Consciousness as generative entanglement

Article (Accepted Version)

Clark, Andy (2019) Consciousness as generative entanglement. *Journal of Philosophy*, 116 (12). pp. 645-662. ISSN 1939-8549

This version is available from Sussex Research Online: <http://sro.sussex.ac.uk/id/eprint/84690/>

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:

Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

CONSCIOUSNESS AS GENERATIVE ENTANGLEMENT*

Recent work in cognitive and computational neuroscience depicts the human brain as a complex, multi-layer prediction engine. Such a device is constantly striving to generate the sensory signal ‘from the top-down’ using stored probabilistic knowledge about the world. This family of models, which I’ll refer to as ‘Predictive Processing Models’ (henceforth PP), has had great success in accounting for a wide variety of phenomena involving perception, action, and attention.¹ It has also found suggestive applications in computational psychiatry. There, these models have suggested new and powerful ways to think about both neurotypical and altered forms of human experiences—for example, in schizophrenia and autism.² But despite their clear promise as accounts of the neurocomputational origins of typical and atypical forms of human experience, they have not yet been leveraged so as to shed light on the so-called ‘hard problem’ of consciousness—the problem of explaining why and how the world is subjectively experienced at all, and why those experiences seem just the way they do.³ To address this issue, I motivate and defend a picture of conscious experience as flowing from ‘generative entanglements’ that mix predictions about the world, the body, and (crucially) our own reactive dispositions. It is this process I argue that reveals a structured world apparently populated by all manner of puzzling ‘qualia’.

¹ See Karl Friston, “The Free-Energy Principle: A Rough Guide to the Brain?,” *Trends in Cognitive Sciences*, XIII, 7 (July 2009): 293–301 and Harriet Feldman and Karl Friston, “Attention, Uncertainty, and Free-Energy,” *Frontiers in Human Neuroscience*, IV, 215 (December 2010): 1–23. For good introductions see Jakob Hohwy, *The Predictive Mind* (New York: Oxford University Press, 2013) and Andy Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (New York: Oxford University Press, 2016). For a recent review, see Floris De Lange, Micha Heilbron and Peter Kok, “How Do Expectations Shape Perception?,” *Trends in Cognitive Sciences*, XXII, 9 (June 2018): 764–779.

² See Paul Fletcher and Chris Frith, “Perceiving is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia,” *Nature Reviews Neuroscience*, x, 1 (January 2009): 48–58; Sanders Van de Cruys, Kris Evers, Ruth Van der Hallen, Lien Van Eylen, Bart Boets, Lee de-Wit, “Precise Minds in Uncertain Worlds: Predictive Coding in Autism,” *Psychological Review*, CXXI, 4 (October 2014): 649–675; Karl Friston, Rebecca Lawson and Chris Frith, “On Hyperpriors and Hypopriors: Comment on Pellicano and Burr,” *Trends in Cognitive Sciences*, XVII, 1 (January 2013): 504–505; and Juliet Griffin and Paul Fletcher, “Predictive Processing, Source Monitoring, and Psychosis,” *Annual Review of Clinical Psychology*, XIII, 1 (May 2017): 265–289.

³ Thomas Nagel, “What is it Like to be a Bat?,” *Philosophical Review*, LXXXIII, 4 (October 1974): 435–456; Joseph Levine, “Materialism and Qualia: The Explanatory Gap,” *Pacific Philosophical Quarterly*, LXIV, 4 (October 1983): 354–361; and David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996).

I begin (section I) by rehearsing some core themes from PP models of perception, action, and emotion. I show (Section II) that these approaches already capture many aspects of experience, suggesting powerful new bridges between phenomenology and the sciences of mind. Sections III and IV use these resources to tackle key aspects of the so-called ‘hard problem’ itself, asking under what conditions a being will infer the presence of qualia and (crucially) begin to find them puzzling. These sections develop the picture I call ‘generative entanglement’. The main treatment ends (Section V) by further discussing what kind of solution (if any) to the traditional ‘hard problem of explaining consciousness’ is here on offer.

I. CORE THEMES IN PREDICTIVE PROCESSING

There are now many good introductions to the PP story (see footnote 1 above). Rather than repeat such an exercise here, I simply highlight three elements that will prove central to the account of generative entanglement.

The first—shared with many other approaches (including Deep Learning applications⁴)—is the use of a *multi-level probabilistic generative model*. A probabilistic generative model is a statistically informed model of how target data streams (in this instance sensory information) were produced. For example, a generative model of the language used in this paper might include information about the statistical structure of English, and (at a somewhat higher level) information about what words and phrases dominate this particular text, and (at a higher level still) even about the stylistic nuances that characterize a specific author’s writing. Crucially, a system that has learnt such a model can use it to create plausible ‘fake’ versions of the training data.⁵

The second feature to stress is one that is more specific to the ‘predictive processing’ (PP) family of approaches. These approaches highlight *precision-weighted sensory prediction error* as the key lever for both learning and online response. Sensory prediction error arises when the multi-level probabilistic generative model targets the sensory flux with a stream of top-down prediction. This amounts to the system seeking to generate a kind of virtual version of the sensory flux (and all its multi-level, multi-area inner resonances) for itself, using

⁴ For example, Alexey Dosovitskiy, Jost Springenberg, Maxim Tatarchenko and Thomas Brox, “Learning to Generate Chairs, Tables and Cars with Convolutional Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, XXXIX, 4 (May 2016): 692–705.

⁵ Such as the photo-realistic images of ‘fake celebrities’ seen in Tero Karras, Samuli Laine and Timo Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” arXiv:1812.04948 (December 2018): 1–12.

what it knows about current context and typical patterns in the world. Obviously, such guessing will often fail in multiple respects, and at multiple levels. Failed guesses generate what are known as 'prediction error signals'. These signal residual error (the sensory information minus the prediction). Prediction error signals are thus highly specific, and are used to recruit new and better rounds of top-down guessing until some acceptable tolerance is reached.

PP argues the information that flows forward as this processing proceeds is always and everywhere constituted solely by these error signals, representing deviations from expected states. These signals provide the information and the impetus required to drive learning or to move the system into new states. They provide a systemically 'self-computable' fulcrum for self-organizing in the service of adaptive success. This is because prediction errors are quashed not just by moving to a better model of the world, but also (complementarily) by altering the world so as to conform to a trusted model. This process occurs at many levels and at many timescales, enabling PP to offer a remarkably unified picture of perception, action, and reasoning.⁶

Crucially, PP models do this flexibly, in ways that are sensitive to changing goals and contexts. This is achieved by means of a further suite of generative-model-based processes whose purpose is to select and amplify all (and only) the prediction error signals that are estimated as most important and reliable given some task and context. This is the process (or better, the heterogeneous, multi-timescale, suite of processes⁷) that implement the so-called 'precision-weighting' of prediction errors in the brain. These precision-weightings are themselves inferred products that are learnt as part and parcel of the predictive (generative) model. The deep benefit of precision-weighting is

⁶ For a satisfying sketch, see Giovanni Pezzulo, Francesco Rigoli and Karl Friston, "Hierarchical Active Inference: A Theory of Motivated Control," *Trends in Cognitive Sciences*, XXII, 4 (April 2018): 294–306; Giovanni Pezzulo, Francesco Rigoli and Karl Friston, "Active Inference, Homeostatic Regulation and Adaptive Behavioural Control," *Progress in Neurobiology*, CXXXIV (November 2015): 17–35.

⁷ Key players in the precision-weighting game include neurotransmitters such as dopamine, serotonin, and acetylcholine, neuronal oscillations, and specific sub-cortical structures such as the pulvinar. See Ryota Kanai, Yutaka Komura, Stewart Shipp and Karl Friston, "Cerebral Hierarchies: Predictive Processing, Precision and the Pulvinar," *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, CCCLXX, 1668 (May 2015): 11–16; Karl Friston, Philipp Schwartenbeck, Thomas Fitzgerald, Michael Moutoussis, Timothy Behrens and Raymond Dolan, "The Anatomy of Choice: Dopamine and Decision-Making," *Philosophical Transactions of the Royal Society, Series B: Biological Sciences*, CCCLXIX, 1655 (November 2014): 20130481; and Mark Miller and Andy Clark, "Happily Entangled: Prediction, Emotion and the Embodied Mind," *Synthese*, CXCv, 6 (June 2018): 2559–75.

that it provides a means to put a single generative model to use in a wide variety of different ways, varying moment-by-moment patterns of communication and influence within the nervous system itself, creating transient organizations that best serve some current need or purpose.

This leads neatly to the third and final feature to be highlighted—the *flexibility and heterogeneity* of the predictive web. By this I mean that the generative model that issues in sensory (and other) predictions does not behave as a fixed monolithic entity. Instead, it is realized as a loose and context-sensitive aggregation of neuronal sub-structures and recurring circuit-level motifs, each of which has its own character, targeting different kinds of patterns, at different (usually multiple) timescales.⁸ Within this rich economy, systemically estimated precision plays the role of conductor and arranger, orchestrating the moment-by-moment use of neuronal (and hence bodily) resources. Variable precision-weighting most fundamentally tracks the systemically estimated reliability (inverse variance) of different prediction error signals. But increasing the precision of some signals relative to others can also stand proxy for self-estimated value more generally construed. For example, when I am searching for my car keys on their yellow fob, the PP apparatus should up-regulate the value of information (carried by prediction errors) relating to that color.⁹

Precision variations thus allow the PP system to alter the relative influence of different neuronal areas and different bodies of information according to changing estimates of task, needs, and context. Importantly, ‘context’ here can itself include the estimated state of both the inner (bodily) and the outer (worldly) arenas. For example, a hungry agent may up-regulate exteroceptive information signaling a well-known red and yellow signposted roadside opportunity. In sum, we should not think in terms of a fixed multi-

⁸ For example, recent work on the hippocampus suggests that it is best characterized as a predictive organ specializing in representations of future states given some current state. This broadens the role of the hippocampus beyond its classical (place-encoding) remit, revealing it as part of a complex predictive mosaic—a part that specializes in a ‘successor representation’ that can apply in both spatial and more abstract contexts. See Kimberly Stachenfeld, Matthew Botvinick and Samuel Gershman, “The Hippocampus as a Predictive Map,” *Nature Neuroscience*, XX, 11 (November 2017): 1643–53; and Peter Kok and Nicholas Turk-Browne, “Stimulus Prediction in the Hippocampus,” *bioRxiv* (January 2018): 246843.

⁹ Demonstrating such ‘gating’ effects den Ouden and colleagues showed that changing task demands could alter the strength of the influence of visual over motor areas, in a way that was controlled by prediction error signals computed in a further region (the putamen). See Hanneke den Ouden, Jean Daunizeau, Jonathan Roiser, Karl Friston and Klass Stephan, “Striatal Prediction Error Modulates Cortical Coupling,” *Journal of Neuroscience*, XXX, 9 (March 2010): 3210–19.

level prediction engine, so much as a complex weave of predictive resources whose internal relations are surprisingly fluid, varying both with task and with inner and outer context.

II. THE PICTURE SO FAR

The PP toolkit has already been deployed, to great effect, as a way of understanding many interlocking features of lived human experience. What PP offers, in a nutshell, is a principled account of how sensing puts us in touch with a structured world populated by possibilities for action. This is the story gradually laid out in a large and ever-growing literature.¹⁰

The starting point is the PP schema rehearsed briefly above. Under that schema, perception and action involve meeting the sensory flux with a stream of apt (generative-model based) top-down prediction. This occurs within a multi-area layered architecture that tracks features and events at multiple scales of space and time. But this whole prediction-perception-action cycle is constantly and profoundly impacted by another, more inward-looking cycle of prediction, perception and action. The inward-looking cycle targets our own changing physiological states—states involving the gut, viscera, blood-pressure, heart-rate, and the whole inner economy underlying hunger, thirst, and other bodily needs and appetites. This is the inner realm celebrated by William James in his model of emotions as the perception of our own bodily responses,¹¹ and tweaked and elaborated by recent theorists such as Damasio and Prinz.¹² PP builds very neatly upon such stories, by stressing the importance of two-way predictive traffic between exteroceptive and interoceptive sources of information. Thus Seth argues that feelings and emotions reflect predictions that integrate information involving interoceptive, proprioceptive, and

¹⁰ See Karl Friston, “Hierarchical Models in the Brain,” *PLoS Computational Biology*, IV, 11 (November 2008): e1000211; Andy Clark, “Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science,” *Behavioral and Brain Sciences*, XXXVI, 3 (June 2013): 181–204; Jakob Hohwy, *The Predictive Mind* (New York: Oxford University Press, 2013); Anil Seth, “Interoceptive Inference, Emotion, and the Embodied Self,” *Trends in Cognitive Sciences*, XVII, 11 (November 2013): 565–73; Lisa Feldman Barrett and Kyle Simmons, “Interoceptive Predictions in the Brain,” *Nature Reviews Neuroscience*, XVI, 7 (July 2015): 1–11; Andy Clark, “Surfing Uncertainty,” *op. cit.*; Pezzulo, Rigoli, and Friston, “Active Inference, Homeostatic Regulation and Adaptive Behavioural Control,” *op. cit.*; and Pezzulo, Rigoli, and Friston, “Hierarchical Active Inference,” *op. cit.*

¹¹ William James, *The Principles of Psychology*, Vols. I & II (Cambridge, MA: Harvard University Press, 1890/1950).

¹² Antonio Damasio, *The Feeling of What Happens* (New York: Harcourt, Brace and Co., 1999) and Jesse Prinz, *Gut Reactions* (New York: Oxford University Press, 2004).

exteroceptive cues.¹³ Sensory information about our current heart rate, for example, may be used to help predict the presence or absence of threatening objects and situations.¹⁴ As my bodily state alters, the salience of various worldly opportunities (to eat, for example) alters too. That means I will act differently, harvesting different streams of exteroceptive and interoceptive information, that in turn determine new predictive cascades.

This makes progress with what might be thought of as the problem of ‘simple sentience’. The term sentience was used by Bentham and others to mark a distinction between the capacity to feel and the capacity to think and reason.¹⁵ Creatures lacking the full sweep of human capacities to think and reason, it was argued, might nonetheless be capable of feeling pain and pleasure, and more generally of feeling how things are going in their own species-specific way. The bedrock PP story begins to show us how this might be so. It depicts sentient beings as those beings whose action-selecting generative model of the (organism-salient) world is in constant two-way communication with a generative model of their own changing physiological state. Such creatures will perceptually encounter a world fit for action, in which what actions are selected depend heavily upon an interoceptively-mediated sense of their current needs and embodied state. This constant entwining of interoceptive and exteroceptive prediction offers a compelling picture of creatures that perceive a world in ways constantly nuanced by their own bodily needs and state. They encounter and navigate the world in ways that reflect and respect their changing bodily state (am I in need of food, in danger of damage, in need of rest and repair?). Creatures like that will certainly *appear* sentient. They will respond to their worlds in ways informed by a delicate dance between what they detect in the external world and their own ever-changing bodily needs and states.

This, I argue, is what underlies all the behavioral manifestations of ‘sentience’. We detect sentience in creatures (and potentially in robots) whose take on the external world is subtly but pervasively responsive to their brain or control system’s take on their own inner, bodily worlds and their own states of

¹³ Seth, “Interoceptive Inference, Emotion, and the Embodied Self,” *op. cit.*

¹⁴ Such interactions have been powerfully demonstrated in experiments that present neutral and fearful stimuli while manipulating self-estimated heart-rate, showing that neutral stimuli are more often seen as fearful when heart-rate is estimated as increased. See Eric Anderson, Erika Siegel, Dominique White and Lisa Feldman Barrett, “Out of Sight but not Out of Mind: Unseen Affective Faces Influence Evaluations and Social Impressions,” *Emotion*, XII, 6 (December 2012): 1210–21.

¹⁵ For example, Jeremy Bentham, *Introduction to the Principles of Morals and Legislation*, second ed., (Oxford: Clarendon Press, 1823).

action-readiness. But human consciousness, as we'll next see, involves additional abilities to model ourselves and our own 'reactive dispositions'. The result will be a picture of human consciousness as emerging from an organization that scores highly along three inter-animated dimensions—the depth of a generative model, its inflection by rich swathes of interoceptive information, and (next section) a creature's ability to model itself and its own reactive dispositions. The PP apparatus shows not only how these multiple sources of information can engage and interact, and but also why some aspects of those interactions may remain hidden from introspection and self-report.

Does the inflection of a generative model by bodily and action-specifying information constitute true sentience, or might it still be merely apparent sentience? I am not convinced that this is the right question to ask. Let's just say (for the moment) that PP offers a promising story about how behavioral *patterns like that* might be ushered into being by the constant interanimated effort to predict internal and external sensory variation, and to minimize prediction errors as we do so. We can then ask, what else could be going on when we humans go one step further and start spontaneously to report (and puzzle over) the presence, in our own experience, of all those distinctive looks, feels, and 'qualia'?

III. INFERRING QUALIA

Let's assume, for the sake of argument, that a perfectly good account of the nature and origins of linguistic communication has already been achieved. In the present context, this seems like an innocent move. We can then ask "under what conditions would a PP agent, already competent in the production and decoding of informative linguistic signals, infer the existence of qualia—and perhaps then become puzzled by them—as she engages in the ongoing prediction of her own interoceptive and exteroceptive sensory stream?"

The next step borrows a key move from Daniel Dennett. Dennett suspects that a major part of what goes wrong when we start to theorize about qualia is that we fail to appreciate the possibility of what he neatly dubs a 'strange inversion'.¹⁶ Consider the sweet taste of honey. According to the standard (non-inverted) story, we might say we like the honey because it tastes so sweet. The view from after Dennett's Strange Inversion is rather different. Our liking of honey, Dennett suggests, is nothing but the subtle complex of

¹⁶ Daniel Dennett, "Expecting Ourselves to Expect: The Bayesian Brain as a Projector," *Behavioral and Brain Sciences*, XXXVI, 3 (June 2013): 209–210; Daniel Dennett, "Why and How Does Consciousness Seem the Way it Seems?," in Thomas Metzinger & Jennifer M. Windt, eds., *Open MIND* (Frankfurt am Main: MIND Group, 2015).

reactive dispositions honey evokes—dispositions (in many folk, at least) to seek it, lick it, affirm it to be tasty, and so on.

In other words, the complex of reactive dispositions comes first, and we label things which evoke that complex of responses as ‘sweet’. In other words it’s not the sweetness which explains our response: it’s our response (a complex functional profile linking inputs to outputs via inner states) which explains the diagnosis of sweetness. But not knowing this, Dennett argues, we project the sweetness onto the world, resulting in the self-diagnosis of a perceptual realm which includes not just dogs, cats, and tables but also the sweet taste of honey. This misprojection, Dennett argues,¹⁷ can itself fruitfully be cast in a way that thus puts it in direct contact with the PP apparatus. To see how, notice first that PP claims that dogs, cats, and senators are all detected as elements in the distal realm simply in virtue of their being inferred as useful predictive nodes (or ‘latent variables’¹⁸) that enable apt prediction of the sensory flux. Just as a PP system might infer, for dealing with a large corpus of English text, that there exist specific English-language words (such as the word ‘cat’) that help explain and predict the arrangements of letters on a page, so it might infer the existence of items such as dog, cats, and senators as a way of predicting rather more complex patterns in the sensory stream. But notice that we ourselves are complex structured causes in the world, and causes, moreover, whose effects are of great adaptive significance. As such, we humans are prime objects for our own prediction machinery. To grapple with our own futures, and those of other agents like ourselves, we need to predict what we, and other agents, will do next.

Suppose you are a being that nature has designed to seek out and favor sugary things, and that as a result you exhibit the pattern of reactive dispositions associated with ‘sweetness’. You then set about the important task of predicting your own and others behaviors. ‘Sweetness’ now emerges as nothing other than a kind of high-level shorthand for those distinctive reactive complexes. But it now becomes a shorthand deployed by the system itself, as a

¹⁷ Dennett, “Why and How Does Consciousness Seem the Way it Seems?,” *op. cit.*

¹⁸ A latent variable is an inferred (or ‘hidden’ or ‘fictive’) cause, generated so as to regiment or simplify some pattern of data or sensory information. Thus, exposed to a corpus of English language text, a machine learning algorithm might infer the existence of nouns and verbs as a way of explaining and predicting patterns in the data. See, for example, Thomas Landauer and Susan Dumais, “A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge,” *Psychological Review*, CIV, 2 (April 1997): 211–40. And for useful general discussion, see Denny Borsboom, Gideon Mellenbergh and Jaap van Heerden, “The Theoretical Status of Latent Variables,” *Psychological Review*, CX, 2 (April 2003): 203–19.

means of better predicting its own and others behaviors. The same applies to features such as cuteness, which now appears as a kind of high-level neural shorthand or latent variable that tracks a different set of sensed features (big round eyes, head disproportionately large relative to body, etc.)—this time, ones designed to elicit approach and nurture responses in adult humans. It is that linked complex of object-features and associated reactive dispositions (here learnt as high-level latent variables that predict sensory flows) that language-enabled beings capture by speaking of themselves as experiencing subjective ‘cuteness’. Such ways of speaking and thinking thus flow from the use of simple internal models of ourselves and others—models that are useful for our daily purposes of prediction and the coordination of joint action.¹⁹

What about qualitative experiences of color—do these perhaps present a harder puzzle? The short answer is that they may (and should) be treated in just the same way.²⁰ In the case of redness, and here simplifying a little for the sake of easy exposition, the relative amounts of each wavelength (estimated against a backdrop of prior knowledge) constitute a lower-level sensory signature that becomes associated with swathes of distinctive reactive dispositions. The predictive compression achieved in all these cases is every bit as good as the predictive compression that (PP claims) underlies the detection and exploitation of every other hidden cause inferred to explain regular patterns within the perceptual stream. Inferred qualitative states are thus on a level footing with inferred states such as ‘being a dog’ or ‘being a test-tube’. Redness, sweetness, and cuteness all pass what Dennett calls “the Bayesian test for being an objective structural part of the world we live in.”²¹ From the point of view of the predicting agent, the world now simply contains brute visual episodes presenting dogs, cats, and senators, but also redness, sweetness, and cuteness.

Finally, notice that just as I may easily spot a never-before-seen object as being a new kind of car or a new kind of food mixer, so too I can spot never-before-encountered cases of qualia, such as when I experience a never-before-encountered kind of animal as cute. The reason, in each case, is that the high-level shorthand is locking on to a complex of cues and features that are largely unknown to the subject, but that can be spotted when they occur, even in

¹⁹ Such a perspective fits well with that of Friston (2017) who suggests that conscious experience is a process found only in creatures whose generative models display *rich temporal thickness*. Such creatures predict how their own and others future actions will impact and alter their own encounters with the universe. This means they can minimize expected prediction error (surprise) on many interacting timescales, from choosing today’s dinner to choosing a partner or pursuing a certain career.

²⁰ Dennett, “Why and How Does Consciousness Seem the Way it Seems?,” *op. cit.*

²¹ *Ibid.*, section 6.

brand new instantiations. If a new kind of animal has large eyes and a disproportionately large head (and meets a host of other interlocking, context-sensitive, ‘soft constraints’),²² then it will be apt for prediction using the high-level latent variable most closely corresponding to the brute linguistic label ‘cute’. This will also recruit all the existing behavioral dispositions associated with cuteness, as they were already linked to that latent variable.

IV. THE PUZZLINGNESS OF QUALIA

Let’s briefly rehearse the story so far. PP depicts perception as governed by a multi-faceted generative model that weaves together information about physiological states, possible actions, and worldly causes. That already shows how perceiving animals encounter a structured world apt for action and intervention, and do so in ways inflected at every level by an interoceptively-mediated sense of their own changing embodied dispositions. But some creatures make further headway by explicitly ‘expecting themselves’. They do this by generating high-level latent variables (representations of hidden causes) that associate complex reactive dispositions, in themselves and others, with complex combinations of feature in body and world.²³

The next move, however, is the crucial one. For suppose we now ask, what will beings thus organized say if asked what they find and detect during their perceptual encounters with the world? Before addressing this issue, it is important to remind ourselves that the presence of the kind of organization just described is consistent with the complete absence of capacities for verbal rehearsal and report. We must not assume that every creature that predicts its own web of reactive dispositions need be able to speak and report at all. But we may legitimately ask what a language-enabled agent, thus organized, might say if they were asked. Given a certain visual input, such an agent might say, for example, “I detect a large cute dog, probably some kind of Labrador cross, whose coat is reddish-brown.”

²² I borrow this term from work on connectionism and artificial neural networks, where it is used precisely to indicate the complex interlocking mesh of small cues that together, in context-sensitive ways, underpin our capacities to recognize and discriminate people, objects, and situations.

²³ Or only in the body: for example, in the case of certain pains. For an excellent treatment of the case of pain, see Colin Klein, *What the Body Commands: The Imperative Theory of Pain* (Cambridge, MA: MIT Press, 2015). Much of the story there could rather easily be translated into the idiom of complex action-driving prediction. That, however, is a project for another day.

The first point to note is that every property and feature here reported has been extracted in exactly the same way. Each feature and property is inferred as part of the multi-level stream that best predicts the current waves of sensory stimulation. Redness, largeness, cuteness, dog-ness, and Labrador-ness all emerge as inferred properties of the distal scene. They are all just latent variables (‘fictive causes’, in the terminology of Hobson and Friston²⁴) that serve to organize and predict whole swathes of interacting lower-level sensory information. Nonetheless, an agent constructed in this way might very well come to be puzzled by the so-called ‘qualitative dimensions’ of her own experiences. There are two reasons for this. The first—common to both qualitative (‘red’) and judgmental (‘dog’) aspects—is that agential access is to final estimations and not to intervening processes. The second, peculiar to the more intuitively qualitative aspects, is that the web of soft constraints that determine those estimations makes deep and essential reference to our own and others’ predicted responses and reactive dispositions.

Concerning the first, notice that an agent, when engaged in reflection or verbal report, has access only to how the predictive model currently estimates distal and bodily states to be, and never to the full details of the processing that leads to those estimations. This makes good design sense. The whole point of the process of probabilistic inference is to estimate how things are in the twin arenas (body and world) relevant to survival and action. Knowledge that looks computationally inwards, at the details of the processing stream itself, would be adaptively redundant. An agent with that kind of genuine but limited access to her own processing will experience a world populated by dogs, cats, chairs, hurricanes—but also by sweet tasting snacks and scary movies. The second cause for puzzlement now comes to the fore. These latter patterns should strike the intelligent agent as especially strange and hard to analyze. For these agents the patterns that matter for these predictive hidden causes (‘sweet-tasting’ ‘scary’) involve worldly ‘content’ items such as cakes and movies. But they also lock on to those items in ways that involve (unseen by the agent) complex cascades of interoceptive and proprioceptive predictions—ones that track our own and others reactive dispositions.

The upshot is that the contents that constitute our qualitative experiences are subtly responsive to our own reactive dispositions and current physiological state in ways that typically remain hidden. They function only as what Koch called ‘executive summaries’²⁵ whose cognitive role is to support

²⁴ Allan Hobson and Karl Friston, “Consciousness, Dreams, and Inference: The Cartesian Theatre Revisited,” *Journal of Consciousness Studies*, XXI, 1 (January 2014): 6–32.

²⁵ Christof Koch, *The Quest for Consciousness* (New York: Roberts and Co., 2004).

choice and action. In the PP story, they are elements in a simplified model of ourselves—a model that (like any other model learnt using the PP apparatus) is under pressure to be both accurate and concise,²⁶ capturing only the complexity needed to support behavioral and adaptive success.

This may be the deep reason for the mismatch, nicely noted by DeLancey,²⁷ between the overall complexity of the neuronal processing that may underlie a specific experience and the complexity of the pictures of that processing that we can actively manipulate when reasoning about ourselves and others. This mismatch, Delancy notes, destabilizes traditional arguments that seek to press substantive conclusions from the apparent conceivability of qualia-free ‘zombie twins’ (physical duplicates of us but lacking qualitative experience).²⁸ For that apparent conceivability may reflect nothing other than the low-complexity of our own model and theoretical reasoning relative to the actual processing, the upshot being that “we lose information about the phenomenal kind in order to reason about it.”²⁹

The PP story helps fill in the cognitive scientific details of exactly the kind of mismatch that Delancy identifies. It shows what it is about the processes that construct systemic response that both weave together huge webs of interacting and often unexpected materials, and then productively ignore the details of the intervening processing for the purposes of constructing a useful model of ourselves as knowing agents—the model used for self-reflection, social interaction, and the selection of action. This also suggests potentially rich contact with work on the so-called ‘social neuroscience theory of consciousness’.³⁰ According to that theory, human consciousness is intimately tied to the prediction of our own behavior, and human agents construct simplified models of their own and other’s mental states that are experienced as reflective conscious awareness.

The PP agent thus finds herself in somewhat the situation described by Chalmers³¹ in a revealing passage that asks how the process of perception

²⁶ This can be achieved by, for example, adding a so-called ‘occam factor’ favoring the simplest (fewest parameter) model that does the job. In biological brains, the process of synaptic pruning during sleep may implement just such a process. See Hobson and Friston, “Consciousness, Dreams, and Inference,” *op. cit.*

²⁷ Craig Delancey, C., “The Modal Arguments and the Complexity of Consciousness,” *Ratio*, XXVI, 1 (March 2013): 35–50.

²⁸ For a nice illustration, *ibid.*, pp. 47–8.

²⁹ *Ibid.*, p. 45.

³⁰ Michael Graziano, *Consciousness and the Social Brain* (New York: Oxford University Press, 2013).

³¹ Chalmers, *The Conscious Mind*, *op. cit.*

would strike an advanced agent that has access only to the products of complex inner processing. Such an agent, when asked how they *know* that the honey tastes sweet, may be forced to say they know this directly, in some brute but puzzling manner. In Chalmers' own words:

Given this kind of direct access to information states... it is natural to expect the system to use the language of 'experience' and 'quality' to describe its own cognitive point of view on perception. And it is unsurprising that this will all seem quite strange to the system: these immediately known, ineffable states which seem so central to its access to the world, but which are so hard to pin down. Indeed, it is natural to suppose that this would seem odd to the system in the same sort of way in which consciousness seems odd to us.³²

Chalmers himself does not think this move delivers the goods. Specifically, he wonders why the agent's experience then strikes her as having a perceptual 'feel' at all—why is it not like 'just knowing what is there' without any perceptual character whatsoever? Here, it helps to reflect that the intelligent PP agent will know a lot more than simply that there is a dog somehow given in her evidential base. She will know the shape, color, texture, and behavioral tendencies of the animal. She will know (roughly at least) that some of her own knowledge comes from sight, and other parts from hearing and touch, as she can predict how her evidential base would alter and fluctuate were she to cover her eyes or ears. She will know that (for example) dogs have four legs and tails, and will be able actively to seek out additional visual evidence for each such feature by attending more closely to the right spatial regions, finding additional tactile evidence by touching the right features, etc. She will also know (implicitly) a lot about how her own sensory inputs should flow and alter as she saccades around the dog-containing scene. That means she will have access to all the rich bodies of object- and modality- specifying information highlighted by work on 'sensorimotor contingency theory'.³³ Moreover, because her estimates of worldly states are constructed in ways that take into account the constant flow of interoceptive and proprioceptive information, all this will be accompanied by a background sense of action-readiness.³⁴

³² *Ibid.*, p. 291.

³³ Kevin O'Regan and Alva Noë, "A Sensorimotor Approach to Vision and Visual Consciousness," *Behavioral and Brain Sciences*, XXIV, 5 (October 2001): 883–975.

³⁴ See Julian Kiverstein, Mark Miller and Erik Rietveld, "The Feeling of Grip: Novelty, Error Dynamics, and the Predictive Brain," *Synthese*, CXCVI, 7 (July 2019): 2847–69.

In sum, nothing in the emerging PP picture suggests ‘judging in the dark’, in the sense of suddenly simply knowing (merely intellectually, as it were) that you confront a dog—as if by some kind of blindsight. But whereas the sensory patterns and sensorimotor predictions that manifest as estimations of ‘dog-ness’ and ‘cat-ness’ are indeed mostly rooted in the nature and behavior of stuff in the world, the patterns manifest as high-level estimations of ‘sweetness’ and ‘scariness’ also owe a great deal to our own physiological responses and reactive dispositions. This is why they can seem (despite the deep commonality in mode of construction) so much more puzzling than the others, resulting in talk of ‘irreducible qualia’ that are simply given in experience.

Our strong species-tendency towards dualist world-views and the persistent diagnosis of a profound ‘explanatory gap’ dogging the scientific explanation of consciousness both flow from this single source. This suggests an important layer on top of the ‘complexity mismatch’ explanation suggested by DeLancey.³⁵ For what really drives our Cartesian intuitions, if the PP story is on track, is not the complexity mismatch per se, so much as the way that mismatch is constructed—a way that involves hiding the core role of predictions concerning our own reactive dispositions, hiding the ongoing role of interoceptive information, and constructing a simplified self-model to help guide future choice and action.

V. SOLVING, DISSOLVING, REVISING?

It might be asked³⁶ what kind of ‘solution’ (if any) to the traditional ‘hard problem’ is here on offer? On the one hand, the PP story makes plain how it is that adaptive agents can (in an experientially neutral sense, of the kind described in for example Chalmers³⁷) encounter a structured world apt for action. But instead of then asking the standard next question (“How can such beings enjoy qualitative experiences, and why are those experiences the way they are?”) we instead address a different issue—under what conditions would creatures like that infer that they are home to puzzling ‘qualia’? Notice that we must here read ‘infer’ in an experientially neutral way too.³⁸ Just as the PP agent infers the existence of dogs, cats, tables and chairs (as a neat way of making

³⁵ Delancey, “The Modal Arguments and the Complexity of Consciousness,” *op. cit.*

³⁶ Thanks to an anonymous referee for this JOURNAL for pursuing this issue, in ways that have greatly helped to clarify the intended picture.

³⁷ David Chalmers, *The Conscious Mind*, *op. cit.*

³⁸ We might thus ask “Why, do sufficiently intelligent agents of the PP stripe exhibit what Chalmers (forthcoming) describes as ‘dispositions to make quasi-phenomenal reports, where reports are understood as outputs that even a non-conscious being could make.’”

predictive sense of the sensory flux) so too she infers the presence of rather puzzling ‘qualitative features’ such as the cuteness, brownness, or annoyingness of the dog.³⁹

By moving to explain these patterns of inference, we effectively shift one part of the debate from a full-on confrontation of the traditional hard problem to something different—a response to what Chalmers calls the ‘meta-puzzle of consciousness’.⁴⁰ This is the problem of explaining various behaviors such as gasps of appreciation of doggy brown-ness, doggy cuteness, and (crucially) verbal expressions of puzzlement over how such features could possibly be being detected by purely material systems, as well as tendencies to say things like “I’m a dualist” and so forth.

If the story outlined earlier is on track, there is a fully satisfying account of all these behaviors⁴¹ available using the Bayesian/PP apparatus. To summarize, PP agents whose ongoing patterns of inference reflect constant internal (interoceptive) self-monitoring and the shape of their own ongoing preparations for action will encounter (in that neutral sense) a world that is valenced through and through. They will encounter a world some of whose structures seem to be inviting them to approach and engage, while others invite them to stay back. They will also encounter a world in which the discriminations they can make (concerning color, shape etc.) are made using cascading and multi-faceted inferences whose details are (for good adaptive reasons) opaque and hidden from view. For the adaptive agent needs to discriminate what’s out there, and in what ways it matters to her. She does not

³⁹ This issue marks a point at which our treatment might diverge from Dennett’s. We hold that qualia are just more content items in a good predictive model, so in that sense our ideas about qualia turn out to be very much like our ideas about dogs and cats. But we also (like Dennett) think that we find qualia more puzzling due to the extent to which they involve greater degrees of hidden reference to our own inner physiological states and reactive dispositions. In other words, the sensorium patterns corresponding to dogs and cats are less critically woven with our bodily responses. That said, it is not clear that this difference should imply any deep difference in metaphysical status. This is a nice issue for future discussion (and thanks to an anonymous referee for this JOURNAL for pressing this question).

⁴⁰ *Ibid.*

⁴¹ There is one such behavior that deserves separate treatment. It is our ability to imaginatively hold all the perceptual facts firm while varying our take on the distal world. This allows us to do things such as say that we could be having the same ‘experience’ even if we were brains in a vat, or prisoners of the Matrix. For an extended treatment of this capacity using the PP apparatus, see Andy Clark, Karl Friston, and Sam Wilkinson, “Bayesing Qualia: Consciousness as Inference, not Raw Datum,” *Journal of Consciousness Studies*, XXVI, 9-10 (October 2019): 19–33.

need to know how that information is being computed, or in what ways her own reactive dispositions and preparations for action are being factored in to the rolling pattern of inference that carves up her world. It is this combination of opacity of and multi-facetedness that, I have argued, leads sufficiently intelligent agents to infer that they are home to puzzling ‘qualitative states’.

By depicting brains as fundamentally multi-faceted hierarchical inference engines, the Predictive Processing story explains why it is that we are driven to make these inferences. The inference to qualia emerges as a good way of capturing useful coarse patterns in our own behavior: ones that can inform our own reasoning and planning, and help us co-ordinate with other similarly designed agents. At the same time, PP shows why these inferences should not be accorded any special authority or status. They emerge in the same way, and for the same reasons, as all our other inferences—to cats, tables, and poker-games. The combined effect is to take the metaphysical sting out of the quale’s tail, but to do so in a way that respects the full complexity and variety of human behavior—including the behaviors that express puzzlement and the undeniable pull of dualism. By showing how such behaviors emerge naturally from PP’s independently-motivated picture, we put engineering, neuroscientific, and information-theoretic flesh on the familiar Dennett-style picture, adding the full weight of that progressive, testable, and neurophysiologically plausible framework.⁴²

Does this solve, dissolve, or simply fail to address the hard puzzle? It probably depends where you start. If you start with the idea that experiential facts are brute givens, then you may conclude that the most we have done is to show why we find them so deeply puzzling. That should at least help render our puzzlement less metaphysically salient. One might go further and recommend that we therefore revise our conception of what qualia are, understanding them as *nothing but* the products of the processes of inference described. For this is a process that simultaneously delivers the ability to respond and act in a structured world, and the ability (given sufficient intelligence) to become increasingly puzzled by those very responses, permeated as they are by large yet invisible swathes of interoceptive and self-predictive information. Either way, this is progress. But a lot of hard work remains to be done. That work will involve unraveling the fine details of all

⁴² The present work should thus be seen (along with Clark, Friston, and Wilkinson, “Bayesing Qualia,” *op.cit.*) as a sustained attempt to cash out the picture briefly imagined in Dennett, “Expecting Ourselves to Expect: The Bayesian Brain as a Projector,” *op. cit.*

those key generative entanglements, and seeing how they map (or fail to map) onto existing proposals in the science of consciousness.⁴³

VI. CONCLUSIONS

What we experience as qualia are simply content items in our best generative model of the world. Otherwise expressed, they are inferred latent variables (hidden or fictive causes) posited by the brain to capture and predict useful patterns in the sensory stream. But some of these patterns involve deep generative entanglements implicating exteroceptive, interoceptive, and proprioceptive (action-specifying) sensory flows. Such hidden entanglements combine information about objective features of the world with information about our own physiological states, our action-readiness, and our own long-term reactive dispositions.

This suggests a graded three-dimensional picture of the core determinants of conscious experience. The first dimension reflects the depth and breadth of a generative model of worldly causes.⁴⁴ The second (the one

⁴³ This is clearly a very large project, whose shape I can only gesture at here. But within PP, it is the ‘precision-weighting’ process that maps most naturally onto the contours and insights of leading work in the science of consciousness. Variable precision-weighting provides PP with the necessary and sufficient means of implementing the core functionality of the popular ‘global workspace’ story about conscious experience since high-precision predictions and prediction errors form a temporary web of increased influence (see Bernard Baars, *A Cognitive Theory of Consciousness* (Cambridge, MA: Cambridge University Press, 1989) and Stanislas Dehaene and Lionel Naccache, “Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework,” *Cognition*, LXXIX, 1–2 (April 2001): 1–37). It is plausible that such temporary precision-based webs of influence are characterized by increased ‘informational integration’ among the communicating populations (see Giulio Tononi, “Consciousness as Integrated Information: a Provisional Manifesto,” *The Biological Bulletin*, CCXV, 3 (December 2008): 216–42). Finally, recurrent processing (yet another prominent thread in the science of consciousness) can be both a consequence and a cause of enhanced precision (see Victor Lamme, “Towards a True Neural Stance on Consciousness,” *Trends in Cognitive Sciences*, x, 11 (November 2006): 494–501). PP thus suggests that these three leading stories are consistent, providing interlocking means by which conscious contents emerge as precision-driven best-estimates of organism-salient (potentially action-driving) states of affairs.

⁴⁴ A generative model is deep when it involves many layers of compression and recoding, enabling low-level patterns to be predicted using a rich downwards cascade. Such cascades will tend (see Friston, “Hierarchical Models in the Brain,” *op. cit.*) to separate out regularities at ever-larger scales of space and time, integrating information across greater timescales at each higher level of processing. But not every deep model is thereby wide. To be wide, a generative model needs to be able to deal with many different kinds of worldly events, changes, and states of affairs, using old knowledge in new ways as contexts morph. See Karl

most naturally associated with ‘simple sentience’) concerns the delicate ongoing inflection of those generative model-based predictions by a stream of interoceptive and proprioceptive information. The third dimension (more closely linked to what might be thought of as reflective self-consciousness) adds to that heady mix the ability of advanced cognizers to model themselves and their own reactive dispositions, so as to co-ordinate with others, and to explore possible futures conditioned on their own likes, wants, and actions. Only creatures scoring quite highly on the first dimension will seem to be truly knowledgeable about their worlds. Only creatures scoring quite highly on the second dimension will systematically and pro-actively factor their own bodily states and needs into their estimations of distal opportunities and their choices of action. But only creatures scoring quite highly on all three dimensions are poised to become puzzled about the nature and origins of their own ‘qualitative experiences’. Human agents score highly (and in inter-animated ways) along all three dimensions.

One might ask, as you ascend the scales from zero to infinity along each of the three dimensions, where does ‘true sentience’, or ‘real consciousness’, or real ‘self-awareness’ begin to be present? Would a simple robot that manifests modest entanglement between a low-grade version of the three dimensions enjoy a correspondingly low-grade sentience or conscious awareness? What about very simple creatures with rich interoceptive capacities, poor exteroception, and no forward-looking self-model? Perhaps the best we can do, in such cases, is to note how the target beings compare along the various dimensions, and how deeply entangled the processing along those dimensions has (or has not) become. Such an approach highlights multiplicity and diversity rather than any single core feature.⁴⁵ Yet by stressing the role of multiple generative entanglements, and showing why the resulting ‘inferred qualia’ will seem so very puzzling to advanced agents, we position this as a preliminary sketch of the nature and possibility of human consciousness itself.

ANDY CLARK

Dept of Philosophy, and Dept of Informatics, University of Sussex
And Dept of Philosophy, Macquarie University, Sydney, Australia.

*This work was carried out with the support of Horizon 2020 European Union ERC Advanced Grant XSPECT - DLV-692739. Special thanks to the two anonymous

Friston, Marco Lin, Chris Frith, Giovanni Pezzulo, Allan Hobson and Sasha Ondobaka, “Active Inference, Curiosity and Insight,” *Neural Computation*, XXIX, 10 (October 2017): 2633–83.

⁴⁵ For a defense of this kind of approach, see Aaron Sloman, “An Alternative to Working on Machine Consciousness,” *International Journal of Machine Consciousness*, II, 1 (June 2010): 1–18.

referees; to Karl Friston, Sarah Garfinkel, Jakob Hohwy, Anil Seth, and Lucia Melloni; and to the members of the XSPECT project team – Mark Miller, Kate Nave, George Deane, David Carmel, Frank Schumann, and Sam Wilkinson.