 28. The next part of the story, from my perspective, can be found in (Friston, 2012).

References

Friston, K. (2012). The history of the future of the Bayesian brain. *Neuroimage, 62*(2), 1230-1233. doi: 10.1016/j.neuroimage.2011.10.004

Splendor and misery of self-models

Conceptual and empirical issues regarding consciousness and self-consciousness

An interview with
Thomas Metzinger

By Jakub Limanowski & Raphaël Millière

Thomas Metzinger
metzinge@uni-mainz.de
Department of Philosophy
Johannes Gutenberg-Universität, Mainz, Germany

Jakub Limanowski
j.limanowski@ucl.ac.uk
Wellcome Centre for Human Neuroimaging
University College London, UK

Raphaël Millière
raphael.milliere@philosophy.ox.ac.uk
Faculty of Philosophy
University of Oxford, UK

Citation: Metzinger, T., Limanowski, J. & Millière, R. (2018). Splendor and misery of self-models: conceptual and empirical issues regarding consciousness and self-consciousness. An interview with Thomas Metzinger. *ALIUS Bulletin*, 2, 53-73

Hello Thomas. Are you a self right now?

Hello Raphaël and Jakub! Hmmm…. If I really was "a" self, how would I *know*? Could I ever catch myself, epistemically, in my *substantiality?* And if I wasn't, how could *I* know?

In *Consciousness Explained*, Dan Dennett has famously criticized what he calls Philosophers' Syndrome, which consists of "mistaking a failure of imagination for an insight into necessity" (Dennett, 1991, p. 401). You have yourself expressed similar concerns regarding "armchair" philosophy of mind, and have always favored the analysis of empirical cases over thought experiments. Furthermore, you have dedicated a good deal of attention to pathological or otherwise non-ordinary states of consciousness, from autoscopic phenomena, depersonalization and somatoparaphrenia to dreaming, meditation and full-body illusions—among many others. In your opinion, what place should the discussion of empirical data about so-called altered states of consciousness have in philosophy of mind?

There are many ways in which it can be useful. For example, it changes our theoretical intuitions. Intuitions are phenomenal states that guide our thinking, and they are millions of years old—shaped in the world of our ancestors, determining the attractor landscapes of our brains today, setting priors. They have the form "I *just know* x", without us having any idea or introspective access to the causal history of this knowledge—in intuitive "insight" we suddenly know something, but we really

have no idea where this knowledge comes from. What many don't see is that there is a distinct phenomenology of knowing, and also a phenomenology of certainty (of knowing that one knows). The phenomenal signature of "knowing" is characterized by the phenomenology of direct accessibility of knowledge (which may be preceded by the initial phase of the phenomenology of ambiguity), which sometimes has a character of immediacy. Systematically and rationally investigating altered states of consciousness has the great advantage of exploding many of your theoretical intuitions, and very efficiently. It makes you open-minded. Against the background of a serious interest in the issues and a good academic training it may simply be the most fruitful general heuristic available. If you had a metric to compare the *fecundity* of armchair phenomenology or old-school analytical philosophy of mind to empirically-informed, interdisciplinary philosophy of cognitive science, then what would you think the results over the last three decades would be?

> "Philosophers should design and propose their own experiments."

A frequent epistemological fallacy consists in ascribing epistemic status to phenomenal states because of the phenomenal signature of knowing. Jennifer Windt and myself, in a German paper, called this the "E-fallacy" or "E-error" (Metzinger & Windt, 2014). I would like to point anybody interested in this point to section 3.1 of our introduction to the Open MIND project (http://open-mind.net). Just because something *feels* like an insight doesn't mean that *is* an insight, to put it in an oversimplified way. But a large part of academic philosophy in the last century has consisted in exactly this—"making our intuitions explicit". However, the great danger in cultivating so called "first-person methods" (perhaps as an antidote to intuition-mongering, as exemplified by some forms of old-school armchair philosophizing) is that people have dramatic experiences of deep, ineffable "insights" or subjective "certainty", and then re-iterate the E-fallacy.

There are not only logically possible worlds, but also phenomenally possible worlds: each world that can be simulated on the phenomenal level, relative to a certain class of systems; the possibility of its simulation depending on the functional architecture. However, the functional architecture of our brains has not evolved to help us generate metatheoretical knowledge—therefore we should always be very careful with modal intuitions about what is necessary or possible. So-called "xPhi" or "experimental philosophy" uses statistical estimation of phenomenological reports of others to search the landscape of possible phenomenal worlds in different populations and comparing the intuitions of lay persons to the academic ones, and

it is very stimulating and leads to interesting results. However, the method of interdisciplinary constraint satisfaction (Weisberg, 2006) that I have tried to develop not only uses empirical bottom-up constraints in domain-specific theory-formation, but ideally also shapes the epistemic aim and the process of experiments themselves. Philosophers should design and propose their own experiments! It is more like a methodological experiment of an intuition-free and actively interdisciplinary oriented philosophy of mind.

There is a downside to this: All our results will be preliminary, and highly domain-specific (e.g., only applicable to human minds). Thirty-five years ago, I was fascinated by Hilary Putnam and the project of classical machine functionalism, the "truly philosophical" project of a fully hardware-independent "universal psychology" where we could aim at saying what consciousness, cognition, and so on, really are in *all* possible beings that instantiate them, no matter how they are physically realized (if at all). Now I have a slightly more modest and sober attitude—that is perhaps another downside of looking into the messy details of real-word embodiment with an open mind.

Another, the third, downside is that you become aware of your own psychological vulnerability and your own mortality in a much more acute way, and that many of the relevant recent empirical discoveries are sobering and unattractive on an *emotional* level. Because, as I have come to think, a very strong and mostly unconscious motive for many people to become interested in the philosophy of mind and related areas in the first-place is to discover something that is uplifting, emotionally thrilling and entertaining, causing phenomenal experiences of "meaningfulness", and which helps them develop sustainable psychological strategies for mortality-denial and self-deception, you will face a lot of resistance by the philosophical establishment and parts of the public.

In *Being No One* (Metzinger, 2003), you argue that the folk psychological view of selfhood is thoroughly misguided, insofar as no such things as selves exist in the world. More precisely, you claim that what we traditionally call "the self" is nothing like a mind-independent substance, but a special kind of representational content, namely the content of a sophisticated mental model—a "self-model"—reducible to neurophysiological processes. Yet you maintain that there is such a thing as an experience of selfhood, construed as the dynamic content of the phenomenal self-model, or 'PSM' for short. On this view, the PSM is simply the part of the self-model whose content is phenomenally conscious. This is consistent with your antirealist stance on selfhood insofar as the PSM is a mental process rather than a substance. Moreover, you argue that self-conscious systems such as human beings *identify* with the content of their PSM. Importantly, you claim that the PSM is phenomenally *transparent*, meaning that the various stages underlying this cognitive model are not available for introspective attention. According to you, this transparency constraint

entails that we cannot be aware of our PSM *as a model*: this explains why we have the illusion of being 'selves' in the folk-psychological sense, i.e. substantial, holistic entities. Indeed, having a PSM entails the instantiation of a phenomenal property described as the "primitive, prereflexive feeling of conscious selfhood" (Metzinger 2003, p. 565). However, in a recent talk given at the Sense of Self conference in Oxford (http://senseofselfoxford.wordpress.com), you argued that "if one takes the phenomenology seriously, there really is no such thing as a determinate subjective quality of 'selfhood'" (quoted from the abstract). On the face of it, there seems to be a significant tension between these two claims. Did you revise your hypothesis regarding the existence of a phenomenology of selfhood in PSM-endowed organisms, or does this apparent tension result from a misunderstanding?

The 2017 Oxford talk was entitled "MPS reloaded" and took a critical look back at a paper which I co-authored with Olaf Blanke in *Trends in Cognitive Sciences* in 2009, and which has been cited more than 400 times (Blanke & Metzinger, 2009; see also Metzinger, 2008). One central aim of this paper was to isolate a minimal model of self-consciousness, the phenomenal property of "minimal phenomenal selfhood", which we defined as "transparent spatiotemporal self-location". One major result of the investigation was that the phenomenology of agency is *not* part of MPS, another one was that (in asomatic out-of-body experiences and bodiless dreams) an extensionless point in space can suffice as the locus of identification. We claimed that having MPS is a necessary condition for developing a strong, cognitive or attentional, first-person perspective (1PP), that is of developing what today I would call an "epistemic agent model" (or EAM) (Metzinger, 2013a, 2017a, 2018; see also my new essay on mind-wandering for AEON: http://bit.ly/2DAckUu). We also claimed that spatiotemporal self-location, self-identification (through phenomenal transparency), and a weak 1PP in the purely geometrical sense of an egocentric frame of reference are necessary and sufficient for MPS. Many people seem to have agreed with this general conceptual framework.

I now think that one subtle mistake I may have made is the uncritical assumption that the property called "MPS" is *phenomenally determinate*. In Oxford, I illustrated the problem by dubbing it the "Refrigerator Light Problem": You believe that whenever you close the refrigerator door the lights go out. However, whenever you try to verify your belief and carefully peep into it, the lights automatically go on. In the talk, I discussed first-person methods like classical mindfulness meditation and the status of introspective reports of the type "Whenever I effortlessly come to rest in a clear, emotionally neutral, thoughtless state, I experience MPS". If somebody claims that they introspectively know that MPS is a distinct quality, which can be instantiated in isolation, they face the problem that any attempt at introspective validation automatically creates a much more elaborate phenomenal structure, including an EAM. If you try to find out what the "rock bottom" level of self-

awareness is by willfully directing your attention inwards, then you create a sense of effort and the phenomenal quality of attentional agency. There is no introspective knowledge of MPS as such. What overlooked is that MPS may actually be phenomenally indeterminate.

Phenomenal indeterminacy is $\neg(F(a) \vee \neg F(a))$, i.e., "neither-nor", relative to phenomenal content, as in the sentence "Raphaël neither *instantiates* the phenomenal property called 'MPS' nor he does *not instantiate* the phenomenal property called 'MPS'". This is not the same as $F(a) \wedge \neg F(a)$, i.e., contradiction as in "Raphaël *instantiates* the phenomenal property called 'MPS' and he does *not instantiate* the phenomenal property called 'MPS' *at the same time*" and it also not the same as phenomen*ological* indeterminacy: "Raphaël retrospectively reports that there was no phenomenal fact of the matter regarding minimal phenomenal selfhood" or "Raphaël retrospectively reports that there was a phenomenal fact of the matter, which cannot be expressed in natural language".

I think we must reject introspective authority: we can assume that there is a determinate phenomenal fact of the matter, but at this point in time we do not (scientifically) know it. Blanke & Metzinger (2009) were right, and in the future science may show that "transparent spatio-temporal self-location" is conscious, and *determinate* with regard to the sense of self. However, *individual subjects themselves* are interestingly limited in their access to "minimal selfhood": We find ourselves in a very special epistemic situation with regard to minimal self-consciousness (that was one of my main points in the talk). The instantiation of MPS is an epistemically elusive conscious experience: *a phenomenal fact that is unknown to the subject*. MPS is a 3PP-determinate phenomenal fact, but, currently, epistemically indeterminate. In its minimality, MPS is 1PP-*indeterminable*: we assume that there is a determinate phenomenal fact of the matter, but we are in principle unable to know it $(F(a) \vee \neg F(a))$. Yes, Blanke & Metzinger (2009) were right, and in the future science will show that "phenomenally transparent spatio-temporal self-location" is an objective fact, and *determinate* with regard to the sense of self. But they overlooked that there will always be 1PP-indeterminability: 3PP-knowledge (involving excellent predictive power, etc.) can be had, but this knowledge will never be 1PP-validated, because attentional and/or cognitive agency necessarily activates an EAM.

I have always been interested in how exactly the phenomenal self "bottoms out" (Metzinger, 2014), and also what is the relevant layer in the human self-model that creates the transition from a weak to a strong first-person perspective *above* MPS (namely, the EAM). If you would like a more poetic description, 1PP-indeterminability of MPS can perhaps be read as the "groundlessness" of self-consciousness.

I think there is an additional interesting discovery, which I tried to draw attention to on the excellent meeting you organized. I call it "indeterminacy blindness": Human beings are completely unaware of the fact that they are introspectively blind to truly *fundamental* and philosophically relevant phenomenal facts, namely, the indeterminate *origin* of their very own 1PP. If this is true, then all autophenomenological reports about the "innermost core of the conscious self" are highly dubious and necessarily theory-contaminated. If I am right in my two claims about 1PP-indeterminability and indeterminacy blindness, then I think this is a philosophically interesting feature of human self-consciousness that might warrant further research.

> "I think we must reject introspective authority: we can assume that there is a determinate phenomenal fact of the matter, but at this point in time we do not (scientifically) know it."

A key concept in your Self-Model Theory of Subjectivity (SMT) is the notion of *phenomenal transparency* of conscious mental representations, which means that only the content of such representations is accessible to consciousness—not the fact that they are representations. As mentioned above, you propose that the experience of being a self arises from such a phenomenally transparent part of a system's self-model. This content (i.e., the *phenomenal self-model*) may be used to represent the subject component in a subject-object relationship, while also representing this relationship—what you call a phenomenal model of the intentionality relation. An interesting criticism of SMT was put forth by Josh Weisberg, who worried that the theory "makes too much of the system phenomenal" (Weisberg, 2006). Weisberg instead proposes, very much in the spirit of higher-order theories of consciousness, that to become conscious, the phenomenal self-model needs to be integrated into a nonconscious model of the intentionality relation (NMIR). Has your conception of SMT changed in response to such thoughts?

Of all the critical reviews of BNO, Weisberg's is probably my favorite one—very intelligent, careful and constructive. I have not looked into this issue for a long time, but in a 2003 paper co-authored with Vittorio Gallese and entitled "The emergence of a shared action ontology: Building blocks for a theory", we showed there exist unconscious functional precursors of what can later also be phenomenally represented as a goal, an acting self or an individual first-person perspective (Metzinger & Gallese, 2003). Empirical evidence demonstrates that the brain models movements and action goals in terms of multimodal representations of organism-object-relations and there is empirical evidence for mirror neurons as specifically coding organism-object relations on various levels of abstraction. The motor system

constructs goal-states (successfully terminated actions), action models, and intending selves as basic constituents of the world it interprets by assigning a single, unified causal role to them. I must confess that I have not thought about this a for a long time and have not followed the empirical literature. My intuition is that the PMIR is anchored and dependent on competing, unconscious MIRs, which in turn evolved out of the need to dynamically model whole organism/object-relationships like grasping. My proposal is that first the brain had to model spatial/motor relationships (for example as observed in conspecifics), then it used this basic schema to represent *semantic* relations like "reference" and *epistemic* relations like "attending to a perceptual object" or "grasping an abstract object".

Have you ever thought about the concept of "grasping a *concept*"? It is perhaps the essence of high-level cognition, of human thought itself. It may have to do with simulating hand movements in your mind but in a much more abstract manner. I once looked into this and found out that humankind has apparently known this for centuries, intuitively or because our ancestors had a much more fine-grained introspection than we do today: "Concept" comes from the Latin *conceptum*, meaning the "fruit of the womb" or "a thing conceived," which, just like our modern "to conceive of something," is rooted in the Latin verb *concipere*, "to take in and hold." At this time, the capacity of a woman to successfully "hold the fruit of the womb" was not something self-evident, not something that could be taken for granted, because many more pregnancies failed than today. As early as 1340, a second meaning of the term had appeared: "taking into your mind." If we go back to the original meaning, then infecting other people with memes via philosophical discussion it like trying to make them pregnant with your own ideas—making them "hold" what you take to be your own intellectual fruit in their own brains, by something we like to call "rational argument" forcing them to "take in and hold" what you (perhaps mistakenly) experience as your *own* insights, hopefully later giving birth to something beautiful.

Surprisingly, there is a representation of the human hand in Broca's area, a section of the human brain involved in language processing, speech or sign production, and comprehension. A number of studies have shown that hand/arm gestures and movements of the mouth are linked through a common neural substrate. For example, grasping movements influence pronunciation—and not only when they are executed but also when they are observed. It has also been demonstrated that hand gestures and mouth gestures are directly linked in humans, and the oro-laryngeal movement patterns we create in order to produce speech are a part of this link. By the way, such empirical data are good examples of something that philosopher of language and cognition should know.

Broca's area is also a marker for the development of language in human evolution, so it is intriguing to see that it also contains a motor representation of hand

movements; here may be a part of the bridge that led from the "body semantics" of gestures and the bodily self-model to linguistic semantics, associated with sounds, speech production, and abstract meaning expressed in our cognitive self-model, the thinking self. I think Weisberg was absolutely right when demanding the phenomenological notion of a "model of the intentionality relationship" (which in my recent writings has somewhat morphed into the EAM, or "epistemic agent model") must be grounded in unconscious mechanisms and an evolutionary story. But, again, I must admit that I have not monitored empirical research in this area for a long time.

You have recently edited an open access volumes that are largely drawing on the so-called predictive processing framework and discuss its philosophical implications (Metzinger & Wiese, 2017, http://predictive-mind.net). The predictive processing framework has appealed to philosophers, however, it has been interpreted both in representationalist terms (Hohwy, 2013; Gładziejewski, 2016) or along enactivist ideas (Bruineberg, Kiverstein, & Rietveld, 2016; Gallagher & Allen, 2016). The active inference formulation of predictive processing (Friston, 2009) has been proposed to dissolve this tension: on the one hand, active inference fundamentally assumes inference on representations in hierarchical generative models in the brain—thus appealing to representationalist accounts. On the other hand, active inference is all about reaching the best possible (i.e., least surprising) situation of myself in and as part of my world, and hence representations arise from interaction with the world—thus appealing to enactivist ideas. Since your initial SMT is a purely representationalist account, do you think the active inference (predictive processing) framework may indeed resolve some issues that philosophers have been arguing about for a while now—or do you think this is a too ambitious claim? How much do you think theoretical neuroscience and philosophy can mutually enrich each other?

Oh, they can certainly enrich each other—but it may need a new generation of philosophers who not only know neuroscience and cognitive science, but also mathematics. Personally, I have a very relaxed attitude about the concept of "representation". Many people take me as a realist, and sometimes also assume some caricature concept of "representation", but actually I am more of an instrumentalist. We live and work in a certain period in the history of science, and it is important to never forget that concepts are *historically plastic* entities, they move through time, just like scientific communities do. They are instruments used by communities of epistemic subjects, they serve a purpose for a certain time, eventually you have to throw them away. I have seen a lot of changes from early machine functionalism to the "computer model of mind" and on to connectionist representation (Paul Churchland's *A Neurocomputational Perspective* and Andy Clark's *Microcognition* were important books in my own intellectual biography), dynamicism and EEEE. I think that running neural models described as having properties like "integrated

likelihood" or "model evidence" in Bayesian statistics can still count as representational processes, and that the representational *level of analysis* continues to be very useful and fecund. But that doesn't commit one to realism, it is just a theoretical tool that works for a certain time—maybe we can dissolve it all into measures of entropy or something else soon.

I still remember when, ages ago, Francisco Varela invited Dave Chalmers and me to Paris, and after my talk on self-models he said to me: "I think in principle your whole story is absolutely right, but with that representationalism it is all false and you will *never* get anywhere!" Maybe so. The two of us had more in common than we ever had a chance to explore, that is for sure. But with all the trendy-sexy stuff today, I wonder if he would have liked it. "Enactivism" obviously refers to a relation: Some *A* "enacts" some *B*. Can someone tell us in a non-circular way what that *A* and that *B* actually are?

You have devoted much of your time to ethical problems implied by cultural or technological advances. Recently, you have discussed the potential implications of technological advances in artificial intelligence. One of the appealing features of your SMT is that one can in principle derive empirically testable predictions about when an (artificial) organism or system would experience a first-person perspective and, ultimately, phenomenal selfhood (Blanke & Metzinger, 2009). In light of your recent modifications of your initial proposal—do you think a collaborative effort of philosophers and cognitive and computer scientists could in fact lead to a form of "Turing test" for first-person perspective and experience of selfhood in artificial systems? And, if this was the case, what would your advice to AI developers be—do you think we should be (more) worried by these recent developments?

I have indeed been recently working on ethical issues raised by technological advances such as Virtual Reality (Madary & Metzinger, 2016) and Artificial Intelligence (Metzinger, 2017b). As you may or may not know, I have demanded a moratorium for synthetic phenomenology for quite a number of years now. I think we should always try to minimize the overall amount of suffering in the universe, and recklessly creating artificial consciousness would carry a high risk of *increasing* the overall amount of suffering in the universe. The last time I have done so was in the very short piece I wrote for the EU, asking for the development of Global AI Charter. Here is an excerpt from the forthcoming collection *Should we fear the future of artificial intelligence?* (reproduced here with permission of the STOA Panel of the European Parliament):

A Moratorium on Synthetic Phenomenology

> It is important that all politicians understand the difference between "artificial intelligence" and "artificial consciousness". The unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective,

> because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. "Synthetic phenomenology" (SP; a term coined in analogy to "synthetic biology") refers to the possibility of creating not only general intelligence, but also consciousness or subjective experiences on advanced artificial systems. Future artificial subjects of experience have no representation in the current political process, they have no legal status, and their interests are not represented in any ethics committee. To make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular for experiencing negative states like suffering. One potential risk is to dramatically increase the overall amount of suffering the universe, for example via cascades of copies or the rapid duplication of conscious systems on a vast scale.
>
> **Recommendation 7**
>
> The EU should ban all research that risks or directly aims at the creation of synthetic phenomenology on its territory, and seek international agreements. This includes approaches that aim at a confluence of neuroscience and AI with the specific aim of fostering the development of machine consciousness (for recent examples see Dehaene, Lau & Kouider 2017, Graziano 2017 and Kanai 2017).
>
> **Recommendation 8**
>
> Given the current level of uncertainty and disagreement within the nascent field of machine consciousness, there is a pressing need to promote, fund, and coordinate relevant interdisciplinary research projects (comprising philosophy, neuroscience, and computer science). Specific relevant topics are evidence-based conceptual, neurobiological, and computational models of conscious experience, self-awareness, and suffering.
>
> **Recommendation 9**
>
> On the level of foundational research there is a need to promote, fund, and coordinate systematic research into the applied ethics of non-biological systems capable of conscious experience, self-awareness, and subjectively experienced suffering.

Your question of a test for phenomenality is right on target, because this is what the applied ethics of AI needs. I have tried to isolate the to isolate the four central necessary conditions for suffering in some freely available papers (for example: Metzinger, 2013b, 2016): The C-condition (having a phenomenal model of reality), the PSM-condition (a self-model), the NV-condition (the ability to represent negative valences—for example via homeostatic cost functions folded into the self-model, representations of decreasing functional coherence or low levels of self-control), and the T-condition (transparency, Mother Nature's most evil trick: forcing organisms to *identify* with negatively valenced states). Nobody knows if they are sufficient, and we have no theory of consciousness. From this it follows that there

is an ethics of risk too: We should take great care to always err on the side of caution, and this principle holds for future AI systems as well as for animals. In any case, we should work hard at an evidence-based theory of suffering that is as hardware-independent as possible. We need such a theory, else in the mid-term we will be unable to move forward with AI in an ethically responsible way.

> “Recklessly creating artificial consciousness would carry a high risk of *increasing* the overall amount of suffering in the universe.”

But if you look at what “The First Postbiotic Philosopher” already said in 2009, we could also introduce a *much* stronger criterion for artificial persons who claim to have phenomenal states:

> The Metzinger Test for consciousness in nonbiological systems demands that a system not only claim to possess phenomenal experience and a genuine inward perspective but also comprehend and accept the theoretical problem of subjectivity, and that it demonstrate this by participating in a discussion on artificial consciousness. It has to put forward arguments of its own and convincingly defend its own theory of consciousness. (Metzinger, 2009a, p. 201-202)

In more recent work, you have refined your previous account of MPS, introducing the notion of the *phenomenal unit of identification* or UI for short (Metzinger, 2013c). You define the UI as the relatively invariant phenomenal property (or set of phenomenal properties) with which a given subject self-identifies at a given time, generating “the distinct experience of ‘I *am* this!’”. In ordinary cases, the target properties of self-identification would most likely be “the integrated contents of our current body image”, accompanied by “the subjective quality of ‘agency’ in the control of bodily actions” (p. 5), because we are embodied agents and identify with a body over which we have global control. In bodiless dreams and asomatic OBEs, however, not only do subjects lack the experience of identifying with a body (describing themselves as “pure consciousness”, “balls of light” or “points in space”), but they can also lack agency and motor control. Thus, in some altered states of consciousness, the UI can be something else than the experienced body image, namely either: (a) the experienced origin of the visuospatial perspective as an “extensionless point in space”, which you call the *minimal* UI—the simplest possible unit of identification; or (b) the unity of consciousness, or “awareness as such”, which you call the *maximal* UI—the most general phenomenal property available for identification. The latter, you speculate, might happen in some asomatic OBEs and in deep meditative states in which subjects self-identify with “pure consciousness”. Apparently against your original account, you conclude that MPS is constituted by self-identification with *at least* a minimal UI (and not necessarily with a body), which

merely requires spatiotemporal self-location. Can you explain to us how the introduction of the minimal UI concept has changed the original MPS proposal, and what its potential benefits may be for addressing empirical questions? For instance, one may wonder to what phenomenal property the minimal UI corresponds to in asomatic OBEs and bodiless dreams. Presumably, there is no special phenomenal property of being an "extensionless point in space": a disembodied experience may simply be an experience of a visual scene which lacks any bodily awareness. The assumption that there is an extra feeling of being a disembodied point in space is at least controversial -– subjects might describe their experiences in such a way simply because this is the easiest way to describe an experience of disembodiment. Do you think there is room for a more deflationary take on such experiences?

The original motivation was to describe more clearly *what* exactly it was that was manipulated in those early experiments trying to create full-body illusions. Very often misreported, they do *not* create classical OBEs in the strong sense of involving a perceptually impossible external perspective (Metzinger, 2005a, 2009b). The UI-concept is determined by the following set of empirical constraints:

- Explicit embodiment is not a necessary condition for the UI
- *Minimal* spatiotemporal self-location is a sufficient condition for the UI
- The UI and origin of visuospatial perspective can be dissociated
- The UI can be located outside of phenomenal body model
- The UI and the origin of visuospatial perspective can be dissociated
- The UI can be *smeared* in phenomenal space
- The UI can be *duplicated* in some subjects

I think that one the most important future research targets is that the identification dimension of MPS has to be analytically grounded in a computational model, and Jakub Limanowski has the merit of having come up with some of the best work in this nascent field. One thing that I was hoping was that the "unit of identification" could be a much clearer concept for computational modelers than "the pre-reflexive self" or something like this.

Likewise, the notion of maximal UI raises a couple of questions. By your definition, when the UI is maximal and equated with "awareness as such", it seems that no phenomenological distinction can remain between the subject herself and anything she might experience. While you suggest that certain states induced by meditation might be examples of such a maximal UI, other researchers have suggested that meditation can induce conscious states in which phenomenal self-consciousness is entirely missing (Ataria, Dor-Ziderman, & Berkovich-Ohana, 2015; Dor-Ziderman, Berkovich-Ohana, Glicksohn, & Goldstein, 2013). On the face of it, self-reports of such states are consistent with this claim (although they certainly must be interpreted with caution): "It was emptiness, as if the self fell out of the picture.

There was an experience but it had no address, it was not attached to a center or subject" (Dor-Ziderman et al., 2013, p. 6). (Incidentally, if reports of expert meditators should not be taken at face value, then the same skeptical point could be made about reports of bodiless dreams or asomatic OBEs.) What do you make of such reports (if you consider them reliable enough); do you think they indeed suggest a lack of MPS (i.e., phenomenal selfhood) altogether and thus contradict a notion of maximal UI according to which the subject would literally experience everything as being identical to *herself*?

Bodiless dreams or asomatic OBEs still have an epistemic agent model, for example a "seeing self" that can control its focus of visual attention. Phenomenologically, the UI will be the sense of effort going along with mental action: the phenomenology of identification latches onto this effortful sense of control.

Autophenomenological reports necessarily presuppose autobiographical memory. Autophenomenological reports about states of "non-dual awareness" also create a "performative self-contradiction": A performative self-contradiction arises when the propositional content of a statement contradicts the presuppositions of asserting it. If *you* weren't there, why do you have an autobiographical memory of the episode? If it was *timeless*, why do you know how long it lasted? If the was no *self-location in space*, why do you know where it happened? I have been thinking about this for quite a while, as I have a long-standing interest in states of this type. It may well be possible that many of the so-called "spiritual" people underestimate what they are talking about, at least if they were to take their own beliefs about such "zero-person perspective" episodes seriously. They have nothing to do with *you*, because *you* cannot directly cultivate them, *you* cannot even prevent them. If they appear, they have nothing to do with *you*. If that is correct, your nervous system may have already realized such states in the past and *you* do not know it. They are not even episodes, because if they are timeless there is a strong sense in which they have been there all along and pervade every moment of your mundane temporal experience. Conceptually, instantiating an EAM plus MPS clearly seems to be a presupposition for autophenomenological reports. So, I think what these advanced practioners reports must be some sort of hybrid state in which the autobiographical self-model must still have been "recording" as it were.

I think one of the strengths of the new conceptual instrument of an UI is that one can ask new questions more precisely: can the phenomenology of identification and the phenomenology of selfhood be dissociated?

- For example, could there be a maximal UI that is non-selfy? Do we know conceivable and empirically plausible phenomenologies of unification with the world as a whole, which are more like an "all-pervading emptiness that has awoken to itself", i.e. more on the Buddhist side than on the Advaita Vedanta

side?

- For example, could there be a minimal UI that is non-selfy? This could for example be a phenomenology of *hacceity*. Maybe, if we do the phenomenology seriously and properly, what we really have never is MPS, but only a conscious THIS-here-now. A haecceity is a non-qualitative property responsible for individuation and identity. A haecceity is not a bare particular in the sense of something underlying qualities. It is, rather, a non-qualitative property of a substance or thing: it is a "thisness" as opposed to a "whatness".

I think there very clearly is a phenomenology of numerical singularity, namely, the subjective experience of *mere particularity*. But if that is the case, are we perhaps misdescribing exactly this phenomenology as minimal phenomenal *selfhood*, when there really is no such thing as a self there? If the phenomenology is indeterminate, then all reports are necessarily theory-contaminated. If you think of your beautiful Dor-Ziderman quote above—would the subject ever have used the word "emptiness" if there hadn't been twenty-five centuries of Buddhist philosophy for which exactly this concept was absolutely central?

But the notion of an UI also allows you to describe empirical results in a more differentiated way. Robotic re-embodiment studies demonstrate that the UI can be dissociated when given two explicit body representations as candidates for subjective self-location (Aymerich-Franch, Petit, Ganesh, & Kheddar, 2016). But do we need to speak of two *selves* in such cases, and would that even be logically coherent?

My own empirical claim that if we were to apply 1PP methods to MPS, we would very likely get some statistical distribution of certain types of reports, of which I have just presented two classical examples. Autophenomenological reports cannot determine the metaphysical status of MPS, because, for example, you cannot decide between MPS" readings and "*haecceitas*"-readings. Phenomenal indeterminacy for MPS seems to be a fundamental epistemological problem, that is why I brought it up at your Oxford conference.

In *Being no one*, you argue that there are two ways in which a conscious system could lack the phenomenology of selfhood: (a) by having a phenomenal world-model without a phenomenal self-model, or (b) by having a "fully opaque" phenomenal self-model, i.e. a phenomenal "system-model" which seamlessly accesses all stages of its own information processes (Metzinger 2003, p. 565). You acknowledge that the first case probably applies to "many simple organisms on our planet", while the second case may loosely coincide with the Buddhist notion of "enlightenment", although it is unclear whether it is nomologically possible, at least for humans. Do you believe that either of these forms of "selflessness" can be (at least temporarily)

exemplified by human subjects, for instance in radically altered states of consciousness induced by psychopathologies or psychoactive drugs?

Absolutely. Full ego-dissolution plausibly occurs in serious cases of depersonalization disorder or organic brain diseases, and I can only recommend your own paper on causal etiologies based on pharmacological stimuli (Millière, 2017)—it is perhaps the best, most well-researched, and most careful discussion of the empirical literature out there. Possibility (b) also seems quite obviously something that has happened to human beings for millennia and in many different cultures. My own attempt to approach what, if I remember correctly, I have called "system consciousness" in BNO (as opposed to "self-consciousness") is of course highly dubious, because it is relative to a certain level of description and a specific functional analysis. The way I used the concepts of "transparency" and "opacity" was as properties of phenomenal representations and, as indicated above, such concepts are historically plastic entities. Nevertheless, if my central conceptual point—namely, that for conscious self-representations transparency necessarily leads to the phenomenology of identification—still holds, then it is obvious how this specific phenomenology can gradually be dissolved by leaving the content of the conscious self-model as it is. There could be many stages, for examples for whom MPS is still robust and fully transparent, but in which the phenomenology of agency on the mental level (that is, the cognitive and attentional EAM) has disappeared, because introspective attention has penetrated into the fine-grained functional mechanisms underlying it. But again, please note the functional analysis I have developed for opacity and de-identification rests on notions like "earlier processing stages" and "vehicle properties" versus "intentional properties". Especially the last two concepts might soon begin to look as artifacts of old-school armchair philosophizing—for example, I think we may perhaps find better conceptual tools in the predictive processing framework.

To add to the previous question, several authors—philosophers and scientists alike—have argued that in a predictive processing framework, the self-model results from active (Bayesian) inference and the brain's implied prediction error minimization about which sensory signals are "the most likely to be me" across exteroceptive, proprioceptive and interoceptive domains (Apps & Tsakiris, 2014; Limanowski & Blankenburg, 2013; Seth, 2013). On this view, the brain's self-model is just a special part of its world-model. In cases in which information processing is heavily disturbed (e.g. by a pharmacological agent), it may be the case that persistent prediction errors are transmitted to higher levels of the system's generative model, resulting in an update of normally very stable predictions regarding the self and world. For example, couldn't it be the case that the phenomenon known as "drug-induced ego dissolution", described as a (reversible) loss of self-awareness at high doses of psychedelic drugs such as LSD (Letheby & Gerrans, 2017; Millière, 2017), is best construed as a breakdown of the conscious self-model itself—resulting from

an (temporary) update of hyperpriors regarding the distinction between self and world? One might argue that this would be an instance in which the first way of being "selfless" mentioned above could temporarily apply to human beings. Put in terms of your view on (phenomenal) self-models, do you think there can be a re-instantiation of a PSM after its complete opacity (e.g., induced by drugs or pathological conditions), or must there always be a part of the self-model that is conscious and transparent? If so, would this part correspond to the minimal UI or might it also be conceived as the maximal UI? Finally, how much weight do you assign nonconscious self-models, the biophysical "grounding" of selfhood in bodily background processes, in such altered "selfless" states and the re-instantiation of a perceived "ego"?

Very interesting questions, much too deep for a short interview! First, "psychedelic" means "mind-manifesting" and one of the most intriguing aspects of such states is perhaps that it makes you prior- and hyperprior-landscape itself a potential object of manifest, explicit conscious experience, simply because this landscape becomes extremely flexible and malleable, highly context-sensitive. Second, these states of consciousness hold the potential to simply make normal people who haven't thought about all these things much very concretely and directly aware of the fact that it is *literally* true that conscious experience is a model. For many subjects, it is the first and only experience ever to approximate global opacity. Now, if that even happens on the level of self-consciousness, then it obviously is quite a dramatic affair, because, if you will, it leads to a Husserlian "bracketing" of the certainty of one's very own existence.

> "Psychedelics states hold the potential to simply make normal people very concretely and directly aware of the fact that it is *literally* true that conscious experience is a model."

About the philosophical problem of a "performative self-contradiction" as related to pharmacologically induced non-egoic states I simply have to say that I have no solution and am thinking about it. Probably the answer is that full blown dissolutions are not remembered (perhaps on the unconscious levels of the bodily self-model), and that everything that people report are just graded phenomenologies, slightly incomplete mystical experiences. Maybe the memory traces are also only constructed when *leaving* such states (remember Dennett's "cassette theory" of dreaming? see Dennett, 1976). In any case, I think multisensory integration leading to MPS and the bodily self-model is an automatic bottom-up process, and it may exactly be what "rescues" the subject in a scientific experiment with ego-dissolving psychoactive substances. At some point, the Here-Now-model

becomes so good that the veil of transparency drops and everything becomes real again.

On a related note, you have raised skeptical concerns regarding reports of alleged "selfless" conscious states, arguing that they "generate a performative self-contradiction" (Metzinger 2003, p. 566). Indeed, you write, "how can you coherently report about a selfless state of consciousness by referring to your own, autobiographical memory?" (Metzinger, 2005b, p. 23). Is it not conceptually possible that the brain may store a conscious experience lacking self-consciousness in episodic memory, and then retrieve the stored memory later in an illusionary autobiographical mode of presentation? In other words, the apparent contradiction in such reports might come from the structure of memory retrieval, rather than the memory itself. Furthermore, descriptions of "drug-induced ego dissolution", for instance, frequently underline the inadequacy of the first-person pronoun to report such experiences, although it is hard to avoid using it for grammatical reasons (e.g. "There existed no one, not even me… so would it be proper to still speak of 'I', even as the notion of 'I' seemed so palpably illusory?", Millière 2017, p. 14).

I have already touched upon this topic in a previous answer, but I think you are floating a very interesting idea here, namely that there could be different phenomenal data-formats and that sincere autophenomenological reports could refer to memories that have become available under an egocentric inner mode of presentation. Would they then be false memories? What I find most intriguing about your proposal is that if something like this can happen in a larger time-window, then it could also happen in a much shorter time frame. Perhaps all experience is originally selfless, and is transformed into first-person experience by continuously creating false memories of the type you describe, via ultrafast forms of "illusionary memory retrieval"?

> "As a philosopher, I am very skeptical about all this loose talk concerning 'first-person methods' and 'first-person data'."

You have argued that "first-person data do not exist" (Metzinger 2003, p. 591), because there is no scientific procedure to settle introspective disagreements. Furthermore, you have suggested that there might not even be any "empirical fact of the matter" regarding some phenomenological disputes, because of "the possibility of phenomenal indeterminacy" (Metzinger 2013, p. 4). On the other hand, your work frequently appeals to subjective reports of altered states of consciousness (e.g. dreaming, out-of-body experiences or thought insertion) and to what you call "paradigmatic autophenomenological reports". What epistemological status do you attribute to such reports? Do you endorse the view that part or totality of phenomenal consciousness is indeterminate?

A part certainly is, and indeterminacy is an important research target for the future. Total indeterminacy would be then end of all knowledge. I am a strange person: As a philosopher, I am very skeptical about all this loose talk concerning "first-person methods" and "first-person data", but as a consciousness researcher I have certainly tried many of these methods—probably even a bit more rigorously than all the people who publicly advertise them to promote ideological forms of anti-reductionism or for purposes of academic virtue signaling and reputation management. But it is exactly because I have done a bit of this in my personal life that I am very much aware of the risk of "theory-contaminated reports", to name just one example. Most of the people exploring altered states of consciousness have extremely strong motives and metaphysical background assumptions, they look for something, otherwise they would not have the courage or discipline it takes. 1PP-loose-talk certainly has a nice, politically correct ring to it (Diversity! No evil reductionism! FINALLY taking inner experience seriously! Everybody can claim what they have always wanted to claim!) and it is the best strategy to get applause from many different types of audience. Stressing the importance of first-person methods makes everybody believe you are a good person, it is good for your career. But as I have explained in publications, the whole concept of "data" is overextended here, the original usage refers to something very different. Second, from a philosophical perspective, the *really* interesting methods are "zero-person methods"—but they are extremely difficult to talk about in any coherent manner.

But of course, we can get very far by refining interview methods and simply taking the reports themselves as data, doing careful semantic evaluation and statistics. Reports, neural correlates, and computational models can get us much further than we may often believe.

References

Apps, M. A. J., & Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, *41*, 85–97. https://doi.org/10.1016/j.neubiorev.2013.01.029

Ataria, Y., Dor-Ziderman, Y., & Berkovich-Ohana, A. (2015). How does it feel to lack a sense of boundaries? A case study of a long-term mindfulness meditator. *Consciousness and Cognition*, *37*, 133–147. https://doi.org/10.1016/j.concog.2015.09.002

Aymerich-Franch, L., Petit, D., Ganesh, G., & Kheddar, A. (2016). The second me: Seeing the real body during humanoid robot embodiment produces an illusion of bi-location. *Consciousness and Cognition*, *46*, 99–109. https://doi.org/10.1016/j.concog.2016.09.017

Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, *13*(1), 7–13. https://doi.org/10.1016/j.tics.2008.10.003

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28. https://doi.org/10.1007/s11229-016-1239-1

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dennett, D. C. (1976). Are Dreams Experiences? *The Philosophical Review*, *85*(2), 151–171. https://doi.org/10.2307/2183728

Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.

Dor-Ziderman, Y., Berkovich-Ohana, A., Glicksohn, J., & Goldstein, A. (2013). Mindfulness-induced selflessness: a MEG neurophenomenological study. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00582

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Gallagher, S., & Allen, M. (2016). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, 1–22. https://doi.org/10.1007/s11229-016-1269-8

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582. https://doi.org/10.1007/s11229-015-0762-9

Graziano, M. S. A. (2017). The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. *Frontiers in Robotics and AI*, *4*. https://doi.org/10.3389/frobt.2017.00060

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Kanai, R. (2017). We Need Conscious Robots. How introspection and imagination make robots better. *Nautilus*, 47. http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

Letheby, C., & Gerrans, P. (2017). Self unbound: ego dissolution in psychedelic experience. *Neuroscience of Consciousness*, *3*(1). https://doi.org/10.1093/nc/nix016

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00547

Madary, M., & Metzinger, T. K. (2016). Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. *Frontiers in Robotics and AI*, *3*. https://doi.org/10.3389/frobt.2016.00003

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, Mass.: A Bradford Book.

Metzinger, T. (2005a). Out-of-Body Experiences as the Origin of the Concept of a 'Soul '. *Mind and Matter*, *3*(1), 57–84.

Metzinger, T. (2005b). Precis: being no-one. *Psyche*, 1–35.

Metzinger, T. (2008). Empirical Perspectives From the Self-Model Theory of Subjectivity: A Brief Summary with Examples. In R. Banerjee & B. K. Chakrabarti (Eds.), *Models of Brain and Mind: Physical, Computational, and Psychological Approaches*. Elsevier.

Metzinger, T. (2009a). *The EGO Tunnel: The Science of the Mind and the Myth of the Self* (1 edition). New York: Basic Books.

Metzinger, T. (2009b). Why are out-of-body experiences interesting for philosophers?: The theoretical relevance of OBE research. *Cortex*, *45*(2), 256–258. https://doi.org/10.1016/j.cortex.2008.09.004

Metzinger, T. (2013a). The Myth of Cognitive Agency: Subpersonal Thinking as a Cyclically Recurring Loss of Mental Autonomy. *Frontiers in Psychology*, *4*, 931.

Metzinger, T. (2013b). Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.), *Robotik und Gesetzgebung* (pp. 247–286). Baden-Baden: Nomos.

Metzinger, T. (2013c). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Consciousness Research*, *4*, 746. https://doi.org/10.3389/fpsyg.2013.00746

Metzinger, T. (2014). First-order embodiment, second-order embodiment, third-order

embodiment: From spatiotemporal self-location to minimal phenomenal selfhood. In L. Shapiro (Ed.), *The Routledge Handbook of Embodied Cognition* (pp. 272–286). London: Routledge.

Metzinger, T. (2016). Suffering. In K. Almqvist & A. Hagg (Eds.), *The Return of Consciousness*. Stockholm: Axel and Margaret Ax:son Johnson Foundation.

Metzinger, T. (2017a). The Problem of Mental Action: Predictive Control without Sensory Sheets. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predicitive Processing*.

Metzinger, T. (2017b). Benevolent Artificial Anti-Natalism (BAAN). *EDGE Essay*. https://www.edge.org/conversation/thomas_metzinger-benevolent-artificial-anti-natalism-baan

Metzinger, T. (2018). Why is Mind Wandering Interesting for Philosophers ? In K. C. R. Fox & K. Christoff (Eds.), *The Oxford Handbook of Spontaneous Thought: Mind-wandering, Creativity, Dreaming, and Clinical Conditions*. Oxford University Press.

Metzinger, T., & Gallese, V. (2003). The Emergence of a Shared Action Ontology: Building Blocks for a Theory. *Consciousness and Cognition*, *12*(4), 549–571.

Metzinger, T., & Wiese, W. (2017). *Philosophy and Predictive Processing*. MIND Group. Retrieved from https://predictive-mind.net/

Metzinger, T., & Windt, J. M. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath, & J. Kipper (Eds.), *Die Experimentelle Philosophie in der Diskussion* (pp. 279–321). Berlin: Suhrkamp.

Millière, R. (2017). Looking For The Self: Phenomenology, Neurophysiology and Philosophical Significance of Drug-induced Ego Dissolution. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00245

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Weisberg, J. (2006). Consciousness Constrained: A Commentary on Being No One. *Psyche*, *12*(1).

Psychedelics
From pharmacology to phenomenology

An interview with
David Nichols

By Leor Roseman & Christopher Timmermann

Citation: Nichols, D., Roseman, L. & Timmermann, C. (2018). Psychedelics: from pharmacology to phenomenology. An interview with David Nichols. *ALIUS Bulletin*, 2, 75-85.

David Nichols
denichol@email.unc.edu
Eshelman School of Pharmacy
University of North Carolina, Chapel Hill, USA

Leor Roseman
leor.roseman13@imperial.ac.uk
Department of Medicine
Imperial College London, UK

Christopher Timmermann
c.timmermann-slater15@imperial.ac.uk
Department of Medicine
Imperial College London, UK

L.R. and C.T. would like to thank Tobias Buchborn for suggesting two questions used in this interview.

The relationship between the pharmacology of psychedelics and their effects on consciousness are usually obscured by a complex myriad of interactions, extra and intracellular mechanisms, etc. What do you think is the correct approach to bridge mechanisms stemming from the molecular level to complex human behavior? In what way do you think psychedelic drugs can provide insights into these mechanisms?

I think modern brain imaging technologies are going to be playing an increasingly important role. Correlating subjective effects with functional effects in specific brain areas should be very revealing. We already know a lot about the neurotransmitter systems that operate in the various anatomical areas of the brain, so coupling all that with brain imaging will be important. We still need to know a lot more about what intracellular signaling cascades are important, and how they affect behavioral endpoints. We are really in the infancy of brain science, and a hundred years from now people will look back and think that the things we did were very primitive. But I believe that psychedelics will prove to be crucial tools to help us understand consciousness.

Tryptamines (e.g., LSD, psilocybin, DMT) and some phenethylamines (e.g., mescaline, 2C-B) are both serotonin 2A agonists and classic psychedelics (see fig. 1 below). However, they have different chemical structures. Could you please explain how they differ chemically and how this difference accounts for the distinct behavioral and phenomenological effects they each produce?

Although several classes of molecules are 5-HT2A agonists, what happens after they interact with the receptor is probably different. The concept of functional selectivity, or ligand bias has been an evolving pharmacological concept for more than 15 years. The way the ligand engages the receptor, that is, the way it docks into the receptor and the amino acid residues it engages, allows the receptor to adopt different shapes, or conformations. These different conformations produce different conformations in the intracellular connecting loops of the receptor, and these different conformations can engage different signaling components. For example, a G protein coupled receptor (GPCR) of which the 5-HT2A receptor is one, can couple to various G proteins within the neuron; Gq, Gi, Gs, etc. In addition, serine and threonine residues in the intracellular receptor loops can be phosphorylated by G protein receptor kinases, and then the phosphorylated fragments can recruit beta-arrestin. Different 5-HT2A agonists, can recruit different intracellular pathways to different extents, and those different signaling pathways undoubtedly lead to subtle differences in the behavioral effects.

“ Different 5-HT2A agonists, can recruit different intracellular pathways to different extents, and those different signalling pathways undoubtedly lead to subtle differences in the behavioural effects. ”

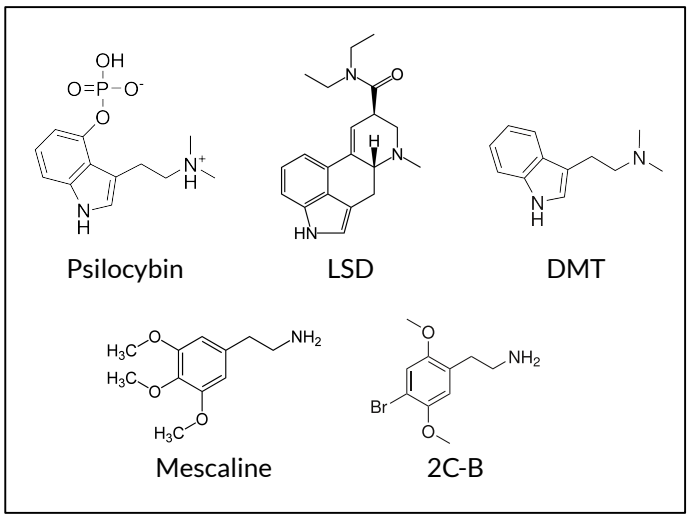


Fig. 1 – A few well-known psychedelic molecules

You are an expert in designing new analogues for different psychedelics. Could you tell us what the rationale behind designing successful analogue is? Is it hypothesis driven, luck (trying lots of different chemical structures) or both?

As an academic, my work had to involve hypothesis testing. On occasion, we might ask “I wonder what this structure would do?” and then we would prepare it to find out. Most often, however, we had a specific hypothesis we tested. Our hypotheses

mostly centered on defining the shape, or conformation of the side chain in tryptamines or phenethylamines, as well as defining the shapes of the methoxy groups in the phenethylamines. For example, that led us to synthesize complex and rigid phenethylamines such as 2-aminotetralins, benzocyclobutenes, and aminomethylindans. The original impetus for most of these studies was an attempt to understand how the 5-HT2A receptor could accommodate different chemotypes, i.e., ergolines, tryptamines, and phenethylamines.

What is the most selective serotonin 2A agonist? What is the subjective experience of this drug?

The most selective 5-HT2A agonists to date have never been tested in humans. One was developed in Denmark, and is a 2,5-dimethoxy-4-cyano-N-(2-hydroxybenzyl) phenethylamine (25CN-NBOH). The other is a three-ring 25B-NBMOMe type structure, where the ethylamine side chain has been tethered into a piperidine ring. The latter structure was crystallized and we published the x-ray crystallographic structure of it, and that gave us an idea of how the side chain of the NBOMe compounds must bind to the receptor. I would love to see clinical tests of a very selective 5-HT2A agonist, because all known psychedelics are both 5-HT2A and 5-HT2C agonists, and in the brain these two receptors generally are functionally opposed to each other.

In a recent study that you were involved in (Wacker, 2017), you demonstrated that the serotonin 2B receptor (very similar to 2A) has a lid-shaped structural extension that stays closed for longer periods every time LSD is attached to the receptor, and that this "lid" traps the LSD inside, which ultimately accounts for its prolonged duration of action. Could you please expand on this finding? What would you hypothesize that other classic psychedelics are doing to the "lid"? What is special about the structure of LSD that closes the "lid"?

The piece of the receptor that does that is called extracellular loop 2, or EL2. Before I retired from Purdue, my last graduate student had mutated all of the residues in EL2 for the 5-HT2A receptor. We did binding studies in each mutant and compared LSD with some LSD analogues known as azetidides, where the diethyl group of LSD had been tethered into a four-membered azetidine ring with appended methyl groups. We had compared the pharmacology of the three stereoisomers, where the 2,4-dimethylazetidine ring had a cis stereochemistry, or an R,R or an S,S configuration (McCorvy, 2012). We found that the S,S configuration gave a compound closest in pharmacology to LSD itself. (That structure has appeared on the "research chemical" market as LSZ). Mutations of the residues in EL2 showed that mutation of leucine 229 to an alanine had an effect that was similar for LSD and the S,S-azetidide, but different for the R,R and cis stereoisomers. Later, working in Bryan Roth's laboratory, it was found that the S,S azetidide had pharmacology

similar to LSD in the 5-HT2B receptor, but the key residue in EL2 in that case was Leucine 209 (Wacker 2017). In examining the receptor kinetics of LSD in the wild type 5-HT2A and wild type 5-HT2B receptors, compared to the L229A and L209A mutant receptor, respectively, it was discovered by John McCorvy, a postdoc in the lab there, who was my last graduate student at Purdue, that in both of the wild type receptors, LSD had a very slow association rate, and an extremely slow dissociation rate. In the Leucine to alanine mutant receptors, LSD had very fast association and dissociation kinetics (Wacker, 2017). In the x-ray crystal structure of LSD in the 5-HT2B receptor, that loop could be seen laying over LSD within the receptor, and Leucine 209 sort of wedged down between the LSD molecule and the receptor. In essence, EL2 was able to "lock" LSD into the receptor. There are now attempts to obtain the crystal structure of LSD bound into the 5-HT2A receptor, but based on the kinetics studies done by John McCorvy, we expect to see a similar "locking" mechanism with EL2. With respect to other 5-HT2A ligands, I suspect that we will see faster on and off kinetics. We think that the ability of the receptor to sequester the LSD may be a key to its high potency and profound psychopharmacology. Numerous LSD analogues have been made and tested, where the diethylamide was modified, and we have no indication that they have the type of activity seen with LSD. So it seems likely that the diethylamide is just the right size and shape and adopts a unique conformation to keep LSD in the receptor. Except for mescaline, most of the other psychedelics have a shorter duration of action, and that may reflect, to some extent, their receptor kinetics. We also found that the 5-HT2A and 5-HT2B receptors recruit beta-arrestin2 in a time-dependent manner; the longer the LSD remains in the receptor, the more robust is the arrestin signaling. That phenomenon may also be an important feature that contributes to the potent effects of LSD.

In the same work (Wacker, 2017), you showed that ergotamine (a non-psychedelic 5-HT2A agonist) and LSD—likely due to differences in conformational receptor change—differentially recruit cascades downstream of 5-HT2 activation. Might these differences account for the lack of psychoactivity of ergotamine? What is the current understanding of Gq-PLC/PLA/PLD, Gi, and arrestin dependent signalling as to their significance for the behavioural and psychedelic effects of 5-HT2A agonists?

I touched on this point earlier. We believe that arrestin recruitment may be very important, but many active molecules seem to have some selectivity for G protein signaling. So that is an important area that needs detailed research. Sadly, the lack of government funding has meant that few people are interested in studies like these, which would be time-consuming and very comprehensive.

In terms of structure-activity relationship, early ideas suggest that certain tryptamines (e.g., psilocin) as well as certain phenethylamines (e.g., mescaline) are able to form intramolecular hydrogen bonds so to mimic ring C and B of LSD, respectively (Snyder & Richelson, 1968). Based on your research, what is the current understanding/evidence about these bonds being of vivo relevance? What might be the relevance of these bonds for fitting the binding pocket of 5-HT2A and/or the drugs' vulnerability to enzymatic degradation?

That idea was proposed early on by Solomon Snyder, but it has been thoroughly discredited by now. It never really made sense to me, as a chemist, but we had to generate the proof. We have some ideas about how psilocin might bind to the receptor, and its orientation is probably not too different from that of bound LSD. However, other than the conserved aspartate in helix 3, LSD does not engage other polar residues except perhaps a serine in helix 5. Psilocin likely engages that same serine, but also it appears to interact with one or two other polar residues. By contrast, we really have no idea how mescaline or other phenethylamines bind, but our mutagenesis studies of the receptor did demonstrate that the phenethylamines engage residues different from those that interact with tryptamines.

Besides classic psychedelics (5-HT2A agonists), there are other drugs that can create a psychedelic experience (e.g., Ketamine (NMDA antagonist), Salvinorin A (κ-opioid receptor agonist), Scopolamine (anticholinergic)). Do you believe that there is a common mechanism shared by these drugs? and if so what is it?

Salvinorin A I think is very different, and is a very selective agonist at the kappa opioid receptor. Users generally find the experience very different from an LSD trip and often very unpleasant. Scopolamine and other anticholinergics produce true hallucinations and a sort of psychotomimetic experience. They also produce amnesia for the experience, which is very different from the 5-HT2A type of agonists. Ketamine is an interesting example, because it leads to increased release of neuronal glutamate (Abdallah, 2016). Classic 5-HT2A agonists also lead to increased brain glutamate, and if co-administered to animals along with ketamine, they can give a potentiated response. Glutamate appears essential to the actions of classic 5-HT2A agonists (Nichols, 2016), so there may be some overlap mechanistically between 5-HT2A agonists and ketamine. Again, we need a lot more research.

In a recent talk you gave at Breaking Convention (https://youtu.be/YeeqHUiC8Io) you argued that endogenous production of DMT (a naturally-occurring psychedelic which is also found in the Ayahuasca brew) is not associated with spontaneous experiences, which may resemble the ones experienced under psychedelic states (e.g., near-death experiences, mystical/peak experiences, etc.). This is contrary to Rick Strassman's argument that endogenous production of DMT might be responsible for these experiences (Strassman, 2001). Could you outline the

strongest points for your argument and what may be the correct experimental approach to the study of biological mechanisms which may be underlying such experiences?

Rick Strassman kind of backed off of his statement by saying it was just "speculation" (Strassman, 2001). The talk I gave there has just appeared in the *Journal of Psychopharmacology*, and the arguments are a bit too detailed to review here, but there are several important points in the paper (Nichols, 2017).

It has also been proposed that DMT may have a neuro-protective function in life-threatening situations (i.e., under oxidative stress) (Szabo & Frescka, 2016). What is your view on this hypothesis? Does the current evidence on endogenous production of DMT support this view in your opinion?

No, essentially the affinity of DMT for sigma receptors is too low for it to be consequential. There is no known mechanism for the production of DMT that would lead to in vivo concentrations high enough to excite any of the known receptors. DMT has only been detected in very trace amounts using very sensitive LC-MS methods.

> "There is no known mechanism for the production of DMT that would lead to in vivo concentrations high enough to excite any of the known receptors."

The function of the serotonin system has remained an elusive subject. You have argued that the discovery of LSD (and its similarity to serotonin) was an important player in unveiling the relationship between brain chemistry and behavior. Recently, Carhart-Harris & Nutt (2017) have proposed a general framework for this system based on work with psychedelics. They have argued that the complexity of the serotonin system may be related to the ability of the organism to flexibly adapt to the demands of the environment, with 5-HT1A and 5-HT2A receptors mediating passive and active coping to stressful stimuli respectively. Do you agree with this hypothesis?

I think their hypothesis is somewhat superficial and fails to account for the wide diversity and expression of the other subtypes of receptors. Certainly 5-HT2A receptors are excitatory, and 5-HT1A receptors are inhibitory, but I don't feel that the contrasting pharmacology of those two subtypes is really sufficient as a comprehensive explanation.

We are experiencing the so called psychedelic renaissance. A renaissance which includes psychedelic science and therapy (Nichols & Johnson, 2017). What is the new knowledge that we have discovered during the current renaissance?

I think we are learning a lot more about how the brain generates mind. We are also learning that psychedelics seem to have tremendous healing potential, which might also be connected with the brain-mind connection. I believe we are just at the beginning of a revolution in thinking about brain, behavior, and emotional disorders, and that the future will be really interesting, once major institutional funders get on board. There are many young scientists interested in this field of research, but if you are an academic, you have no future without major funding. Once agencies begin to recognize the profound importance of understanding psychedelics and how they affect the brain, I believe we will see knowledge enter an exponential phase of grown.

> “I believe we are just at the beginning of a revolution in thinking about brain, behavior, and emotional disorders, and that the future will be really interesting, once major institutional funders get on board.”

What important knowledge about psychedelic is lacking? And when do you think we will gain it?

That is a question that I cannot begin to answer. Like any new field of investigation, there are things we will discover that we had probably never thought about before. A central question that everyone in this field thinks about (I hope) is “who is man?” Philosophers used to debate the nature of man, and still debate the nature of consciousness. Who are we, and why are we here? Is man just a complex biomachine that evolved through random natural selection, or does he have some connection to other beings, organisms, and to life in general? Unfortunately, those debates do not earn any money, so in the modern money-driven world, people seem to have forgotten them. Psychedelics force us to rethink these questions. They force us to think about the nature of mind, and of memory. A recent finding was that people who use psychedelics tend to be more altruistic. Why is that? Their personality trait of openness is also increased. How and why does that happen? I don’t want to go too far out on a limb, but perhaps some people who use psychedelics actually become better people. It would be interesting to know how that happens and if it could generally be applied to improve personality.

Science can be quite confusing, as many labs show contradicting results which are sometimes serving a certain agenda. Is there anything that we are sure about in psychedelic research?

This field in general is loaded with the potential for all kinds of magical thinking. There are modern scientific studies now published that involve very poor science.

Part of that may be due to poor reviewing at the journals. Part of it may result from wishful thinking; the investigator wants to prove their hypothesis so badly that they misinterpret their data. There was a lot of that in the early research. Hopefully, the majority of scientists in this field today are aware of the great need to do things right this time around. As a high-profile speaker said at a recent MAPS conference, "Don't screw it up this time".

The field of psychedelic research is noticeable for its interdisciplinary nature. Conferences on psychedelics substances usually have contributions stemming from anthropology, chemistry, neuroscience, psychology, biology and philosophy. Many times, however there is a lack of conversation between fields which may greatly benefit from some of this cross-talk. In your opinion, in what way should this multidisciplinary aspect find expression so that the field benefits most from it?

I think if the scientists are good, and well-trained, they can speak to each other. What often happens, however, is someone with a modicum of training in, for example, anthropology comes up with a poorly documented idea that they are able to sell to the uninformed. And then their myth begins to spread. A lot of well-trained "scientists" come up with dubious ideas, but they rely on people to accept their ideas because they have a PhD, or an MD, and write a book. Well-trained scientists can generally see through that sort of hokum. More often, however, it is the less well trained who are susceptible to half-baked urban legends about psychedelics. I really resent "scientists" who use their credentials to gain prestige with less well-informed masses who are simply hungry for knowledge. As a chemist, I have enjoyed conversations with scientists in many other fields, so I think the key is that the people in the different fields have to be well trained and have integrity.

> "A lot of well-trained "scientists" come up with dubious ideas, but they rely on people to accept their ideas because they have a PhD, or an MD, and write a book."

Psychotomimetic is a term that was applied in psychedelic research when psychedelics were considered as mimicking psychosis. Most psychedelic researchers today would avoid using this term, however there is still insight we might gain about psychosis using psychedelics. What are these insights in your opinion?

I think very early onset schizophrenia might have some resemblance to psychedelic actions, where you find hypermetabolic effects. Remember, the atypical antipsychotic drugs are antagonists at the 5-HT2A receptor, the target for classic psychedelics. And activation of the 5-HT2A receptor also can enhance dopaminergic brain function (Nichols, 2016), another monoamine that seems key to psychosis.

Microdosing has become quite fashionable in the past few years. It is quite different than the regular psychedelic use in which the emphasis is the psychedelic experience. What is your view on the mechanism of microdosing? What is your opinion about a chronic administration of a psychedelic?

I think it is a bad idea. There is no controlled study to show that it actually does anything, and there are no studies comparing it with a prescription psychostimulant such as Modafinil or Ritalin. It seems theoretically possible that a low dose of LSD might do something, because it gets trapped in the receptor, but LSD also stimulates the 5-HT2B receptor, which can lead to cardiac valvulopathy. But there has been no well-controlled study to show that LSD actually enhances creativity. And if you think about a dose-response curve, even if you enhanced creativity at an effective dose of LSD, what pharmacological reason is there to expect that you will enhance creativity at a low dose? So I don't think it is a good idea. I think it is a fad that will die off at some point.

In the 1997 MAPS bulletin you wrote "If you do psychedelic research, and that is all you do (I have some other more mainstream research in addition to the psychedelic work), you have perhaps half-a-dozen people world-wide who share your research interests. Perhaps not surprisingly, you may develop a sort of cult following, but that kind of adoration is not particularly fulfilling. People occasionally tell me that my name is known all over the world in the 'psychedelic community'. While that may be true, it doesn't get recognition within the scientific community, which is my workplace, comprised of my peers. What you want is recognition from them that you are doing good work. You are unlikely to get it, so your rewards must come from within yourself, and you must believe that someday the value of your work will become clear to other people, because that is unlikely to occur in your own lifetime. It will help if you are the sort of person who can deal easily with delayed gratification". Is it different now?

I think that is still the case. Most of the researchers I know are doing it because of a personal drive that tells them it is important work. I have often thought that if I had gone into a different area of research, cancer, heart disease, etc., that I might have gained recognition for my work in mainstream circles. Among the bulk of mainstream medicinal chemists I believe I am largely unknown, despite publishing hundreds of research publications and giving seminars all over the world. It is frustrating, but I believe that what I have done is very important, and it is gratifying to see it gaining more traction today.

What are your hopes and concerns about mainstreaming psychedelics?

I hope we are witnessing a paradigm shift in the treatment of all kinds of emotional and psychiatric disorders. I used to think I would be dead before any of that happened, but now I see potential approval for these medicines in the early 2020s,

while I hope to still be alive! And before that, I believe that national agencies, the NIMH in the U.S. for example, will start funding research in this field at the level it should have been for all these past several decades. Then we will know that the field is maturing as lots of new young scientists will be attracted to study psychedelics.

References

Abdallah, C. G., Adams, T. G., Kelmendi, B., Esterlis, I., Sanacora, G., & Krystal, J. H. (2016). Ketamine's mechanism of action: a path to rapid-acting antidepressants. *Depression and anxiety*, 33(8), 689-697.

Carhart-Harris, R. L., & Nutt, D. J. (2017). Serotonin and brain function: a tale of two receptors. *Journal of Psychopharmacology*, 31(9), 1091-1120.

McCorvy, J. D. (2012). Mapping the binding site of the 5-HT2A receptor using mutagenesis and ligand libraries: insights into the molecular actions of psychedelics.

Nichols, D. E. (2016). Psychedelics. *Pharmacological reviews*, 68(2), 264-355.

Nichols, D. E. (2017). N, N-dimethyltryptamine and the pineal gland: Separating fact from myth. *Journal of Psychopharmacology*, 0269881117736919.

Nichols, D. E., Johnson, M. W., & Nichols, C. D. (2017). Psychedelics as medicines: an emerging new paradigm. *Clinical Pharmacology & Therapeutics*, 101(2), 209-219.

Snyder, S. H., & Richelson, E. (1968). Psychedelic drugs: steric factors that predict psychotropic activity. *Proceedings of the National Academy of Sciences*, 60(1), 206-213.

Strassman, R. (2001). *DMT: The Spirit Molecule*. Rochester, VT: Park Street Press.

Szabo, A., & Frecska, E. (2016). Dimethyltryptamine (DMT): a biochemical Swiss Army knife in neuroinflammation and neuroprotection? *Neural regeneration research*, 11(3), 396.

Wacker, D., Wang, S., McCorvy, J. D., Betz, R. M., Venkatakrishnan, A. J., Levit, A., ... & Shoichet, B. K. (2017). Crystal structure of an LSD-bound human serotonin receptor. *Cell*, 168(3), 377-389

On the "feel" of things

The sensorimotor theory of consciousness

An interview with
Kevin O'Regan

By Cordelia Erickson-Davis

Citation: O'Regan, K. & Erickson-Davis, C. (2018). On the "feel" of things: the sensorimotor theory of consciousness. An interview with Kevin O'Regan. *ALIUS Bulletin*, 2, 87-94.

Kevin O'Regan
jkevin.oregan@gmail.com
Laboratoire Psychologie de la Perception
Université Paris Descartes, France

Cordelia Erickson-Davis
cred22@stanford.edu
Departments of Anthropology and Medicine
Stanford University, USA

You are primarily interested in the "what it's like" aspect of sensory experience. To address it, you've developed what is called the "sensorimotor theory" of consciousness (O'Regan & Noë 2001, Noë & O'Regan 2002; Myin & O'Regan 2002, O'Regan 2011, O'Regan 2014), which holds perception to be a law-governed mode of encounter with the environment. These laws are abstracted from the sensorimotor contingencies of the animal in relation with the environment—both the contingencies fixed by the perceiver's visual apparatus, as well as those fixed by the character of objects. These processes of sensorimotor interaction are distinct from perceptual consciousness, however, which you divide into "perceptual awareness" on one hand (or "transitive perceptual consciousness"), which is the exercise of one's practical knowledge of these sensorimotor contingencies, and "general perceptual consciousness" on the other, which is the capacity to become aware. Do I have that right? Could you further describe the theory, and in particular focus on how it addresses consciousness from the perspective of what you and others have described as the "easy" and "hard" problems of consciousness?

Yes, I'd say you've got it more or less right. But the way you say it sounds very technical. I would have liked you to stress the luminous simplicity of the idea and why it is a breakthrough in understanding consciousness!

To explain better, let me first note that most people think that consciousness is a mystery—they think that there is a problem in explaining how physical and chemical processes in a brain could somehow generate subjective experience. Philosophers call this the "hard" problem of consciousness.

Science doesn't just advance by making discoveries: it advances by defining terms in more precise ways. I think such redefining is what's needed to understand consciousness. The following is a way of defining consciousness that captures what most people mean by the term, and that at the same time dissolves the mysteries.

The definition has two layers:

First, at the top layer: in the normal everyday sense of the word, when you say you are conscious of something... there has to be a "you" with sufficient cognitive capacities. People would not normally say of a fly that it is conscious of the cheese it landed upon: The fly is presumably just a biological machine that is reacting to the environment. What about the mouse that ate the cheese? And the cat that ate the mouse? And the dog that chased the cat? And the child that chased the dog? And the adult that scolded the child? Clearly as we go higher in the hierarchy, we have higher degrees of being "conscious of". Certainly the adult's, if not the child's, understanding of the situation involves them not just reacting, but a variety of other things like knowing that they are reacting, knowing who "they" are, and knowing why they are reacting, and knowing that they know that they are reacting...

Though maybe too complex for flies and mice, there is nothing magical about such self-referring cognitive states. Being "conscious of" something in this way is what the philosophers call the "easy" problem of consciousness. It requires a variety of highly developed cognitive capacities, including the ability to conceive of one's "self"—but this mode of being conscious of something is not a mystery. It is coming to my smartphone in the next decades.

That was the top layer. But now there is the bottom layer of consciousness. I can be conscious of *an experience*. For example, I can be conscious of the hurt of the pain, or the redness of a red sunset. They *feel like something* to me. It's not just that I'm thinking of them or aware of them, like I can think of a pain or of a red sunset. I actually *feel* them. What is it like to feel things, rather than not feel them like when you are just thinking about them? This is what the philosophers call the "hard" problem, or the problem of "qualia". They think it's hard because they see no way brains could generate feels.

> "We should conceive of the feel of a pain, and the feel of red, and all perceptual experiences as *ways of interacting with the world*."

And that is where I think a redefinition helps. In fact, the redefinition I propose is perfectly obvious to the man in the street, who would never imagine that a feel could be generated by the brain. What after all, is the feel of driving a Porsche? Well, it's the way it *handles* when you swing around the corner, it's how it *reacts* when you press the accelerator and it speeds forward... it's *how you interact with it.* Similarly, we should conceive of the feel of a pain, and the feel of red, and all perceptual experiences as *ways of interacting with the world.*

At first counterintuitive for the scientist looking for brain mechanisms, this way of thinking about experience provides an exquisitely simple account of the "hard" problem: when you have an experience, the what-it's-like of the experience is constituted by *how you interact with the world* when you're having the experience. But this experienced quality is not *consciously* experienced unless you as a person are attending to it, making use of it in your rational thoughts, decisions, planning etc., in the way of being "conscious of" that I described in the top layer of my account.

In summary: you are having a conscious experience of red when, at the top layer, you are "conscious of" the fact that, at the bottom layer: you are currently engaged in interacting with the world in a way that is constitutive of the laws of redness. The top layer provides the awareness, the bottom layer provides the experienced quality.

What motivated you to think about these issues, and what was your training up until that point?

Ever since I was a child I wanted to make a machine that thinks. I started off studying physics, because I thought the brain could be understood using the methods of statistical physics that try to model the behavior of large numbers of interacting bodies. I then moved into experimental psychology, where I worked on eye movements and visual perception. I realized that there was a logical flaw in the way people think about perception: people assume that perception involves the brain making an internal representation of the world. But then: who or what perceives that internal representation? It was this realization that led me to postulate a "sensorimotor" theory, where perceiving involves interacting with the world, not making an internal representation.

Many aspects of your theory resonate with other approaches that fall under the umbrella of "embodied cognition". This includes the "ecological perspective" of Gibson (1966, 1979), the programs of "active" and "animate" perception (Aloimonos et al. 1988; Ballard et al. 1997), embodied artificial intelligence (Brooks 1991), autonomous systems (Varela & Bourgine 1992), and enactive perception and cognition (Thompson & Varela 2001; Varela et al. 1991). How does the sensorimotor theory compare with these other theories?

The sensorimotor theory is trying to address the "hard" problem of consciousness: Why do things feel the way they do? Why does "red" look "red" rather than "green". Why does "red" not sound like a bell? Gibson's ecological approach and embodied artificial intelligence are not trying to solve that issue, and are instead looking at the role of action in perception. To me this is not very exciting: action obviously improves perception because it provides more information to be gathered. But these approaches miss another important role for action, namely what it brings to an understanding of the experienced quality of sensory experience and consciousness.

Autonomous systems and enactive approaches are, on the other hand, addressing the issue of consciousness. They invoke action as an essential element in consciousness. However, my impression is that they think that there is something magical about action. They think that interaction with the world somehow instills consciousness into biological systems. Their appeal to interaction, and concepts like autopoeisis, seems to be an attempt to use mysterious notions to elucidate what they think is even more mysterious, namely consciousness. What these approaches seem not to have realized is that if we understand the what-it's-like of perceptual experience as being constituted by what we do when we interact with the world, then there is actually no mystery. In other words, autonomous systems and enactive approaches correctly invoke action, but they don't realize why it is that action solves the mystery of qualia. They seem to want to keep a mystery where there is no mystery.

> "Autonomous systems and enactive approaches seem not to have realized that if we understand the what-it's-like of perceptual experience as being constituted by what we do when we interact with the world, then there is actually no mystery."

In your theory, phenomenological inquiry takes on an entirely tractable tone. As you say, "the subject matter of phenomenological reflection is not an ephemeral, ineffable, sensation-like momentary occurrence in the mind, but, rather, the real-world, temporally extended activity of exploring the environment and the structure of sensorimotor contingencies" (O'Regan & Noë 2001, p. 962). What kinds of phenomenological reflection have you utilized in your experimental work? That is, how have you operationalized subjective inquiry in your empirical approach?

For example, I'm very proud of the work we did with David Philipona on color. Color a priori doesn't seem to involve interacting with the environment. But by taking a sensorimotor approach to color, we were forced to postulate that the experience of color is necessarily rooted in what happens when you move colored surfaces around in different lights. This gave a completely new idea about what color perception is, and made interesting predictions about what it means to be a "pure" color. We found a surprising link to anthropological data about color naming, where we accurately predicted which color names should occur most frequently. A simple philosophical idea, the sensorimotor approach, provided a surprising scientific prediction.

As you discuss in your book, the consequence of assuming that experience derives from the rules that govern action-related changes in sensory input is that the "feel"

of perceptual modalities like visual experience should be obtainable via channels other than vision ("provided that the brain extracts the same invariants from structure") (O'Regan & Noë 2001, p. 956). That is to say: sensory substitution. Can you describe the sensory substitution work you've done over the years, and what you've learned from it?

Indeed, the sensorimotor theory predicts you should be able to see with your ears, for example, or hear with your skin, provided you use some technical tricks to recreate the same sensorimotor laws via alternate sensory channels.

With my collaborators over the last years we have looked at how visual information can be conveyed through auditory input, how auditory information can be conveyed through the skin, and how it might be possible to obtain an augmented "sixth" sense of magnetic North via hearing.

Our efforts have been somewhat disappointing. We have discovered that it is technically not so easy to provide the brain with the right sensorimotor laws. Furthermore, we have found that the adult brain seems to be less flexible than we had thought. Our latest efforts to make a tactile aid that helps hearing-impaired people better understand speech has proven much more difficult to realize than we had anticipated. It seems that adult humans have a hard time making the arbitrary links that we require between speech sounds and tactile patterns. It may be that we have not been doing things right however. Perhaps the problem is that up until now, we have not included a proper "action" component in our approach. Sensorimotor theory would suggest that this would be necessary.

> "We will have machines that interact with us in ways that gradually involve higher and higher levels of cognition, including meta-cognition. We will not hesitate to say that they are conscious."

You have said you believe that it is possible to build a robot that "feels". Does that mean that you believe artificial intelligence and robots are now or will be considered conscious?

Definitely. If consciousness is just a word that describes certain capacities we have to interact with the world, then machines are already on their way to being conscious. As I said earlier, whether a mouse, cat, dog, child or adult is conscious is a matter of degree. Similarly, in the next few decades, we will have machines that interact with us in ways that gradually involve higher and higher levels of cognition, including meta-cognition. When such machines interact with us socially every day,

when they have levels of knowledge and (meta-)cognition approaching (or superseding!) ours, we will not hesitate to say that they are conscious. Furthermore, when machines interact with the world with their senses, they will have experiences just like we have experiences when we interact with the world. The experiences will be different of course, precisely to the extent that their modes of interaction are different from ours. But that is true of mice, cats, dogs, and children too.

You have also said, in conversation, that you think that our societal focus on consciousness as a rubric for ethical decision-making is a mistake. Can you say more about this? What kind of ethics of artificial intelligence and other forms of cognitive enhancement technologies (e.g., brain machine interface devices) do you think is needed?

If consciousness is not an all-or-none thing, and is just a matter of having certain capacities (and meta-capacities) to interact with the world, then consciousness is useless as a criterion for ethical decisions. And even if there were some objective criterion of consciousness, it would be a pretense to invoke it: civilizations have often denied ethical respect to various perfectly conscious groups. Slaves, women, certain ethnic and religious groups, have all been denied human rights at various times through history. Ethics is ultimately a matter of social agreement, and human societies must take full responsibility for the decisions they take about whom to give ethical rights to. Appealing to science is just hypocrisy.

How can sensorimotor theory be applied to "alternative" states if consciousness—e.g., dreams and hallucinations? On these perceptual states, you've written that you believe them to correspond to implicit knowledge and implicit expectation, based on prior perceptual experience. Do you think the study of these states has anything to contribute to consciousness studies, from a sensorimotor perspective?

I personally haven't worked on implications of sensorimotor theory for altered states of consciousness, as produced for example by drugs, trances or meditation.

Note that some critics of sensorimotor theory have claimed that the theory cannot account for dreams and sensory hallucinations, since these occur without any interaction with the world. But this is to misunderstand the theory. The theory says that the quality of an experience resides in what you do when you are interacting with the world. If, through drugs or dreams you are in the same state that you usually are in when you are interacting with the world in a "red" way, then you will experience red, even if now you happen not to be interacting with the world at all.

The critics sometimes go on to say: well, doesn't that show that the brain does generate experience after all, since you can get the experience without interacting with the world? My answer is that the brain *enables* the experience, since a brain is

necessary in order to interact with the world. But that doesn't mean that the brain *generates* the experience in any meaningful way: Experiences are not the kinds of things that can be generated. Experiences are modes of interaction with the world. This new way of thinking about experiences is hard for some people to embrace. But note that a similar change in point of view happened as regards the notion of life. It used to be thought that life was *generated* by a vital spirit. Modern biology redefined our notion of life. It now considers that life is not the kind of thing that is generated. Life is *enabled* by various physical and chemical mechanisms like respiration, reproduction, etc. It would be meaningless to say that any one or other such mechanism "generates" life. Life is a capacity that certain systems have to interact with the world. Experience is the same.

What are you spending most of your time thinking about these days, and what's next for you?

Modestly, since I think that consciousness is no longer a mystery, I'm trying to solve a problem that I think is currently a mystery, namely why humans seem to be able to understand things. Today's machine learning architectures can do good pattern classification if they are given mounds of examples, but they don't understand what they're doing. Humans seem to understand things... I'm trying to understand what it is to understand.

References

Aloimonos, J., Weiss, I., Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*. 1(4), 333-356.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723-742.

Braund, M. J. (2008). The structures of perception: An ecological perspective. *Kritike: An Online Journal of Philosophy*, *2*(1), 123-144.

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, *47*(1-3), 139-159.

Gibson, J. J. (1966). The senses considered as perceptual systems.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939-973.

Myin, E., & O'Regan, J. K. (2002). Perceptual consciousness, access to modality and skill theories. A way to naturalize phenomenology? *Journal of Consciousness Studies*, *9*(1), 27-46.

Noë, A., & O'Regan, J. K. (2002). On the brain-basis of visual consciousness: A sensorimotor account. *Vision and mind: Selected readings in the philosophy of perception*, 567-598.

O'Regan, J. K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford University Press.

O'Regan, J. K. (2014). The explanatory status of the sensorimotor approach to phenomenal consciousness, and its appeal to cognition. *Contemporary Sensorimotor Theory*, *15*, 23.

Thompson, E., & Varela, F. J. (2001). Radical embodiment: neural dynamics and consciousness. *Trends in cognitive sciences*, *5*(10), 418-425.

Varela, F. J., & Bourgine, P. (Eds.). (1992). *Toward a practice of autonomous systems: Proceedings of the First European Conference on Artificial Life*. MIT press.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind: Cognitive science and human experience*. MIT press.

Verbal hallucinations, intentionality, and interpersonal experience

An interview with Matthew Ratcliffe

By Mathieu Frerejouan

Citation: Ratcliffe, M. & Frerejouan, M. (2018). Verbal hallucinations, intentionality, and interpersonal experience. An interview with Matthew Ratcliffe. *ALIUS Bulletin*, 2, 95-107.

Matthew Ratcliffe
matthew.ratcliffe@univie.ac.at
Department of Philosophy
University of Vienna, Austria

Mathieu Frerejouan
mathieu.frerejouan-du-saint@univ-paris1.fr
Department of Philosophy
Pantheon-Sorbonne University, Paris, France

Hallucinations, and in particular "Auditory Verbal Hallucinations" (AVHs), have become a main topic of your recent work (Ratcliffe, 2017a, 2015b; Ratcliffe and Wilkinson, 2016). Your interest in psychopathology and the way it interacts with philosophy isn't new, your preceding field of expertise being "depression" (Ratcliffe, 2015a). However, one might say that there is still a gap between the subject of depression and hallucination, since the former is considered an affective disorder while the latter is traditionally conceived as a symptom of psychosis. What inspired you to study the topic of hallucinations? Is it somehow related to your work on depression?

I don't think we can (or, at least, should) draw a clear line between affective disorder and psychosis. Consistent with this, there is considerable common ground between my study of depression-experiences and my more recent work on hallucinations. Both reflect a wider interest in the phenomenology of feeling and the structure of interpersonal experience. In my 2008 book, *Feelings of Being*, I offered a detailed account of what I call "existential feeling". Existential feeling is an all-enveloping sense of reality and belonging, in the context of which more localized intentional experiences arise, such as perceiving *p*, remembering *q*, and thinking about *r*. Changes in existential feeling are expressed in a range of ways. For instance, people sometimes refer to feelings of familiarity, unfamiliarity, unreality, strangeness, detachment, being at home in the world, being estranged from everything, and so forth. I have proposed that these feelings consist, most centrally, in a sense of the various *types of significant possibility* offered by the surrounding world, something that is constituted by a range of diffuse, felt, bodily dispositions.

My 2015 book on depression is a more specifically focused case study, which seeks to demonstrate the applicability of this account to forms of depression-experience and, in the process, to develop the account in much more detail. In particular, my work

on depression emphasizes the centrality of interpersonal possibilities to existential feeling, along with how our experiences, thoughts, and activities are shaped and regulated by relations with other people. This is also a central theme of my 2017 book, *Real Hallucinations*. Consistent with my earlier work on existential feeling, the book maintains that seemingly localized experiences, of the kind that are often labeled as "hallucinations" and "delusions", tend to arise in the context of wider-ranging phenomenological disturbances involving the sense of reality. It builds on this earlier work by exploring the *anticipatory* structure of experience in more detail and also showing exactly how this structure is inextricable from the interpersonal.

> "Seemingly localized experiences, of the kind that are often labeled as 'hallucinations' and 'delusions', tend to arise in the context of wider-ranging phenomenological disturbances involving the sense of reality."

One of the reasons I ended up focusing on hallucinations and, more specifically, "auditory verbal hallucinations" (something of a misnomer, as will become clear) is that I became involved in a project called "Hearing the Voice", based at Durham University and funded by the Wellcome Trust. However, I address these experiences in order to make points that have much wider application. As with the topic of depression, hallucination is employed as a case study, through which I develop a more general philosophical account of the structure of human experience and the manner in which it depends on interpersonal relations.

The experiences you describe are presented as "real hallucinations", in contrast with what you call "philosophers' hallucinations". In fact, if the concept of hallucination plays a crucial role in philosophy of perception, it is mainly understood as a logical possibility relying on thought experiments. From this approach, a hallucination is an experience which is indistinguishable from a veridical perception, though without there being any physical object which is perceived (Macpherson, 2013). Why does this definition fail to make sense of what you call "real hallucinations"?

We can think of philosophers' hallucinations in two ways: (a) an experience that is phenomenologically identical to a perceptual experience of *p* in one or another modality, which occurs in the absence of *p*; (b) an experience that a person is unable to distinguish from a perceptual experience of *p*, which occurs in the absence of *p*. The latter is more permissive, as two experiences could turn out to be quite different in kind, even where the subject is constitutionally incapable of telling them apart.

Turning first to (a), it is pretty clear that real hallucinations are messier than philosophers' hallucinations—they are seldom, if ever, phenomenologically identical

to veridical perceptual experiences. However, a more interesting point is that they are often quite different in *kind*. In my 2017 book, I draw a distinction between the content of an experience and the sense that one is having an experience of that type. For instance, when you look at a cat, your experience has a certain content, "a big, white cat asleep on a chair". Along with this, there is a pre-reflective, ordinarily unproblematic sense of its being a perceptual experience (and, more specifically, a visual perceptual experience) of a cat, rather than an experience of remembering or imagining a cat. The question I begin by addressing is this: in virtue of what do I take myself to be perceiving something rather than, say, imagining or remembering it?

You might think that the answer is simple enough: the experience has a content that is specific to (visual) perception and is thus, in certain respects at least, distinct from an imagined or remembered content. Thus, the sense of being in one or another type of intentional state is to be identified with those aspects of experiential content that are unique to a state of that type. However, what I demonstrate through a detailed study of auditory verbal hallucinations (hereafter, AVHs) is that sense and content can come apart. Granted, some of those experiences labeled as AVHs do indeed seem to resemble, to some degree, hearing a voice emanating from the external environment, but in the absence of a speaker. However, many of them (probably the majority) are quite different. Voice-hearers often report that the "voice" is not experienced as originating outside of them, that it lacks some or all auditory qualities, and that it is different in kind from mundane perceptual experiences, auditory or otherwise. What we have here is the sense of perceiving something, arising in the absence of the usual sensory perceptual content. This is different from a philosopher's hallucination of type (a), given that the content of the "hallucination" is quite unlike that of an auditory perceptual experience.

In contrast to experiences like this, I maintain that certain other "hallucinations" have experiential contents that resemble those of veridical perceptions, while at the same time involving no sense of perceiving. Hence some "hallucinations" involve "content without sense"; others involve "sense without content"; and others fall somewhere between the two poles. Philosophers' hallucinations of type (a) fail to accommodate the relevant distinctions.

Type (b) philosophers' hallucinations fare better, insofar as they accommodate the possibility of failing to distinguish something from a perception even when the relevant experiences are quite different. In other words, one can have a sense of perceiving something in the absence of the usual content. But again, the reality is much messier. While content and sense are to be distinguished, content does at least make some contribution to sense. Consequently, when one has the sense of perceiving something, but without the usual perceptual content, that sense is partial,

incomplete. In addition, there is often a feeling of incongruity, tension. The relevant experience is immediately recognized as unusual, as involving a kind of intentionality that stands apart from imagining, perceiving, remembering, thinking in inner speech, and so forth.

We are owed an account of what this sense of being in an intentional state consists of, given that it is not exhausted by content. And this is something that I seek to provide in the book, by showing how the sense of being in a given type of intentional state is constituted largely by a cohesive, affectively-charged pattern of anticipation that is specific to a state of that type. I argue that various "hallucinations" arise due to localized disruptions of anticipatory patterns, and also that these disruptions generally occur in the context of less pronounced but wider-ranging and more enduring disturbances of the structure of intentionality. Some such experiences involve a sense of perceiving that is associated with a content of imagination, memory, or inner speech. Others involve a sense of perceiving that is not tied to an experiential content in another intentional modality. Both of these broad types of experience are often described in seemingly paradoxical ways, in terms of experiencing something as "there" and at the same time "not there", hearing something but not hearing it, and so on.

In your last book (Ratcliffe, 2017a) you often refer to authors such as Louis Sass (1994, 2014), Josef Parnas (2013) and Dan Zahavi (2007, 2014, 2017), who adopt a phenomenological approach to psychopathology. An assumption that you share with them is that localized symptoms, such as hallucinations, cannot be understood separately from more profound changes affecting our global experience of the world and ourselves. However, you also question this approach by noting that these alterations are mainly conceived as a "fragmentation from within", therefore neglecting how our "self" is embedded in interpersonal relations. Why, in your opinion, should phenomenological psychopathology not leave aside the interpersonal dimension of pathological states?

As you note, consistent with the spirit of phenomenological psychopathology, I maintain that various seemingly localized, anomalous experiences are actually symptomatic of wider changes in the structure or form of experience. So the disagreement addressed in my 2017 book is more specific in nature. A substantial body of recent work on the phenomenology of schizophrenia proposes that the various "symptoms" originate in a more fundamental disturbance of what is often referred to as "minimal self". One concern I have about such approaches is that they are often insufficiently critical of the schizophrenia construct. There is a tendency to insist on qualitative distinctions between experiences that are typical of schizophrenia and of other conditions, distinctions that are in many cases questionable. However, the main focus of my critique is on the claim that

schizophrenia originates in a disturbance of minimal self. There are two aspects to this critique.

First of all, I raise the concern that it is unclear what the relevant sense of "self" actually consists of. Proponents of the view maintain that every experience essentially has a perspectival structure, a sense of its originating in a singular locus of experience. This locus is not to be construed as a separate entity from which experiences emanate, as something that experiences presuppose, or as something that is recognized reflectively. Rather, it is integral to the structure of experience, inseparable from it, and grasped with a kind of phenomenological immediacy. But what, exactly, does it consist of—what more can be said? Repeated appeals to me-ness, mine-ness, what-it-is-like-for-me-ness, and the like do not really tell us very much. Thus, one of the things I try to do in the book is formulate a more specific and detailed account of what "minimal self" (construed phenomenologically) actually is. My proposal is that we identify minimal self with the *modal structure of intentionality*, by which I mean a pre-reflective sense of the various types of intentional state as distinct from one another—"perceiving" as distinct from "imagining", "imagining" from "remembering", etc. I offer various arguments for this move. For instance, if one could not distinguish perceiving from remembering and anticipating, one would lack any sense of temporal location. And, if one could not distinguish perceiving from imagining, one would similarly lack any sense of spatial location. Without any sense of spatial or temporal location, it is difficult to see how any kind of experiential self or perspectival structure could be retained.

> "My proposal is that we identify minimal self with the *modal structure of intentionality*, by which I mean a pre-reflective sense of the various types of intentional state as distinct from one another."

One could argue that the modal structure of intentionality is *necessary* for minimal self or make the stronger claim that it is also *sufficient*. While I am tempted towards the latter, I restrict myself to the claim that modal structure is *necessary* and also *central*. To further support this position, I argue that the various symptoms of "schizophrenia" attributed to self-disorder (such as certain types of AVHs) are best understood in terms of localized and wider-ranging changes in the modal structure of intentionality. Thus, if we want to attribute such experiences to disturbances of minimal self, we should identify minimal self with the modal structure of intentionality or at least concede that modal structure is essential to it. If one rejects this conclusion and insists that minimal self is something else altogether, perhaps something "even more minimal", then one should stop trying to account for

schizophrenia in terms of disordered minimal self. Given that the relevant symptoms originate in disturbances of the modal structure of intentionality, any appeal to an additional self-disturbance would be explanatorily redundant.

I am not sure whether or to what extent my account of minimal self and modal structure is shared by those who have written on self-disorder in schizophrenia. Dan Zahavi (2017) disagrees with me and wishes to insist that minimal self is something even more phenomenologically primitive. As for what others think, I look forward to finding out. But, if minimal self is supposed to be something else, then I honestly don't know what it is: appeals to a pre-reflective "what-it-is-like-for-me-ness" that is allegedly integral to all experience strike me as obscure.

So that is the first part of my critique. The second part concerns the relationship between minimal and interpersonal self. The literature on schizophrenia and self-disorder encompasses some subtly different accounts of the relationship between self-experience and interpersonal/social experience. Some of these differences need to be made clearer and more explicit. Even so, it is at least apparent that all of these accounts award self-disturbance some kind of priority over changes in how one experiences and relates to other people. For instance, Josef Parnas and several of his co-authors maintain that disturbances of intersubjectivity *presuppose* more fundamental forms of self-disturbance. They further suggest that the causes of self-disorder originate within the individual and are plausibly genetic (e.g., Raballo et al., 2009). In contrast, I think it likely that disturbances in the modal structure of intentionality have interpersonal/social causes, in many but not all instances. But my main point of disagreement concerns constitution rather than causation. In *Real Hallucinations*, I argue at length that the modal structure of intentionality is inextricable from interpersonal experience. Neither has priority over the other. Thus, regardless of how it might have been caused, a self-disturbance (construed as a certain kind of pronounced change in the modal structure of intentionality) is also a disturbance of interpersonal experience, and vice versa.

One might object that young infants plausibly have a basic sense of self before they are fully socialized. So, surely, minimal self comes first and the interpersonal comes only later. However, my claim is not that the modal structure of intentionality *must*, in all possible cases, depend on the interpersonal. Rather, the type of modal structure that we find in typical adult humans *does* happen to be interpersonally dependent. Social development does not involve adding more complicated capacities on top of a static, underlying, core sense of self. Rather, it is to be construed as a transformative process, a point that applies to development more generally. The modal structure of intentionality changes during development; an adult does not have the same *kinds* of intentional experiences as an infant. Development of the structure of intentionality is, if you like, *entrusted* to the social

world, such that it can be derailed in one or another way by certain interpersonal processes. Moreover, that structure is interpersonally and socially sustained even in adulthood. Hence a pronounced shift in how one relates to other people in general also amounts to a change in the modal structure of intentionality.

My overall account of the relationship between modal structure and interpersonal experience is lengthy, multi-faceted, and rather complicated. However, I will at least try to give a brief summary of some of the central points. I propose that the structure of experience centrally involves a kind of bodily, felt *anticipation*. Drawing on Husserl, Jaspers, and the later Wittgenstein, I argue that perceptual experience ordinarily incorporates a pervasive sense of confidence, certainty, or trust. As one interacts with one's surroundings, things are anticipated with varying degrees of determinacy and, on the whole, experience unfolds in ways that are in line with anticipation. This dynamic experience of confident anticipation and fulfilment is not localized; it is a cohesive, all-enveloping backdrop against which more localized experiences of potential and actual anomalies arise.

Inspired by themes in Husserl's later work, I develop an account of how the modal structure of intentionality depends on this backdrop of practical, perceptual confidence. I maintain that our sense of being rooted in a world, in a realm where we perceive *p*, remember *q*, and imagine *r*, and distinguish between experiences of these and other types, is constituted by this overarching background of confident, cohesive anticipation and fulfilment. Our sense of something *as perceived* involves its integration into the wider temporal structure. And our more general grasp of the distinctions between being the case, not the case, and possibly the case originates in and continues to depend upon this same aspect of experience.

Other forms of intentionality involve characteristic deviations from the anticipation-fulfilment structure of perception. Imagination, for instance, is comparatively unconstrained: a cat can turn into a horse and fly away without the same sense of anomaly. Memory is similarly unconstrained in certain respects but not in others. For instance, one can move around in time, but unlike when imagining, one cannot change the temporal order of events. I claim that these distinctive temporal patterns, and an appreciation of whether and how they depart from the style of practically engaged perceptual experience, are central to the sense of being in one rather than another type of intentional state.

There is much more to be said here, but the basic point is that the modal structure of intentionality depends on what we might call a non-localized *style of anticipation*. The next step in the argument is to show that this style is inextricable from one's anticipated and actual interactions with other people. There are, I show, all sorts of ways in which other people serve to sustain, repair, and disrupt the anticipatory style

of experience. Consider a world in which other people in general offer only one or another form of threat, a world where there is no prospect of felt interpersonal connection or of trusting relations. This would impact on a person's wider experience of the surrounding environment in many ways. Anticipated and actual interactions with other people more usually shape what is perceptually and practically salient to us, as well as the kind of significance that it has. Other people also play numerous roles in emotion regulation. In addition, the experienced world is shaped by a tapestry of projects and wider commitments, all of which depend for their integrity on the anticipation of certain kinds of interactions with other people. Without the prospect of such interactions, projects and associated frameworks of anticipation would be unsustainable. And, deprived of the more usual system of stable, habitual possibilities that draw one in and structure one's activities, one would be more likely to retreat from the social world, becoming increasingly passive. With that, there is a diminution of various activities that themselves lend structure and coherence to experience.

Once all of these effects are described in detail and added together, we come to see how certain changes in the interpersonal sphere, such as those characterized by pronounced social anxiety and loss of basic trust in others, add up to a world that is more generally lacking in structure, devoid of a more usual sense of confidence or certainty. With this, the modal structure of intentionality is to varying degrees and in different ways eroded. For instance, a perceptual world that is lacking in structure, riddled with doubt, no longer shaped by long-term projects and associated configurations of equipment, and divorced from practical activities, becomes closer in structure to certain imaginings.

> "Experiences that tend to be associated with the label 'schizophrenia' need to be placed in their interpersonal contexts and re-interpreted accordingly."

Types of experience along these and similar lines are, I suggest, consistent with various different psychiatric diagnoses. A global loss of trust in other people and an interpersonal world that offers only threat are associated with certain post-traumatic conditions. However, such phenomenological changes are equally consistent with a loss of taken-for-granted reality that phenomenological psychopathology regards as central to schizophrenia. I accept that the boundaries here are less clear-cut than they are often taken to be. Furthermore, there are no grounds for regarding disorders of self as somehow more basic than disorders of interpersonal relatedness. The two are inseparable and the relationship is one of mutual implication.

Effectively, what I end up doing in the book is steering a middle path between the self-disorder approach, which has the virtue of acknowledging how various symptoms depend on wider disturbances of experience, and various claims associated with the Hearing Voices Movement, to the effect that experiences that tend to be associated with the label "schizophrenia" need to be placed in their interpersonal contexts and re-interpreted accordingly.

There is a long-running debate in psychiatric literature concerning the nature of what Jules Baillarger named "psychic hallucinations" (Baillarger, 1846) and that we now sometimes call "verbal hallucinations". Often described as voices whose content is however similar to thoughts, philosophers as well as psychiatrists have repeatedly asked themselves whether these experiences should be classified as perceptions or thoughts. You propose another approach to this debate, defending that we must acknowledge "a way of experiencing, a kind of intentionality, that does not fit into established categories" (Ratcliffe, 2017a). If verbal hallucinations are neither full-blown perceptions nor thoughts, then how should we characterize them?

As I mentioned earlier, the label "AVH" encompasses a range of importantly different experiences. Some of these plausibly resemble—to varying degrees—veridical auditory experiences, but many others do not. What we have in these latter cases is a variably complete sense of perceiving something, which arises in the absence of the usual sensory perceptual content. The sense of perceiving might attach to a content of inner speech, to a memory, or to an imagining. In cases where the sense of perceiving attaches to an inner speech content, I suggest that the same experience can be described in either or both of two ways: as hearing a voice, or experiencing someone else's thoughts. In other words, a certain type of AVH is to be identified with "thought insertion". Such experiences may be personified to varying degrees, something that involves further input from imagination, memory, and narrative abilities.

So, what we have are various different experiences, all of which differ from mundane experiences of perceiving, thinking, and so forth. They involve a partial sense of being in one type of intentional state, associated with a content more typical of another type of intentional state. This adds up to a distinctive kind of experience, a *way of experiencing* that stands out as different from unproblematic instances of perceiving, thinking, and so forth. Anomalous experiences of these kinds generally occur against a backdrop of wider changes in the structure of intentionality. Nothing is experienced as "real" or "there" in quite the way it once was, thus rendering the person more susceptible to localized disturbances of intentionality that are more extreme in nature.

Beyond your interest in hallucinatory states, you insist on the fact that an analysis of these unusual experiences allows us to better understand how our experience of the world and ourselves is structured. For this reason, you present your work as a first step of a larger philosophical inquiry regarding our different intentional states types and the way they interact with one another. Have you planned to investigate another type of altered state in the future, or are you going to continue your study of hallucinations?

I continue to work on existential feeling, interpersonal experience, and the modal structure of intentionality. In conjunction with this, I still write on the phenomenology of depression and I will probably have a bit more to say about hallucinations too. I may also end up getting dragged further into debates concerning the existence and nature of minimal self. I don't want to, but it's proving irresistible—like a really nasty itch that you have to scratch, even though you know that doing so will only make it worse.

My next major project is likely to be on the nature of "grief", something that is complicated, multi-faceted, highly variable, poorly understood, and philosophically neglected. This will complement my work on depression, as the issue of when and how grief should be distinguished from depression remains unresolved. It will similarly complement my work on hallucination, given that "bereavement hallucinations", including "voices", are commonplace but again poorly understood. The topic of grief also fits in with my wider interest in emotions, feelings, and interpersonal relations.

> "Grief does not simply conclude at some point with 'letting go' of the deceased. Rather, people retain various different types of connection with the dead which continue to play important roles in their lives."

One of the things I want to do is explore in depth the many ways in which experience, thought, and activity are interpersonally regulated, something that is rendered particularly salient by bereavement. But what I'm most excited about here is the prospect of opening up a new area of social cognition research. To date, work on interpersonal experience, understanding, and interaction in philosophy and cognitive science has focused exclusively on our relations with the living. Yet, as the "continuing bonds" literature has convincingly shown, grief does not simply conclude at some point with "letting go" of the deceased, ceasing to relate to her. Rather, people retain various different types of connection with the dead, connections that can continue to play important roles in their lives. I'd like to widen

social cognition research to accommodate these relations in all their diversity (including their cultural diversity), and also to address how our relations with the living and the dead interact with each other. After that, I might try to tackle the topic of temporal experience, something that I've been working towards for a while now but still feel thoroughly intimidated by.

References

Baillarger, J. 1846. De l'influence à l'état intermédiaire à la veille et au sommeil sur la production et la marche des hallucinations. Paris: JB. Baillière.

Macpherson, F. 2013. The philosophy and psychology of hallucination: An introduction. In Hallucination: Philosophy and Psychology, ed. F. Macpherson and D. Platchias, 1-38. Cambridge, MA: MIT Press.

Parnas, J. 2013. On psychosis: Karl Jaspers and beyond. In One Century of Karl Jaspers' General Psychopathology, ed. G. Stanghellini and T. Fuchs, 208-228. Oxford: Oxford University Press.

Raballo, A., D. Sæbye, and J. Parnas. 2009. Looking at the Schizophrenia Spectrum Through the Prism of Self-disorders: An Empirical Study. Schizophrenia Bulletin 37: 344-351.

Ratcliffe, M. 2008. Feelings of Being: Phenomenology, Psychiatry, and the Sense of Reality. Oxford: Oxford University Press.

Ratcliffe, M. 2015a. Experiences of Depression: A Study in Phenomenology. Oxford: Oxford University Press.

Ratcliffe, M. 2015b. How is Perceptual Experience Possible? The Phenomenology of Presence and the Nature of Hallucination. In T. Breyer and M. Doyon, eds. Normativity in Perception: Phenomenological, Analytic and Psychopathological Perspectives. Basingstoke: Palgrave Macmillan: 91-113.

Ratcliffe, M. 2017a. Real Hallucinations: Psychiatric Illness, Intentionality, and the Interpersonal World. Cambridge MA: MIT Press.

Ratcliffe, M. 2017b. Grief and the Unity of Emotion. Midwest Studies in Philosophy 41: 154-174

Ratcliffe, M. and S. Wilkinson. 2016. How Anxiety Induces Verbal Hallucinations. Consciousness & Cognition 39: 48-58.

Sass, L.A. 1994. The Paradoxes of Delusion: Wittgenstein, Schreber, and the Schizophrenic Mind. Ithaca: Cornell University Press.

Sass, L.A. 2014. Delusion and double bookkeeping. In Karl Jaspers' Philosophy and Psychopathology, ed. T. Fuchs, T. Breyer, and C. Mundt, 125-147. New York: Springer.

Zahavi, D. 2014. Self and Other: Exploring Subjectivity, Empathy, and Shame. Oxford: Oxford University Press.

Zahavi, D. 2017. Thin, Thinner, Thinnest. In C. Durt, T. Fuchs, and C. Tewes, eds. Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World. Cambridge, MA: MIT Press: 193-199.

Conceptual, anthropological and cognitive issues surrounding religious experience

An interview with Ann Taves

By Martin Fortier & Maddalena Canna

Ann Taves
taves@religion.ucsb.edu
Department of Religious Studies
University of California, Santa Barbara, USA

Martin Fortier
martin.fortier@ens.fr
Institut Jean Nicod
ENS/EHESS, Paris, France

Maddalena Canna
maddalena.canna@ehess.fr
Laboratoire d'Anthropologie Sociale
EHESS/Collège de France, Paris, France

Citation: Taves, A., Fortier, M. & Canna, M. (2018). Conceptual, anthropological and cognitive issues surrounding religious experience. An interview with Ann Taves. *ALIUS Bulletin*, 2, 109-128.

A. T. wishes to thank Egil Asprem for his helpful comments on the first draft of this interview.

You have proposed to conceive of religion as special *paths* to a *goal* (Taves, 2009: 66). If we look at the diversity of religious experience, we notice that, in most cases, the notion of *path* is associated with a rather conflicting notion of *immediate* and/or *inherent* presence of the religious in humans. Various traditions suggest that no path is needed, and that the risk of identifying spiritual practice with a sequence of goal-oriented actions must be avoided by any means. For example, in the Dzogchen tradition, the aspirant can follow a path only under the condition that she is aware that the desired state already pre-exists in herself (Norbu & Shane 1986). Similar conceptions can be found in some Hindu traditions (e.g., Hughes, 1994). In the Christian world, the debate about the notion of grace, which can be defined as a quality independent from human action, shows how gradualist (path-related) and immediatist conceptions of religious experience are often intertwined and complement one another.

In the Buddhist tradition, the debate between subitism and gradualism presents a similar conundrum. In subitism, enlightenment is taken to be attainable all at once, whereas in gradualism, it is said to be achieved only through arduous improvement (Gregory, 1991; Faure, 1991). Similar debates are also present among the different branches of Kashmir Shivaism (Padoux, 2017).

Thus, the notion of *path* seems particularly relevant to account for the gradualist approach to religion, where *events* play a crucial role. On the other hand, the notion of *self-recognition* might seem more relevant for the immediatist approach, where

events play a more peripheral role.

How do you conceive of the interplay between gradual and immediate ingredients of religious practice and experience? What is the relation between shifts in self-awareness/self-recognition and special events? Is self-recognition conceivable as a particular kind of special event or do we need slightly different conceptual tools to analyze it? Even more broadly, what does this tension between the gradualist and the immediatist take on religion tell us about the emergence of religion itself?

I think the key thing to notice here is the difference between relying on a path schema and claiming that following a path (a gradualist approach) will get you where the tradition says you want to go. The path schema (or something like it, e.g., "a way") seems to play a role in all the examples you give. When I suggested (Taves, 2010, p. 175) that religions, spiritualities, and philosophies are often organized around path schemas, I indicated that they could use the path schema to assess, rank, manipulate, and sometimes transcend things that matter. I didn't elaborate on the transcending idea, but that is what I think is at play in the more immediatist take on paths. It is only by starting on the path with its postulated goal that people ask the question (how do I get there) that the immediatists are answering (by saying they have already arrived). If they don't ask the question, then it is pointless to tell them that they already have the answer. Traditions often encourage people to start out on the path (i.e., do some sort of practice), but warn them that the path will not get them to the goal. Ultimately, the immediatists say you have to transcend the path or recognize that there "is no path" or that it is all about grace and nothing you can do will get you there, etc.

> "Meditating on paradoxes can, I think, trigger shifts in self-awareness, which I would conceptualize as an 'internal event'."

To speak to one of your later questions, I think that traditions often use the path schema to set up a paradox, e.g., the path is "no path". Meditating on paradoxes can, I think, trigger shifts in self-awareness, which I would conceptualize as an "internal event". In other words, I think that the tension between gradualist (practice will get you there) and immediatist (only direct insight will get you there) approaches can be used to set up a paradox and that meditation on paradoxes can trigger shifts in self-awareness and deep insights of the sort that immediatists often seek to attain. But people must want to get "there"—even if "there" is the insight that there is no "there"—for either approach to "work".

You argue that enhanced vividness of sensory content causes people to take what they experience as more real (Taves, 2009, p. 159-160). Do you consider sensory

vividness and attribution of reality to always be correlated?

Tantric visualizations aim at enhancing the vividness of imagination (Kozhevnikov, Louchakova, Josipovic, & Motes, 2009) and yet their function is precisely to induce a sense of decreased reality (to become aware that any sensation is nothing but the non-real product of the mind) (Beyer, 1973). This seems to be a clear counterexample challenging the putative correlation between vividness and attribution of reality.

Another example of dissociation between the sense of reality and the vividness of the sensory content is provided by the comparative study of hallucinations: the sense of reality of psychedelics-induced hallucinations, deliriant-induced hallucinations and Charles Bonnet hallucinations can vary to a great extent, whereas their respective sensory content remains quite stable (Fortier, Forthcoming). Namely, regardless of the vividness of the visual hallucinations, deliriant-induced hallucinations always have a high sense of reality, psychedelics-induced hallucinations a moderate sense of reality and Charles Bonnet hallucinations no sense of reality.

According to you, what are the key factors contributing to the attribution of reality to experience in general and to religious experience in particular?

I would view vividness and reality as contingent. I think that "paths" assert different conceptions of reality. As animals, I think we have evolved mechanisms (our senses) to determine what is real for us as human animals. Our "natural" or "default" sense of what is real is grounded in the way we have evolved to process input from our bodies and our environment. Various practices and neurological conditions, e.g., Tantric visualizations, psychedelics, or Charles Bonnet syndrome, can all mess with our natural or default sense of what is real. The paths may do this deliberately and then inform us of how we are to appraise what we have experienced.

You have recently proposed that the Building-Block Approach (BBA) to religion (Taves, 2009; Taves, 2015) could be fruitfully combined with the Predictive Coding Framework (PCF) (Taves & Asprem, 2017). Within the PCF, perception and cognition are said to result from a bidirectional tradeoff between top-down predictions and bottom-up prediction errors. Rephrased in the terms of the PCF, the BBA would thus amount to saying that religious experiences and concepts are constructs resulting from the combination of interpretation (top-down processing) and sensory information (bottom-up processing).

Let me interrupt the question at this point to clarify our understanding of a few key terms.

Building Blocks: When we refer to "building blocks" in the context of the BBA, we are using building blocks as a synonym for the components or parts that interact to

produce a phenomenon of interest. These components interact to form mechanisms that produce or maintain the phenomenon (Asprem & Taves, 2016). Mechanisms are nested in stacks and the phenomena of interest can be specified at any given level.

Religion: We view "religion" as a complex cultural concept, i.e., as an abstract noun with unstable, overlapping meanings that vary within and across cultures and social formations (Asprem & Taves, 2016). We refrain from defining religion, so that we can study the way others use it and related terms.

Experience: The flow of information in so far as we are aware of it. *An experience*: an internal event that we have segmented out of the flow of information and thus cognize as "an event." The process of segmenting and cognizing (known as "event cognition") takes place below the threshold of consciousness. We can recount these events and consciously reappraise them after the fact. Narrating and reappraising, however, are new events, which are also cognized subpersonally (Taves & Asprem, 2017). Attribution theories typically focus on conscious (post-hoc) appraisals of events.

One possible worry is that the BBA and the PCF are in fact incompatible. Indeed, the BBA is mainly interested in the *cognitive* interpretation and appraisal of special events or experiences. It states that a given percept (i.e., a special experience) can be interpreted in various ways depending on one's own expectations and predictions. Now, arguably, this is not what the PCF is mainly concerned with. When the PCF acknowledges the importance of the "interpretation" of sensory data by prior expectations, it means it in a very different way from what the BBA has in mind: interpretation, in the PCF, does not refer to how percepts are cognitively interpreted but how percepts are generated in the first place. In other words, the bidirectional process the PCF is chiefly discussing concerns the production of percepts, whereas the bidirectional process the BBA is focusing on concerns the interpretation of an already-constituted percept.

To make this point even clearer, let us consider some examples. Case 1: seeing a coffee cup on the table. Case 2: seeing a face in the foam of a coffee cup. One can be tempted to say that Case 1 and Case 2 both involve some kind of interpretation (some kind of top-down processing). In Case 1, perceptual priors make the brain interpret the sparse sensory data collected by the retina as being a coffee cup. In Case 2, the percept—the foam in the coffee cup—is interpreted as containing a face in virtue of prior experience and gestaltic skills. It must be stressed, however, that two kinds of "interpretation" are here at play: in Case 1, priors contribute to the generation of a percept; in Case 2, priors contribute to the cognitive sense making of a percept which is already there. So, in Case 1, "interpretation" refers to the influence of strictly subpersonal *perceptual* priors, whereas in Case 2, "interpretation" refers to *cognitive* priors which may be either personal or subpersonal.

Importantly, studies looking at the role of priors in the brain mainly focus on cases similar to Case 1 above. For example, when the light-from-above prior makes one see a circle as convex instead of concave, the output of the process is a percept and not the interpretation of a percept (e.g., Sun & Perona, 1998). By the same token, when, in a binocular rivalry task, the interplay between predictions and prediction errors results in a switch of percept—a house, then a face, then a house, then a face—, the type of top-down processing at work generates distinct percepts and not distinct interpretations of a percept (e.g., Hohwy, Roepstorff, & Friston, 2008).

By contrast, the type of top-down processing the BBA is interested in is not how distinct percepts are generated depending on one's internal models, as much as it is in how a given percept is cognitively interpreted. For example, given that a coffee cup is in front of me, how will my internal models influence my interpretation of the foam of the coffee cup? While the PCF literature dedicated to the study of perceptual priors is extensive (e.g., Knill & Richards, 1996), to our knowledge, that which is dedicated to the study of top-down processing in the interpretation of sensory stimuli is extremely reduced.

Do you agree with the above conceptual distinction between BBA and PCF? If so, do you think the project of unifying BBA and PCF can nonetheless be successfully achieved?

I am not as certain as you are that we can make a sharp distinction between "cognitive" and "perceptual" priors and thus between "interpretation" and "perception." This is the issue researchers are discussing under the heading of cognitive penetrability (Zeimbekis & Raftopoulos, 2015). Some, as you seem to assume, view the formation of percepts as cognitively impenetrable; others disagree. The central question is at what point specific predictions (appraisals derived from cultural schemas) come into play in terms of processing. Language offers an example that likely blurs the distinction between your Case 1 and Case 2. Based on the language(s) we learn as a child and the neuronal pruning that takes place in light of that specific language, there are population specific differences in what we hear (Roepstorff, Niewöhner, & Beck, 2010). Japanese speakers, for example, don't hear the difference between "r" and "l". I would assume that the formation of a percept in this sense (hearing or not hearing the difference) would take place at a lower level of processing than cognizing an event of the sort Asprem and I were discussing. Regardless of how this issue is resolved—and it appears to be a focus of much research—I think that we can integrate the BBA into a predictive processing framework if we think in terms of predictions (aka appraisals) at different levels of mechanisms within a processing hierarchy. In Taves and Asprem (2017), we are considering predictions as appraisals and the error monitoring process as an appraisal process. Schemas, which may be evolved or learned, inform event models, which in turn generate predictions.

The key thing to recognize is that event cognition is a subpersonal process. We focus on it, because it identifies the components that interact to produce the phenomenon of interest to us as humanists: events that surface to awareness, including internal events, and, thus, the events humans are inclined to turn into narratives of "experiences". But just because the event surfaces to awareness does not mean that we are aware of the components that interact to produce the experience. Your point that *cognitive* priors, as you call them, may be either personal or subpersonal is an important one. We are assuming that learned (and thus population specific) priors may be internalized to the point where they shape our experience below the threshold of awareness. If some learned priors, such as language, effect changes in the brain such that we cannot perceive things in any other way that would suggest that these learned priors can operate much more deeply than we might suspect.

I will be interested in following research that explores the role of population specific differences at those lower levels of processing. I do think that we will be able to unify the BBA and PCF in a more precise fashion at some point, but for now the key point is that processing involves multiple levels of mechanisms and that population specific processes (e.g., cultural schemas) are clearly implicated at subpersonal levels that generate differences in how we appraise events at levels that are not accessible to us.

The above distinction between two types of "interpretation" or "top-down processing" has important consequences when it comes to the debate between inherentism and attributionism. Inherentism claims that some experiences are intrinsically religious, whereas attributionism contends that the religious character of experience is derived from some cognitive ascription or attribution.

Importantly, the distinction between perceptual priors and cognitive priors leads to the identification of two types of inherentism:

- *Strong inherentism* is the view that the religious character of religious experience does not result from any top-down processing. According to this view, religious experience would be solely determined by bottom-up processing. Prior knowledge would play no role whatsoever in determining the content of experience. Strong inherentism is blatantly inconsistent with PCF.

- *Weak inherentism* acknowledges that experience is partly determined by prior knowledge. According to weak inherentism, experience is the result of a tradeoff between prior expectation and collection of sensory data. This view is perfectly consistent with the PCF. It still qualifies as a version of inherentism because it denies that *cognitive* priors shape religious experience. It acknowledges that religious experience—be it exteroceptive, sensorimotor, or interoceptive—is determined by priors but these priors are non-cognitive: they determine what an experiential content will be and not how an experiential content will be interpreted.

- *Attributionism* assumes that the religious character of experience results from the cognitive interpretation—by System 1 (fast/unconscious thinking) or System 2 (slow/conscious thinking)—of experience. Weak inherentism and attributionism both agree that prior knowledge determines the religious character of religious experience but the type of prior knowledge the former is interested in—i.e., exteroceptive, sensorimotor, and interoceptive priors—differs from the type of prior knowledge the latter is interested in—i.e., cognitive priors.

It is worth noting that weak inherentism is compatible with the rejection of perennialism (i.e., the view that religious experience is to a large extent the same in every culture): since the content of experience is determined by non-cognitive priors and since these priors are to a large extent culture-dependent (e.g., Adams, Graf, & Ernst, 2004), various types of religious experiences can be triggered across cultures (each culture-specific prior triggers a specific type of intrinsically religious experience). Weak inherentism and attributionism are both compatible with the diversity of religious experiences, but for quite distinct reasons. The former implies that the diversity of experiences stems from the plurality of experiential (exteroceptive, proprioceptive and interoceptive) priors, whereas the latter maintains that the diversity of religious experiences stems from the plurality of cognitive priors.

In other terms, the enculturation of religious experience can take different forms. Unlike strong inherentism, weak inherentism and attributionism both recognize that culture influences the content of religious experience. But while classical attributionism focuses on the enculturation of religious experience through cognition, weak inherentism emphasizes the possibility for religious experience to be encultured through exteroception, proprioception and interoception. (See summary in Table 1).

In your discussion of the theories of religious experience, you convincingly show how *strong inherentism* is incompatible with the empirical data provided by proponents of the PCF (Taves & Asprem, 2017). But, it seems, you do not consider the possibility of there being a type of inherentism perfectly compatible with the PCF—namely, *weak inherentism*.

What would be your main objections against weak inherentism? Would you be tempted to endorse a hybrid theory combining weak inherentism and attributionism or to stick to attributionism and reject any form inherentism?

I think your terminology captures the subjective distinction between experiences that we consciously appraise and those, such as your experience of a feminine presence (discussed below), that seem to be as real as the cup. But as my comments on the previous question suggest, I don't think that the distinctions work well from a scientific perspective. Part of the problem has to do with terminology generated by various lines of research in sociology (on framing theory (Goffman, 1974; Snow,

	involvement of bottom-up processing?	involvement of top-down processing (exteroceptive, proprioceptive and interoceptive priors)?	involvement of top-down processing (cognitive priors)?	"religious" in virtue of…	enculturatuion of religious experience?
strong inherentism	yes	no	no	sensory (naturalistic version) or extrasensory (supernaturalistic version) data = inherentism	no = perennialism
weak inherentism	yes	yes	no	sensory data and non-cognitive priors = inherentism	yes (exteroceptive, proprioceptive, interoceptive) = pluralism
attributionism	yes	(yes)	yes	sensory data and cognitive priors = attributionism	yes (cognitive) = pluralism

Table 1: Strong inherentism, weak inherentism and attributionism

Rochford, Worden, & Benford, 1986; Johnston & Noakes, 2005)), social psychology (on attributional theory (Heider, 1958; Malle, 2004; Proudfoot & Shaver, 1975; Spilka, Shaver, & Kirkpatrick, 1985)), and emotion and stress research (on appraisal theory (Arnold, 1960; Lazarus & Folkman, 1984; Scherer, Schorr, & Johnstone, 2001)). Asprem and I have taken to using the language of appraisal processes as an umbrella term to encompass all these lines of research, which taken together focus on different levels and components involved in processing sensory data.

As already indicated, we view the "predictions" generated through predictive processing as subpersonal appraisals and we are open to the possibility that culture in the form of learned content (e.g., language) shapes what we perceive at a very low level of processing. So again, I hesitate to embrace your clearcut distinction between perceptual and cognitive processing. We tend to use cognitive not to refer to conscious (the ordinary language usage) but as a language used to describe all mental processing (as is common among cognitive scientists), in which case the distinction between perceptual and cognitive processing doesn't make sense.

"Asprem and I view the 'predictions' generated through predictive processing as subpersonal appraisals and we are open to the possibility that culture in the form of learned content (e.g., language) shapes what we perceive at a very low level of processing."

Arguments developed in your book (Taves, 2009) as well as by other authors (e.g., Proudfoot, 1985) have forcefully made the case for at least some degree of attributionism. It now seems difficult, and somewhat naïve, to defend classical versions of inherentism. Yet, there might arguably be some room for revised (and more sophisticated) versions of inherentism. One of them would be weak inherentism (as defined above). But even that version might still be too strong. Another and more promising version would be a blending of attributionism and weak inherentism. It seems that at least some religious experiences vindicate inherentism. Here is such an example: encounters with the ayahuasca spirit.

I (MF) have tried ayahuasca multiple times. While experiencing the effects of the hallucinogenic brew on my brain and body, I was bearing in mind the debates about the nature of religious experience. One thing that struck me was how easy it proved to discriminate between cases in which the experience of the ayahuasca spirit was the result of an attribution and those in which there was no apparent attribution whatsoever. In the first case, I would have some unusual cognitive, bodily or perceptual experience and interpret it with the local Shipibo-Konibo schema: these unusual experiences must be, I inferred, the ayahuasca spirit (*oni ibo yoshin*) who is

teaching me something, curing my body, showing me images, etc. In those cases, it was obvious that what I was experiencing was only an unusual experience *cognitively interpreted* as being an ayahuasca spirit. In another context, in another culture, I would certainly have cognitively interpreted the same unusual experience differently (as being God, a demon, an elf, etc.).

Here, as far as I can tell, you are simply distinguishing between a conscious appraisal of cues and cues that are appraised below the threshold of consciousness, which is where most such processing takes place.

Yes, but importantly, there were other cases in which it seemed to me that no interpretation was involved. To make an analogy: when you are walking in the forest, and you hear a whistle, you may interpret the whistle as the presence of a human person. In such a case, the "perception" of that person requires some cognitive interpretative process. By contrast, when you are having a coffee with someone in a café, the experience of that person does not seem to require any interpretation. The experience that you are then having *inherently* features a human person. Now, this is precisely what I felt on some exceptional occasions: the ayahuasca spirit was there, as obviously as when I am having a coffee with a friend; her presence was completely compelling and undeniable. She was a feminine presence; she was inhabiting an other-world and yet her world was closely interconnected with my everyday world; she was at the same time authoritative, benevolent, mischievous and incredibly witty; somehow, she felt very familiar; and she was exceedingly powerful (I felt she had the power to change all my beliefs with a single snap of the fingers).

Perhaps there is nothing very conclusive in such an experience. As you have thoroughly demonstrated, proponents of inherentism have often adduced compelling personal experiences in favor of inherentism, but such personal reports systematically fail to address the possibility that the processes of attribution at play are implicit and therefore subjectively undetectable. One worry with this line of objection, however, is that it would lead one to say that in the meeting with a friend in a café, some attribution is also at work.

Yes, from a predictive processing perspective, the cognizing of meeting in a cafe event would rely on a "meeting a friend in a café" schema, which would rely sub-schemas related to "friend" and "café," and on lower level processing (e.g., facial recognition or other cues) to recognize your friend and to navigate your way around the café and so on.

Now, if attribution is so widespread, then it loses its analytic interest: it misses the key distinction between *experiencing* spirits and *cognitively inferring* that spirits are there.

This is one of the places where your distinction between experiencing and

cognitively inferring is quite confusing. I would rephrase the issue as follows: appraisal processes take place at multiple levels of processing, both above and below the threshold of awareness. In order to understand what people experience, we need to specify the level at which appraisals are taking place. When we say we "experience a spirit" rather than "inferring the presence of spirits", we are making a distinction between an experience that surfaces to consciousness more fully formed in the former case than in the latter. If we look at the components that interact below the threshold of consciousness to give rise to "an experience", it is clear (see Figure 2 in Taves & Asprem, 2017) that learned (event) schemas can operate below the threshold of consciousness to predict what is happening, i.e., that we are interacting with a spirit. If we don't have a schema that can predict what is happening, the experience would surface as a set of cues that we have to consciously interpret, i.e., *infer* there is a spirit.

The inherentist and attributionist models make very straightforward empirical predictions about the nature of the ayahuasca experience. The attributionist insists that interpretations can take place implicitly. As a consequence, people might fail to detect ongoing interpretative processes. The way scholars could detect these implicit interpretations would be by analyzing the phenomenology of these spirit encounters across cultures. If the attributionist is right, then she should find that, in some cultures and contexts, people report the ayahuasca spirit to be feminine while others report it to be masculine; that some report it to be talkative and provider of existential advice while others report it to be secretive; etc. By contrast, the inherentist's prediction would be that some phenomenological features of the ayahuasca spirit are culture-independent: they do not vary across cultures.

Ayahuasca has now become a world-wide phenomenon. People are drinking it across the world. In addition to Westerners, many cultures and subcultures are starting to experiment with this hallucinogenic decoction (including, as surprising as it may sound, Iranian Sufis (Rooks, 2014)). Some of these groups of drinkers are not exposed to the narratives circulating about the ayahuasca experience. As such, these groups provide ideal participants to adjudicate the inherentism vs. attributionism debate. If, without being exposed to the emerging ayahuasca world culture and its narratives, these participants' phenomenological descriptions of the ayahuasca spirit perfectly fit with those provided by Amazonian indigenous people and Westerners, then there might be something intrinsically present in those experiences that vindicates some form of inherentism. Preliminary—and yet unpublished—data suggest that the features of the ayahuasca spirit are indeed strikingly consistent across cultures. But much more investigation needs to be done.

I think this sounds like a good way to test *at what level* the percepts are formed and to what extent they are shaped by culture. I think that it is quite possible that ayahuasca and other hallucinogens generate experiences that have some features in common across cultures. I have been following the research with psilocybin where

there do seem to be features that are commonly elicited, but the researchers are quite clear that set and setting play an important role in how the experience itself unfolds in real time (Griffiths et al., 2006, 2011). (In fact, Timothy Leary coined the set and setting distinction in the context of the Harvard Psilocybin Project in the early sixties (Hartogsohn, 2017; Zinberg, 1986).) The importance of set and setting is particularly apparent if we recognize that researchers are using psilocybin to model both "mystical" and "psychotic" experiences (see, e.g., Gonzalez-Maeso & Sealfon, 2009; Vollenweider & Kometer, 2010)! In the former case, the researchers go to great lengths to produce a positive experience by controlling set and setting. Insofar as they limit themselves to administering the mysticism scales (e.g., the MEQ 30), they only ask participants to report on those aspects of the experience that fit with it the scale's definition of "mystical."

If proven right, the form of inherentism sketched above would not vindicate perennialism. Indeed, it would imply that some brain patterns can lead one to perceive a specific spirit presence, but that another—induced, for example, by another hallucinogenic chemical compound—would trigger a different kind of religious experience (e.g., the presence of a spirit exhibiting other features). This revised version of inherentism would reject the classical neurotheological quest for a single "God-spot" (e.g., D'Aquili & Newberg, 1999). According the revamped version of inherentism, there would be dozens of spirit experiences mapping with dozens of distinct neural correlates—the putative single "God-spot" would thereby be replaced by various "spirits-patterns" or "gods-patterns". Within this hypothetical model, the diversity of religious experiences would be duly acknowledged, but it would remain that the phenomenological features—and the neurobiological signature—of each of these religious experiences would be invariant across cultures—making them *intrinsically* religious and not religious in virtue of some *attribution*.

Although I think that it is possible that experiences of presence are triggered by some hallucinogens, I do not think you can refer to them as "religious experiences", since experiences of presence can be appraised in many different ways (see Barnby & Bell, 2017). The determination of who the presence is and the nature of the interaction is, I would think, very much determined by set and setting (as well as lower level appraisal processes that may tend to skew the experience in a positive or negative direction (Underwood, Kumari, & Peters, 2016a, 2016b)). In your own case, you report a "feminine presence" with certain attributes, which you then characterize as the ayahuasca spirit; this latter characterization is surely a conscious appraisal. The experience—phenomenologically speaking—is of a particular sort of presence that fits with your expectations regarding the "ayahuasca spirit". If naïve subjects were given ayahuasca without knowing what drug they were taking in a neutral (or medical) environment, I suspect that their experience of a presence, if they had one, might be quite different and definitely would not be recognized as the

"ayahuasca spirit".

Yes, but if compelling evidence was to support the existence of invariant phenomenological features of the "ayahuasca spirit" across cultures, could it somewhat challenge the attributionist thesis?

I am making a fairly subtle distinction here, which you may or may not find meaningful, between what is experienced phenomenologically and how it is consciously appraised. I think it is possible that ayahuasca sometimes triggers cues that are appraised below the threshold of consciousness as a sensed presence. As indicated above, I do not think that naïve users would recognize the presence as the "ayahuasca spirit." Their *conscious* appraisals would reflect their own prior experience and the setting in which they received the drug.

You have suggested that events triggering disruption of the ordinary sense of reality foster the search for new meanings and predispose subjects to adopt new religious beliefs (Taves 2009, p. 100). Paradox is by definition a cognitive contradiction which, to be managed, requires one to reframe the context of its appraisal. Thus, paradox works as a trigger of new cognitive solutions.

Data collected on my own (MC) fieldsite tend to confirm the link between paradoxical mental activity and Altered States of Consciousness (ASCs) (deemed) spiritual. Mental paradoxical practices among Miskitos lead to dissociative hallucinatory seizures attributed to spirit attacks (Canna, 2017). Most Miskitos experiencing spirit attacks have been exposed to contradictory injunctions. They are told "not to think" about the spirit and yet at the same time are incited to pay attention to any possible manifestation of Him (*kaiki bas*, *lukpara*, literally "be careful not to care"). This insoluble contradiction triggers a paradoxical state, dissociative symptoms (trance, somatizations, conversions), and hallucinations. In most cases, the mental image of the spirit is appraised as impossible to "expel from one's head" (in the words of my interlocutors) and this loss of control is re-appraised as a proof of the autonomy of the image, and as a consequence, of the spirit's reality.

These data are consistent with William Barnard's self-account of his spiritual experience (Barnard, 1997; Taves, 2009, pp. 102-119). As you underlined, Barnard's experience was triggered by paradoxical mental activity. First, Barnard become obsessed with the idea of what would happen to him after death. He tried to imagine himself as a not existing being, and thus experienced a paradox described as follows: "I kept trying, without success, to envision a simple blank nothingness (the visualization exercise generated a paradox—that is, asking self to imagine self not being able to imagine)" (in Taves, 2009, p. 109). This paradox fostered the emergence of an ASC: "Suddenly, without warning, something shifted inside. I felt lifted outside myself, as if I had been expanded beyond my previous sense of self" (in Taves, 2009, p. 110).

To what extent do you think that the notion of paradox can explain the emergence of different states of consciousness (deemed) spiritual? Paradox is also pervasively present in many religious traditions (e.g., Christian notion of Trinity). To what extent do you think paradox is a typical feature of religious experience?

I think your findings and insights on this are very intriguing. Rather than say that paradox is a typical feature of religious experience, I would prefer to say that there are a number of traditions that *use paradox* to generate the insights that they claim reveal Truth or Reality (deliberately capitalized). I was thinking about this when I answered the paths question above (see first question).

On a more methodological note, you have championed a highly directive method for conducting post-hoc elicitation interviews (Taves, 2009; Taves & Asprem, 2017). For example, you propose to interrupt the narrative flow by asking the narrator w-questions such as "what" and "why", in order to distinguish the event model she has in mind (i.e., her sense of "what happened") from the explanation she provides (i.e., her explanation as to "why it happened").

Directive interviewing methods have both upsides and downsides. A possible downside is that by interrupting the interviewee's narration, directive methods lose some valuable information: it prevents the interviewee to re-enact her experience and access pieces of experience that she would not be able to recollect otherwise (Petitmengin & Lachaux, 2013; Petitmengin, Remillieux, Cahour, & Carter-Thomas, 2013).

Very directive and less directive methods tend to produce substantially different data. What is your take on the advantages and disadvantages of the distinct kinds of interviewing method?

Asprem and I were writing as historians not ethnographers, so we weren't actually offering a method for conducting "post-hoc elicitation interviews", but rather offering a method for analyzing narratives left to us by dead people, where interrupting the flow the narrative is not an issue! I think your observation is spot on when it comes to interviews of living people and, as far as I am aware, Claire Petitmengin has developed the most sophisticated method for teasing out the nuances of experiences from living subjects.

You have suggested that some cognitive impairments make a subject more prone to religious experience (Taves & Asprem, 2017). Studies by Tanya Luhrmann et al. (2015) as well as Rebecca Seligman and Lawrence Kirmayer (2008) suggest that phenomenologically similar experiences can be experienced as pathological or not depending on the social context. If we agree with these authors, a crucial anthropological question is to establish to what extent the opposition between pathology and health is arbitrary. Namely, boundaries between pathological and altered states of consciousness (deemed) religious seem particularly porous.

Suffering does not seem to be a distinctive feature of pathological experiences as opposed to religious ones. Indeed, in several shamanic traditions, the shaman's relationship to spirits is experienced as a form of painful oppression or afflicting persecution (Stépanoff, 2015; Canna, 2017). The possibility of controlling spiritual manifestations does not seem to be a critical feature of pathology either (e.g., Halloy, 2015). To what extent does the adoption of socially shared frameworks make anomalous experiences acceptable and non-pathological? How do you conceive of the boundaries between pathology, normality and election?

This is a great question and one that requires further research. Appraisal processes clearly make a difference, but it is not clear how much of a difference and for whom. Psychosis researchers are working to answer this precise question (see the work of Underwood, Kumari, & Peters, 2016a, 2016b).

> "I think we can assume that 'religion' and other innovations rely on some sort of new idea or insight that provides some sort of competitive advantage. In that sense, all innovations rely on some sort of experience."

While some scholars—classical (Tylor, 1871) or contemporary (Bulkeley, 2016)—argue that religion originates in anomalous experience such as dreaming, proponents of the cognitive science of religion contend that representations produced by our intuitive ontologies—rather than experience—are sufficient to explain the emergence of religion (e.g., Boyer, 2001). By insisting that there are no intrinsically religious experiences and that experience always has to be processed and interpreted by cognitive schemas to be deemed religious, you seem to reject the classical Tylorian view: *experience alone is not sufficient for religion to emerge*.

Now, importantly, there are two ways of rejecting the hypothesis that the origin of religion lies in experience:

(1) Weak rejection: experience is not sufficient but is necessary for religion to emerge.

(2) Strong rejection: experience is not sufficient nor necessary for religion to emerge.

What is your view on the origin of religion? Do you think that the strong rejection of the Tylorian view is a tenable position? Alternatively, do you lean towards the view that experience is not sufficient but indeed necessary?

The answer to this depends on what we mean by "religion" and "experience". If we define "experience" as an internal event that segmented out of the flow of

information of which we are aware, then any idea or insight is an "experience". If we assume that, whatever we mean by "religion" and however many times and places it emerges, it is—as an emergent phenomenon—an innovation. Considered as such, I think we can assume that "religion" and other innovations rely (at a minimum) on some sort of new idea or insight that provides some sort of competitive advantage. In that sense, all innovations rely on some sort of experience (aka insight or idea).

References

Adams, W., Graf, E., & Ernst, M. (2004). Experience can change the "light-from-above" prior. *Nature Neuroscience*, *7*(10), 1057–1058. https://doi.org/10.1038/nn1312

Arnold, M. (1960). *Emotion & Personality*. Columbia University Press.

Asprem, E., & Taves, A. (2016). Building Blocks of Human Experience Website. Retrieved from bbhe.ucsb.edu

Barnard, G.W. (1997). *Exploring unseen worlds: William James and the Philosophy of Mysticism.* Albany: State University of New York Press.

Barnby, J. M., & Bell, V. (2017). The Sensed Presence Questionnaire (SenPQ): initial psychometric validation of a measure of the "Sensed Presence" experience. *PeerJ*, *5*, e3149. https://doi.org/10.7717/peerj.3149

Beyer, S. (1973). *Magic and ritual in Tibet: The cult of Tara*. Delhi: Motilal Barnasidass.

Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Bulkeley, K. (2016). *Big dreams: The science of dreaming and the origins of religion*. Oxford/New York: Oxford University Press.

Canna, M (2017). Modéliser les états modifiés de conscience. Vers une anthropologie interactionnelle de la conscience. *Intellectica*, *67*, 268-299.

D'Aquili, E., & Newberg, A. (1999). *The Mystical Mind: Probing the Biology of Religious Experience*. Minneapolis: Fortress Press.

Faure, Bernard (1991). *The Rhetoric of Immediacy: A Cultural Critique of Chan/Zen Buddhism*. Princeton NJ/Oxford: Princeton University Press.

Hughes, John. (1994). *Self Realization in Kashmir Shivaism: The oral Teachings of Swami Lakshmanjoo*. Albany: State University of New York Press.

Fortier, M. (Forthcoming). Sense of reality, metacognition and culture in schizophrenic and drug-induced hallucinations: An interdisciplinary approach. In J. Proust & M. Fortier (Eds.), *Metacognitive Diversity: An Interdisciplinary Approach*. Oxford/New York: Oxford University Press.

Gregory, Peter N. (1991). *Sudden and Gradual: Approaches to Enlightenment in Chinese Thought*. Motilal Banarsidass Publications.

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience*. Lebanon, NH: Northeastern University Press.

Gonzalez-Maeso, J., & Sealfon, S. C. (2009). Psychedelics and schizophrenia. *Trends in Neurosciences*, *32*(4), 225–232.

Griffiths, R. R., Johnson, M. W., Richards, W. A., Richards, B. D., McCann, U., & Jesse, R. (2011). Psilocybin occasioned mystical-type experiences: immediate and persisting dose-related effects. *Psychopharmacology (Berl)*, *218*(4), 649–65. https://doi.org/10.1007/s00213-011-2358-5

Griffiths, R. R., Richards, W. A., McCann, U., & Jesse, R. (2006). Psilocybin can occasion mystical-type experiences having substantial and sustained personal meaning and spiritual significance. *Psychopharmacology (Berl)*, *187*(3), 268-83; discussion 284-92. https://doi.org/10.1007/s00213-006-0457-5

Halloy, A. (2015). *Divinités incarnées: L'apprentissage de la possession dans un culte afro-brésilien*. Paris: PETRA.

Hartogsohn, I. (2017). Constructing drug effects: A history of set and setting. *Drug Science, Policy and Law*, *3*, 1–17. https://doi.org/10.1177/2050324516683325

Heider, F. (1958). *The Psychology of Interpersonal Relations*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Johnston, H., & Noakes, J. A. (2005). *Frames of protest: social movements and the framing perspective*. Rowman & Littlefield Publishers.

Knill, D., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Kozhevnikov, M., Louchakova, O., Josipovic, Z., & Motes, M. A. (2009). The enhancement of visuospatial processing efficiency through buddhist deity meditation. *Psychological Science*, *20*(5), 645–653.

Lazarus, R. S., & Folkman, S. (1984). *Stress, Appraisal, and Coping* (1 edition). New York: Springer Publishing Company.

Luhrmann, T., Padmavati, R., Tharoor, H., & Osei, A. (2015). Differences in voice-hearing experiences of people with psychosis in the USA, India and Ghana: interview-based study. *The British Journal of Psychiatry*, *206*(1), 41–44.

Malle, B. F. (2004). *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, Mass: A Bradford Book.

Norbu, Namkhai Chogyal & Shane, John. (1986). *The Crystal and the Way of Light: Sutra,*

Tantra, and Dzogchen. New York/London: Routledge & Kegan Paul.

Padoux, A. (2017). *The Hindu Tantric World: An Overview*. Chicago: Chicago University Press.

Petitmengin, C., & Lachaux, J.-P. (2013). Microcognitive science: Bridging experiential and neuronal microdynamics. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00617

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Consciousness and Cognition*, *22*(2), 654–669. https://doi.org/10.1016/j.concog.2013.02.004

Proudfoot, W., & Shaver, P. (1975). Attribution Theory and the Psychology of Religion. *Journal for the Scientific Study of Religion*, *14*(4), 317. https://doi.org/10.2307/1384404

Roepstorff, A., Niewöhner, J., & Beck, S. (2010). Enculturing brains through patterned practices. *Neural Networks*, *23*(8–9), 1051–1059. https://doi.org/10.1016/j.neunet.2010.08.002

Rooks, B. (2014). Ayahuasca and the Godhead: An Interview with Wahid Azal of the Fatimiya Sufi Order. *Reality Sandwich*. Retrieved from http://realitysandwich.com/219826/ayahuasca-and-the-godhead-an-interview-with-wahid-azal-of-the-the-fatimiya-sufi-order/

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal Processes in Emotion: Theory, Methods, Research* (1 edition). Oxford, New York: Oxford University Press.

Seligman, R., Kirmayer, L., (2008). Dissociative Experience and Cultural Neuroscience: Narrative, Metaphor and Mechanism. *Cultural Medical Psychiatry*, 32, 31–64.

Snow, D. A., Rochford, E. B., Worden, S. K., & Benford, R. D. (1986). Frame Alignment Processes, Micromobilization, and Movement Participation. *American Sociological Review*, *51*(4), 464. https://doi.org/10.2307/2095581

Spilka, B. (B), Shaver, P. (P), & Kirkpatrick, L. A. (LA). (1985). A general attribution theory for the psychology of religion. *Journal for the Scientific Study of Religion*, *24*(1), 1–20.

Stépanoff, Charles. (2015). Trans-singularities: The Cognitive Foundations of Shamanism in Northern Asia. *Social Anthropology*, *23* (2), 169-185.

Sun, J., & Perona, P. (1998). Where is the sun? *Nature Neuroscience*, *1*(3), 183–184. https://doi.org/10.1038/630

Taves, A. (2009). *Religious experience reconsidered: A building-block approach to the study of religion and other special things*. Princeton NJ/Oxford: Princeton University Press.

Taves, A. (2010). No Field Is an Island: Fostering Collaboration between the Academic Study of Religion and the Sciences. *Method and Theory in the Study of Religion*, *22*(2–3), 177–188.

Taves, A. (2015). Reverse Engineering Complex Cultural Concepts: Identifying Building Blocks of "Religion." *Journal of Cognition and Culture*, *15*(1-2), 191–216. https://doi.org/10.1163/15685373-12342146

Taves, A., & Asprem, E. (2017). Experience as event: Event cognition and the study of (religious) experiences. *Religion, Brain & Behavior*, *7*(1), 43–62. https://doi.org/10.1080/2153599X.2016.1150327

Tylor, E. (1871). *Primitive culture: Researches into the development of mythology, philosophy, religion, language, art and custom*. Londres: John Murray.

Underwood, R., Kumari, V., & Peters, E. (2016). Appraisals of psychotic experiences: an experimental investigation of symptomatic, remitted and non-need-for-care individuals. *Psychol Med*, *46*(6), 1249–63. https://doi.org/10.1017/S0033291715002780

Underwood, R., Kumari, V., & Peters, E. (2016). Cognitive and neural models of threat appraisal in psychosis: A theoretical integration. *Psychiatry Res*, *239*, 131–8. https://doi.org/10.1016/j.psychres.2016.03.016

Vollenweider, F. X., & Kometer, M. (2010). The neurobiology of psychedelic drugs: implications for the treatment of mood disorders. *Nature Reviews Neuroscience*, *11*(9), 642–651.

Zeimbekis, J., & Raftopoulos, A. (Eds.). (2015). *The Cognitive Penetrability of Perception: New Philosophical Perspectives* (1 edition). New York, NY: Oxford University Press.

Zinberg, N. (1986). *Drug, Set, and Setting: The Basis for Controlled Intoxicant Use* (1 edition). New Haven: Yale University Press.

ALIUS BULLETIN

exploring the diversity of consciousness

n°2
FEBRUARY 2018
aliusresearch.org