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