All the Self We Need

Philip Gerrans

I use evidence from psychiatric disorders involving the experience of depersonalisation to decompose the causal and cognitive structure of experiences reported as self-awareness. I combine insights from predictive coding theory and the appraisal theory of emotion to explain the association between hypoactivity in the Anterior Insular Cortex and depersonalization. This resolves a puzzle for some theories raised by the fact that reduced affective response in depersonalization is associated with normal interoception and activity in Posterior Insular Cortex. It also elegantly accounts for the role of anxiety in depersonalisation via the role of attention in predictive coding theories.

Keywords
Affective processing | Appraisal theory of emotion | Bodily awareness | Depersonalisation | Disorders of self-awareness | Identity | Phenomenal avatar | Predictive coding | Self | Simulation

Author

Philip Gerrans
philip.gerrans@adelaide.edu.au
University of Adelaide
Adelaide, Australia

Commentator

Ying-Tung Lin
linyingtung@gmail.com
國立陽明大學
National Yang-Ming University
Taipei, Taiwan

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

> "Who is the I that knows the bodily me, who has an image of myself and a sense of identity over time, who knows that I have propriate strivings?" I know all these things, and what is more, I know that I know them. But who is it who has this perspectival grasp? It is much easier to *feel* the self than to *define* the self (Allport 1961, p. 128)

1 Preliminary remarks

I think Allport has it the wrong way round. It is easy to *define* the self, as he in fact does, as the thing that thinks, feels, perceives and has a sense of identity over time. It is hard, however, to a find an entity that fits the definition. This is so even though, according to Allport, experiencing being a self is unproblematic ("it is easier to *feel* the self"). In fact, the experience of being someone is actually very elusive, phenomenologically and conceptually. On some accounts self-awareness is actually the experience of *Being No-One*[1] (Met-

1 Strictly speaking, the experience is not of being no one, since there is no one to be. Rather it is an experience we cannot help but take to be of being someone, even though there is no entity causing the experience.

Gerrans, P. (2015). All the Self We Need.
In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 15(T). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570078

zinger 2003). In this chapter I use disorders of self-awareness to develop an account of the experience which gives rise to the feelings referred to by Allport. In the final sections we shall see whether our experience is of being someone, no-one, or something other than a self. Perhaps a body. Or the process of thinking.

The conclusion is that self-awareness is *almost* a necessary or inevitable illusion when the mind is functioning smoothly. The experience of being a self is produced by mechanisms that compute the relevance of sensory (including, and especially, bodily) information to a variety of organismic goals represented at different levels of explicitness in a cognitive hierarchy. The computations relate information to those goals, *not to selves*. Those computations of goal relevance produce consequent bodily feelings. Those, and only those, feelings give us the phenomenal information we need to plan, remember, and interact with other people and the world as though we are unified selves. Thomas Metzinger argues that integration of information in experience amounts to the construction of a phenomenal avatar, which the brain uses to manoeuvre the organism through the world (Blanke & Metzinger 2009; Metzinger 2011). I agree, and the rest of the chapter can be seen as an attempt to anatomise that avatar. I use evidence from psychiatric disorders involving the experience of depersonalisation to decompose the causal and cognitive structure of experiences reported as self-awareness.

2 Introduction

So many psychiatric disorders are explained in terms of the way the patient experiences herself that, even if intuitive or philosophical theories which posit a self as the object of experience are not correct, there is an interesting phenomenon there to be explained. My idea is that the best integrative explanation of those disorders is *ipso facto* the best philosophical theory of self-awareness because those disorders cannot be explained other than via a model of the way the experience is generated in normal and abnormal situations.[2] Once we have explained those disorders we can determine the theoretical utility of overlapping folk, clinical and philosophical conceptions of self-awareness. Thus, the approach I take is consistent with that proposed by Dominic Murphy in his plea for a (cognitive neuro) scientific psychiatry: "we arrive at a comprehensive set of positive facts about how the mind works, and then ask which of its products and breakdowns matter for our various projects" (2006, p. 105).

So until the concluding sections I use the term self-awareness to refer to the experience we report in terms of awareness of being a unified persisting entity: the same person at a time and over time. It may turn out that such experiences are illusions or misinterpretations of some other phenomenon, perhaps because there are no such entities as selves, but I delay that discussion until the evidence is assembled. To anticipate, I think the intuitive folk concept of self-awareness is very like the intuitive concept of episodic memory, which is of "re-experiencing" a previous episode. Cognitive neuroscience tells us that in fact episodic memory experiences are constructed to suit current cognitive context rather than retrieved intact. However it does no harm in everyday life to think of episodic memory as content-preserving retrieval of past experience. Similarly the intuitive conception of self-awareness tracks processes which, when they function harmoniously, produce experiences that provide a plausible basis for the concept of a unified and persisting self. That

There is no substantial Cartesian, or bodily, or neural, entity that sustains the properties ascribed by Allport. Thus part of Metzinger's project is to explain why we feel as though we are substantial entities.

2 In other words I take the strong view advocated by Murphy. The ontology of the mind *is* the ontology of cognitive science. The reason is that only with the correct theory of cognitive architecture in place can we understand how neural processes implement the cognitive processes whose operations we experience as personal-level phenomenology. That personal-level phenomenology provides the raw material for intuitive or folk explanations that abstract from cognitive and neural realization. But that abstraction is precisely why, as Halligan and Marshall once memorably said, in the absence of a suitably constrained cognitive model, psychiatry will be consumed by "the expensive and extensive search for non existent entities" (Halligan & Marshall 1996, p. 6). I take the view that mechanistic (in the sense of neuroscientific) and phenomenological (based on reflection on the nature of experience) explanation are not independent projects. One *could* have a purely personal-level phenomenological ontology of mind. But the fact that such ontologies mislead about the sources of psychiatric disorder is a reason to search for an integrative theory. But the only way neuroscience can explain experience is via a detailed computational, cognitive theory.

concept, while not entirely accurate, provides a useful ability to represent and communicate sufficient unity and persistence. If I tell you I will be happy to pick you up at the airport you need to be able to rely on *me* to be at the Arrivals gate. The precise nature of my (dis)unification as a single self is not relevant. If I told you I would send my body but would not be present myself you would phone a psychiatrist. (It would be super to be able to deputise your body to attend departmental meetings, weddings etc. on your behalf, wouldn't it?) Yet something like that phenomenon of alienation occurs in depersonalisation, as a deeply felt and distressing phenomenon. The difference in experience between people with depersonalisation and those without it is an essential *explanandum* both for psychiatry and for philosophers interested in the (possibly illusory) phenomenology of selfhood.

The rest of the chapter proceeds as follows. I first discuss the Cotard delusion, in which people say that they have died, disappeared or do not exist (délire de négation). The Cotard delusion raises a set of questions about the relationship between self-awareness, bodily experience, and affective processing. I outline some suggestive intuitive answers to these questions based on the phenomenology of the disorder but argue that they are insufficient as explanations. A deeper explanation is provided by the cognitive neuroscience of depersonalisation. That explanation relies on a theoretical framework that draws on

I. The appraisal theory of emotion
II. The simulation model of memory and prospection
III. The hierarchical predictive coding model of cognitive processing

This framework allows us to explain how:

- affective experiences provide the basis for self-awareness as a distinct form of bodily awareness *moment to moment*
- those moment to moment experiences of self-awareness can be annexed to cognitive processes whose temporal reach is longer than the present, creating the experience/illusion of a continuing self
- when affective processing is compromised the resultant experience is reported as change, or in extreme cases, loss, of self. Mere absence of bodily or affective response *per se* does not lead to depersonalisation. What leads to depersonalisation is the absence of *predicted* affective responses that normally constitute self-awareness that leads to depersonalisation. This explanation also provides a full explanation of an intriguing phenomenological observation made by Cotard about the role of anxiety in generating depersonalisation.

With this theoretical framework in place I discuss depersonalisation disorder and depersonalisation aspects of the Cotard delusion, resolving some of the questions raised by the initial phenomenological explanation.

Once those questions are answered we can make some comments on the theoretical utility of philosophical theories of self-awareness, which for convenience I classify into four types: Illusory Self, Fat Controller, Embodied Self, Narrative Self. The Illusory Self is a version of the Humean idea that self-awareness is either illusory or a theoretically loaded misdescription of some other experiential phenomenon (perceptual, interoceptive, emotional, somatic). It is quite consistent with the Illusory Self theory that the experience is a "necessary illusion" created by architecture installed by evolution. The Fat Controller theory is that self-awareness is the experience of a genuine *substantial* self, a locus of higher order cognitive integration and top down control (like the aptly-named Will Self's Fat Controller in his *Quantity Theory of Insanity*). Embodied Self theories identify self-awareness with forms of bodily awareness. Finally there are Narrative views of the self, thin and thick. On the thin view the self is a "centre of narrative gravity", a fictive entity generated by the Joycean machine to organize and communicate. On thicker views the self is not a fiction but a genuine cognitive entity whose essence is to construct and communicate its own autobiography as an essential aspect of higher order cognitive control. The Thin view goes

naturally with the Illusory Self view: it explains the persistence of the Illusion, while the Thick view (naturally enough) fits well with Fat Controller views.

Cognitive neuroscience does not vindicate any of these theories. However this does not mean that we should regard the phenomenon of self-awareness as empirically disconfirmed. It turns out that there are cognitive processes that generate experiences with some of the properties ascribed by different theories under different conditions. So, as with episodic memory, rather than explaining self-awareness away, we can describe and explain the nature of the experiences reported as self-awareness in terms of the structure of the processing which generates it. Self-awareness is a cognitive illusion, based on the nature of affective processing. The relevant experience plays a crucial role in higher levels of cognitive control that organise and communicate experience in narrative form: fragments, episodes, chronicles, histories and epics (Currie & Jureidini 2004; Goldie 2011; Jureidini 2012). This conjunction of processes makes self-awareness an irresistible illusion. The nature and necessity of this illusion is shown by the nature of the disorders that arise when it fails.

3 The phenomenology of the Cotard syndrome

In their study of uncommon psychiatric syndromes Enoch & Trethowan (1991) provided a haunting clinical vignette. They described a patient who said that her body was decomposing and disappearing and that eventually she would be "just a voice". Another patient suffering from the same condition described himself as a "dead star" orbiting an inert galaxy. The Cotard delusion, from which these patients suffer, was described by Jules Cotard in 1882 as a "délire de négation", a delusion of inexistence (Cotard 1880, 1882, 1884, 1891; Debruyne et al. 2009). It is also described as a paradoxical belief that one is dead. The current cultural fascination with zombies provides the metaphor of "walking corpse" syndrome to describe the condition. However, as with many psychiatric disorders, perhaps the most telling descriptions and explanations of the phenomenon were provided in the nineteenth century, in this case by Cotard himself. He described his patient thus:

> Miss X affirms she has no brain, no nerves, no chest, no stomach, no intestines; there's only skin and bones of a decomposing body. . . . She has no soul, God does not exist, neither the devil. She's nothing more than a decomposing body, and has no need to eat for living, she cannot die a natural death, she exists eternally if she's not burned, the fire will be the only solution for her. (Translation from Cotard 1880)

Cotard explained this delusion as a consequence of a particular type of psychotic depression "characterized by anxious melancholia, ideas of damnation or rejection, insensitivity to pain, delusions of nonexistence concerning one's own body, and delusions of immortality" (Debruyne et al. 2009, p. 67).

More recently (Gerrans 2000, 2001; Debruyne et al. 2009) the delusion of inexistence has been explained as a consequence of the experience of depersonalisation. The delusion is a personal level response to an intractable and impenetrable loss of affective response to the world. Of course to say that an experience is of depersonalization is not an explanation but an intuitive characterization: the concept expresses the phenomenology of feeling disconnected from the world including one's own body, as though experiences are "not happening to me". Such feelings plausibly originate in what we might call affective derealisation: the failure of emotionally salient events to trigger affective responses in the patient so that the world feels strange and unreal. Since affective responses are a form of bodily experience it makes sense that the Cotard delusion is often expressed as beliefs about alteration in body state: in particular that the body is vanishing, disappearing or dead. And since there is an intimate connection between felt body state and self-awareness this loss of normal affective response is expressed as the idea that the self no longer exists.

But surely it is equally intuitively plausible that a person suffering from derealisation might express the experience by saying that the world (perhaps including her body) feels strange, emotionally inert or unreal? In other words, why does the patient not report derealisation, the feeling that the world is unreal? One possible answer is contained in the following suggestion:

> Cases of the Cotard delusion have been reported . . . in which the subject proceeds beyond reporting her rotting flesh or her death to the stage of describing the world as an inert cosmos whose processes she merely registers without using the first-person pronoun….The patient does not recognize experiences as significant for her because, due to the global suppression of affect [ex hypothesi a consequence of extreme depression], she has no qualitative responses to the acquisition of even the most significant information. These extreme cases of the Cotard delusion are those in which neural systems on which affect depends are suppressed and, as a consequence, it seems to the patient as if her experiences do not belong to her. Thus the patient reports, not changes in herself, but changes in the states of the universe, one component of which is her body, now thought of as another inert physical substance first decomposing and finally disappearing. (Gerrans 2000)

My earlier self suggested that when the patient experiences global affective suppression she experiences her body as simply a body, a physical substance rather than the body which sustains the self or the body *qua* self: Hence the depersonalisation. However this simply begs the question. What is it about affective processing which transforms representations of body states to representations of states of a self?

4 Feelings of self-relevance

Appraisal theory is familiar to theorists of emotion as the theory that emotions are representations of the significance of events for the organism. Fear, for example, results from the representation of objects as dangerous for the organism. Early appraisal theorists assimilated these appraisals to beliefs about the properties of the objects of emotion (Kenny 1963; Solomon 1976, 1993). Consequently appraisal theory has been criticized as overly intellectualistic and as ignoring the felt aspect of emotion. Fear is a visceral state whose essence is a feeling, not a judgement, runs the objection. Equally an emotional feeling may arise or persist in the absence of, or in opposition to, a judgment.

Recent versions of the theory avoid this objection by recognising that most emotional appraisals are in fact conducted by neural circuits that automatically link perception to the automatic regulation of visceral and bodily responses. Consequently appraisals issue almost instantaneously in feelings that reflect the nature of that appraisal. When we recognize a familiar person and see her smile, for example, the significance of that information for us has been represented and that representation used to initiate our own bodily response within a few hundred milliseconds (Adolphs et al. 2002; Sander et al. 2003; Sander et al. 2005; N'Diaye et al. 2009; Adolphs 2010).

The consequence of these appraisals is autonomically-regulated body states and action tendencies that produce changes in visceral and bodily state. These changes are sensed as affective feelings via specialised circuitry that evolved to monitor organismic state. At any given moment we experience a "core affect" which is the product of multiple appraisals along different dimensions at different time scales.

These affective processes essentially represent the significance of incoming information for the organism along a number of different dimensions—hedonic, prudential, dangerous, noxious, nourishing, interesting, and so on. These representations, however, relate an aspect of organismic functioning to a represented object; they do not represent a self *per se*. The detection of danger alerts the organism to the need for avoidance, for example. The consequent feeling of fear is a way of sensing the bodily consequences of that appraisal. The self as an en-

tity need not be represented in either the initial appraisal or the consequent experience. The self-relevance (as appraisal theorists call it) of dangerous objects is however *implicitly represented* in the bodily experience of fear. The same is true of all affective experiences: they carry important information about the world and the way the organism is faring in it in virtue of the appraisal processes which generate them. But they do so without representing a self in any substantial sense. Rather they relate salient information to organismic goals represented at different levels of explicitness for different purposes (Tomkins 1962, 1991; Scherer 2004).

Cognitive neuroscience has identified circuits that function as "hubs" of distributed circuits that determine the subjective relevance of information. Lower-level hubs, of which the amygdala is a central component, implement rapid online appraisals (Sander et al. 2003; Adolphs 2010) and coordinate visceral and bodily responses. These lower level hubs associate affective experiences with online sensorimotor processing of the type often described as reflexive: that is initiated by, and dependent on, encounters with the environment. It follows that such experiences decay with the representation of the stimulus. They are stimulus dependent. Such reflexive affective processes can of course only sustain a feeling of self-relevance moment to moment.

5 Simulation, affective sampling, and the self

By self-awareness, however, philosophers have in mind the experience of being an entity that exists through time, which is not something that can be produced by reflexive processing. The organism needs to be able to represent itself, not just moment-to-moment but as an entity with a history and a future ("to consider itself the same thing at a time and over time".) It must therefore be able to link affective experience to memory and prospection in the same way as it links it to perception and sensory processing moment to moment. That is to say that it must be able to appraise episodes of memory and foresight for self-relevance.

Because the temporal window of human cognition extends beyond the present we have evolved systems that recapitulate important aspects of reflexive affective processing for those higher level cognitive processes involved in planning, recollection, prospection, and decision-making. These systems *simulate* temporally distant experiences by rehearsing some of the same perceptual and emotional mechanisms activated by the simulated situation. As a result we can recall previous episodes of experience and imagine future episodes of experience and link those simulations to other high-level cognitive processes in order to plan and decide. We remember being sunburnt and imagine getting skin cancer when deciding whether to go to the beach at noon (Gusnard et al. 2001; Buckner et al. 2008; Fair et al. 2008; Broyd et al. 2009).

These simulations are the raw material of autobiographical narratives whose structure and duration can vary depending on cognitive context. They may be as simple as recall of a single event that triggers a flash of affect, but can also be assembled into elaborate histories and imaginative rehearsals depending on the cognitive context. This narrative capacity provides a crucial aspect of cognitive control possibly unique to humans. The most important aspect of these simulations is sometimes overlooked in studies that emphasise their quasi-perceptual content. That is the fact that the simulation of perceptual and sensory experience evokes affective associations. We simulate a scene in order to evoke the affective responses that represent the significance of events and objects for us. When we imagine or recall an episode of experience its affective significance is also represented in experience via the offline rehearsal of affective processing. The ventromedial prefrontal cortex is a structure which "traffics" or makes available the affective information. In effect, the ventromedial prefrontal cortex recapitulates at a higher level the properties of the amygdala. In so doing it associates affective information with explicitly represented information used in reflective decision making and planning (Ochsner et al. 2002; Bechara & Damasio 2005). It thus allows the subject to make explicit reflective appraisals. When I lie on the beach I have pleas-

ant feelings produced by low-level appraisal systems. When I imagine or recall lying on the beach while trying to decide whether to holiday in Thailand or Senegal my ventromedial prefrontal cortex makes available the affective information prompted by that simulation.

This is why "pure" episodic memory studies (such as recall of content of visual scenes) do not activate the ventromedial prefrontal cortex, whereas "activations in the ventromedial PFC [prefrontal cortex] … are almost invariably found in *autobiographical* memory studies" (Gilboa 2004, p. 1336; my emphasis). Gilboa (2004) suggests that this is because "autobiographical memory relies on a quick intuitive 'feeling of rightness' to monitor the veracity and cohesiveness of retrieved memories in relation to an activated self-schema." This is consistent with studies showing activity in the ventromedial and related subcortical structures when people make intuitive (that is, rapid and semiautomatic) judgments about themselves. When people make judgments about themselves using semantic knowledge and symbolic reasoning, ventromedial structures are less active.

This idea is supported by studies of patients with lesions to the ventromedial prefrontal cortex. These patients oscillate between various forms of reflexive cognition and more abstract forms of thinking using semantic knowledge and procedural reasoning. What they have lost is the ability, provided by ventromedial structures, to simulate affective and motivational response in the absence of the stimulus, while they retain the ability to process information in an abstract way. Consequently, a ventromedial patient may be able to do a utility calculation about her personal future but be unable to act on that knowledge. It appears that semantic knowledge is motivationally inert. Such results are often used to emphasize the necessary role of affect in deliberation, but they also suggest that what those affective responses do is provide the necessary *personal* perspective on information. They make the information *mine*, so to speak. Furthermore, this diminishment is not just *at* a time, but over time. These patients, although not amnesic in the strict sense of the term, have very limited ability for autobiographical recall or prospection. They have no sense of a persisting self (Damasio 1994; Bechara & Damasio 2005; Gerrans & Kennett 2010).

This suggests that disorders in which people feel a diminished sense of self would be characterized by hypoactivity in the ventromedial prefrontal cortex. In a review of the neuropsychological and imaging literature, Koenigs & Grafman concluded that "one could conceive of the VMPFC patients' selective reduction in depressive symptoms as a secondary effect of *a primary lack of self-awareness and self-reflection*" (2009, p. 242; my emphasis). In other words, patients with ventromedial damage do not "feel" personally affected when considering even quite distressing events because they cannot access or activate the required affective responses.

It seems that "mine-ness" of experience is a cognitive achievement mediated by the ventromedial prefrontal cortex. As we noted above the ventromedial prefrontal cortex is suited to play this role because it recapitulates at a higher level many of the processing properties of lower-level hubs of emotional processing that represent self-relevance. Rather than reinvent the cognitive wheel for controlled processing, evolution has provided pathways that traffic affective and reward-predictive information processed automatically at lower levels to controlled processing coordinated by the ventromedial prefrontal cortex.

In effect, these studies suggest that in both online reflexive and offline reflective processing affective processes are needed to represent the significance of the information for the subject, and it is the consequent bodily feelings that produce the feeling of self-awareness My version of this view is in some ways an amalgam of ideas found in Seth (2013) and Proust (2013). All three of us share the view that the mind is hierarchically organized, and that feelings of self awareness emerge when higher order, metacognitive processes such as planning or deliberation integrate bodily information which signals relevance. On Seth's and my view the Anterior Insular Cortex (AIC) is in some ways specialized for that function in view of its architecture: it does

not merely relay first order bodily information but is involved in the representation of the significance of that information. Thus it is well placed to be the source of some of the metacognitive feelings identified by Proust (2013) as serving crucial indicator functions.[3]

Affective processes represent the relevance of information for an organism and initiate suitable action tendencies and autonomic responses. The bodily consequences are sensed and summarised by specialised systems that inform the organism how it is faring in the world: this is affective information (Prinz 2004). This affective information is made available to other cognitive processes, which operate at different time scales, from instantaneous and automatic, to reflective and controlled. We are able to think and behave as continuing entities because the salience of information for different organismic goals is represented by affective processes at different time scales and levels of explicitness. An organism that can *use that affective information in the process* is a *self*.

This suggests that if the ability to access affective information is lost then self-awareness would also be diminished. Thus as we suggested above a key to the experience of depersonalisation in the Cotard delusion is the profound loss of affect associated with extreme depression. This suggestion is almost correct but it ignores another stage in the production of depersonalisation. After all, from what we have said so far affective processes represent the self-relevance of information. If the consequent feelings are unavailable the world should feel not significant for the subject. That is to say the subject might feel detached from the world or as if the world was emotionally inert. But it seems an extra step from a lack of affective experience to the feeling or thought of non-existence. Of course the step might be a small one. This was the idea of Gerrans in his pioneering work at the dawn of the millennium. He suggested that there was such an intimate connection between affective experience and the self that any profound involuntary change in affect would be felt as a change to the self. However since then interesting work on depersonalisation disorder has provided a deeper understanding of the phenomenon. That work draws on the predictive coding theory of cognitive function.

6 The predictive coding hierarchy

The mind is organized as a hierarchical system that uses representations of the world and its own states to control behavior. According to recently influential Bayesian theories of the mind, all levels of the cognitive hierarchy exploit the same principle: error correction (Friston 2003; Hohwy et al. 2008; Jones & Love 2011; Clark 2012, 2013; Hohwy 2013). Each cognitive system uses models of its domain to *predict* its future informational states, given actions performed by the organism. When those predictions are satisfied, the model is reinforced; when they are not, the model is revised or updated, and new predictions are generated to govern the process of error correction. Discrepancy between actual and predicted information state is called *surprisal* and represented in the form of an error signal. That signal is referred to a higher-level supervisory system, which has access to a larger database of potential solutions, to generate an instruction whose execution will cancel the error and minimize surprisal (Friston 2003; Hohwy et al. 2008). The process iterates until error signals are cancelled by suitable action.

This is a very basic outline of the predictive coding idea dodges a crucial question: the extent to which Bayesian formalisations actually describe neurocomputational processes rather than serving as a predictive calculus for neuroscience (Jones & Love 2011; Hohwy 2013; Clark 2012; Park & Friston 2013; Moutoussis et al. 2014). It also blurs an important distinction which is not salient to formalisations such as Bayesian theory: namely the fact that not all higher level control systems can and do smoothly cancel prediction errors generated at

3 There is an interesting debate to be had here. On the views of e.g., Damasio and Bechara affective feelings are not metacognitive but experiences produced by lower level or first order processes *associated* with metacognitive processes (such as planning and decision making). Proust refers to feelings generated by metacognitive processes. On the view proposed here the AIC metarepresents the *significance* of first order bodily information (e.g., visceral or tonic muscular state) in the context of self-relevant metacognition. It allows the subject to experience not just body state but the relevance of that body state.

lower levels. For example vision and motor control are good examples of predictive coding systems (Hohwy 2013). Often however experiences best explained as carrying information about prediction error are not cancelled by the adoption of a higher-level belief. Consider déjà vu experiences which signal mismatch between an affect of familiarity and perception of a novel scene (O'Connor & Moulin 2010). We know the scene is novel, but it still feels familiar. The point is just that the higher order belief does not always smoothly cancel prediction error. And this should be expected. Coding formats are not uniform across cognitive systems, which is why sensory and higher-level cognitive integration is such a cognitive achievement for the mind.

From our point of view what matters are the key ideas of hierarchical organization, upward referral of surprisal and top-down cancellation of error. Also crucial is the idea that the highest levels of cognitive control involve active, relatively unconstrained, exploration of solution space. This is the level at which attention can be redirected to alternative solutions and their imaginative rehearsal. Phenomena such as delusion represent a high level response to an obstinate signal of prediction error that cannot be simply cancelled from the top down. This way of thinking of the mind weds a version of predictive coding theory to insights from neurocomputational theory that treat executive systems as specialized for the resolution of problems which cannot be solved at lower levels. Thus at low levels in the hierarchy the structure of priors and errors and referral of surprisal is constrained, modularized some might say. At the so-called personal level of belief fixation predictive coding best describes the idea that those experiences which command executive resources are those which signal prediction error which cannot be resolved at lower perceptual and quasi perceptual levels. This is at least one level at which predictive coding involves active sampling of information (active inference) as well as the routine cancelling of surprisal according to a well defined prior model. The latter almost defines perception. The former, according to O'Reilly & Munakata 2000) as well as predictive coding theorists (Spratling 2008) is definitive of executive control.

Thus most of the detection and correction of error occurs at low levels in the processing hierarchy at temporal thresholds and using coding formats that are opaque to introspection. Keeping one's balance, parsing sentences and recognizing faces are examples. We have no introspective access to the cognitive operations involved and are aware only of the outputs. This is the sense in which our mental life is tacit: automatic, hard to verbalize, and experienced as fleeting sensations that vanish quickly in the flux of experience. This is the "Unbearable Automaticity of Being" (Bargh & Chartrand 1999). However even these relatively automatic processes generate experiences of which we can become aware. The recognition of faces, for example, produces an affective response within a few hundred milliseconds. When that affective response is absent or suppressed due to malfunction a prediction is violated and the discrepancy between familiar face and lack of familiar affect is referred to higher levels of executive control to deal with the problem.

At the higher levels of cognitive control, surprisal is signalled as experience that becomes the target of executive processes. These metacognitive processes evolved to enable humans to reflect and deliberate to control their behaviour. The highest levels of cognitive control involve reflection, deliberation, rehearsal and evaluation of alternative courses of action and explicit reasoning. When for example a predicted affect is absent we might find ourselves in the position of a patient described by Brighetti who lost affective responses to her family and her professor. She had "identity recognition of familiar faces, associated with a lack of SCR [SCR is skin conductance response, a measure of electrodermal activity consequent on affective processing]" (Brighetti et al. 2007). In other words her predicted affective response to familiars was absent, which resulted in an experience becoming the target of higher-level control processes. Such patients sometimes produce the Capgras delusion that the familiar person has been replaced by an imposter or double. A truly florid delusion such as is sometimes seen in schizo-

phrenia might elaborate the delusional thought into an epic paranoid narrative.

The aim here is not to enter into the controversy about the explanation of the Capgras delusion but to note the role of the architecture that generates it (Young et al. 1994; Breen et al. 2001; Ellis & Lewis 2001). Higher levels of cognitive control are engaged to deal with error signals referred from lower levels in the hierarchy. Perhaps the most important level in the hierarchy for personal and social life is the level at which subjectively adequate narratives are generated to make experience intelligible and by which we communicate our experiences to others. This is the level at which delusional thoughts originate. By subjectively adequate here I merely mean "fits the experience of the subject". At even higher levels of cognitive control we can revise and reject those subjectively adequate autobiographical narratives, replacing them with empirical theories that draw on publicly available norms of reasoning and semantic knowledge to produce objectively adequate responses to subjective experience (Gerrans 2014). Delusions are best conceptualized as higher-level responses to prediction error which, however, cannot cancel those errors. In fact as Clark (2013) points out such delusory models in effect "predict" further experiences of that type, which means that the delusion will be strengthened.

A very important point to note for the subsequent explanation of depersonalization and the Cotard delusion is that it is not the absence of affect *per se* which produces the error signal and engages higher-level cognition. Lack of affective response alone does not require a high level response unless that lack of affect is unpredicted. That is why we are not bothered by lack of response to strangers (we don't predict it at any level in the control hierarchy) but if a new mother has no affective response to her baby the experience can be part of a syndrome of post-natal depression.

The example of post-natal depression allows us to make another important point about the relationship between predicted affect and psychosis. Mothers most vulnerable to post-natal depression are those who had powerful positive expectations of motherhood and the bond with the infant. When that bond does not materialize for some reason they are confronted with a distressing lack of predicted affective response. Sometimes this will produce a kind of Capgras delusion regarding the baby. The mother might say that the baby has been replaced or is an alien (Brockington & Kumar 1982). Interestingly, and tellingly, if the mother is also extremely anxious the condition can be even more serious. Anxious attention to the experience tends to magnify the problem.

This role for anxiety is nicely elucidated by the predictive coding framework. Formal considerations aside, the concept of predictive coding places a huge emphasis on the signaling of error. This means that incoming information must be compared to a prediction and the difference computed and referred to a control system. At higher levels those error signals take the form of experiences. These experiences are often imprecise and opaque since they are produced by lower level systems that encode information in different formats to those used by explicit metarepresentational capacities. They also compete for metarepresentational resources among the constant flux of experiences that engage attention. Thus they create a problem of working out for any experience how much is signal and how much is noise.

It is very important for high-level cognition to be targeted as precisely as possible for only as long as required. Thus any vagueness in experience needs to be resolved. Attention is the process which solves this problem. Hohwy (2012, p. 1; my emphasis) makes the point for perceptual inference but it applies in general:

> *conscious* perception can be seen as the upshot of prediction error minimization and *attention* as the optimization of precision expectations during such perceptual inference.

Clark (2013, p. 190) makes a similar point:

> Attention, if this is correct, is simply one means by which certain error-unit responses are given increased weight, hence

becoming more apt to drive learning and plasticity, and to engage compensatory action.

The point is that attention is directed to error signals in order to make them more precise by increasing the signal to noise ratio. Attention amplifies the signal and maintains it while higher-level systems try and interpret the experience and manage appropriate responses. If the response works the error signal is cancelled and attention can be directed elsewhere.

Within this framework we can make an observation about anxiety that can be overlooked by approaches that concentrate on the arousal, hypervigiliance or the associated beliefs concerning threat or danger. These approaches de-emphasise a crucial element. That is uncertainty. Anxiety is an adaptive mechanism that primes the organism cognitively and physiologically to resolve uncertainty. Thus, if a prediction cannot be verified, or an error signal disambiguated, anxiety in this sense will result. Of course what we call pathological anxiety is the dysfunctional activation and maintenance of these mechanisms. The point is that someone who is anxious in this way will continue to misallocate attentional, cognitive and physiological resources to experiences. Another point about anxiety is that, in pathological cases, action does not cancel the signal or the dysfunctional allocation of resources to it. This may be why the role of anxiety in depersonalisation is not straightforward. Some recent studies have not found a strong correlation between anxiety and depersonalisation (e.g., Medford 2012). However the scales used to measure anxiety give a score that sums scores for self-report of feelings, behaviour and cognition. The suggestion here is that what really matters is the allocation of attention to signals which cannot be resolved, perhaps because they are intrinsically noisy, ambiguous or have insufficient information. It is also important that the patient cannot resolve the uncertainty by revising the predictive model that generates it since that is usually maintained low in the predictive hierarchy by mechanisms that are not accessible. The person with Capgras delusion, for example, automatically predicts affective response to familiar faces and when it goes missing there is nothing she can do to revise that prediction. Instead she is confronted with an anomalous experience, which automatically captures attention. Similarly with depression. Loss of affective response is not something that can be restored from the top down.

In some cases of post-natal depression all these factors seem to be operative. The mother expected to bond with the infant but in fact perhaps birth was traumatic, the baby did not attach straightaway, and the mother needed more support and reassurance than she received. She was left distressed and unable to cope which made bonding and attachment even more difficult. This would be bad enough but if the mother had a strong prior expectation that motherhood would be straightforwardly rewarding a prediction is violated. If the mother is also anxious she will attend intensively to the resultant experience of absent affect, but she will encounter only further feelings of emptiness and panic. The presence of the baby and the expectations of family and friend only compound the sense that she is not feeling what she should be feeling. What happens next depends on context and support but it is not really surprising, especially given the relationship between massive hormonal fluctuation and emotional regulation, that in some cases new mothers develop psychotic symptoms (Spinelli 2009).

7 Depersonalisation

Depersonalisation Disorder (DPD) is characterized by "alteration in the perception or experience of the self so that one feels detached from and as if one is an outside observer of one's own mental processes" (American Psychiatric Association 2000). Critchley points out that DPD is often accompanied by alexithymia, a condition in which conscious awareness of emotional states is compromised or absent. This is consistent with findings summarized by Medford that "de-affectualisation", a reduction or absence of affective response, presents as a core feature of clinical cases. Depersonalisation is a separate disorder to derealisation (the feeling that the world is inanimate or unreal) but derealisation

is often an important aspect of depersonalisation. Indeed, as Medford describes their relationship, depersonalisation can sometime be a response to derealisation (Sierra et al. 2002; Hunter et al. 2004).

Seth et al. (2011, p. 9; my emphasis) summarize a range of findings about DPD as follows: "In short, DPD can be summarized as a psychiatric condition marked by the selective diminution of the *subjective reality* of the self and world". They explain this diminution as the result of the loss of "sense of presence", the feeling of being engaged in experience. This is what they mean by subjective reality: the condition is not like an hallucination or delusion in which objective reality is misrepresented by faulty perception or belief fixation. In fact the patient correct represents "objective reality" but loses the sense of herself as the subject of experience.

In the attempt to explain the loss of the sense of presence cognitive neuroscience has developed a theoretical picture that considerably augments older theories. On those older theories DPD represented a suppression or inhibition of emotion as a response to trauma or distress. On this view DPD activates mechanisms which might in other circumstances be adaptive. For example, if the subject of violent attack deactivated those mechanisms which produce the experience of distress that would qualify as an adaptive response to trauma. Of course such a response is only adaptive in the short term. Inability to feel distress might also reduce avoidance behavior with disastrous consequences.

It seems that the deactivation is accomplished by inhibitory activity in the Ventrolateral Prefrontal Cortex (VLPFC). The VLPFC is a structure which plays a crucial role in the regulation of affective feeling, especially as part of a process of reappraisal (Füstös et al. 2013). The adaptive aspect here is that it allows the subject to redirect attention and divert cognitive resources to alternative interpretations of self-relevance and response behaviour by inhibiting an experience that would otherwise monopolise cognition. This role has been tested in tasks which involve the top down regulation of negative affect but, as Medford says, "In DPD such suppression is apparently involuntary (and largely resistant to volitional control), but it is reasonable to suppose that this will nevertheless engage similar inhibitory networks" (2012, p. 142). Thus the patient with DPD experiences the result of *involuntary* deactivation of systems that produce the bodily experience of emotion.

These ideas are consistent with the evidence from cognitive neuroscience about other primary neural correlates of DPD. *Hyper*activity in VLPFC leads to *hypo*activity in the Anterior Insular Cortex (AIC). That reduced activity in the AIC produces the loss of a sense of presence. This hypothesis results from findings that it has a primary role in higher order representation of interoceptive (visceral, autonomic, bodily) states. It generates the bodily feelings that signal how we are faring in the world moment to moment consequent on affective processing. Activity in the AIC produces what Damasio called the "core self" and what Critchley calls "the sense of presence". As Critchley says,

> evidence from a variety of sources converges to suggest a representation of autonomic and visceral responses within anterior insula cortex, where, particularly on the right side, this information is accessible to conscious awareness, influencing emotional feelings (2005, p. 162).

When Damasio made his contributions to the neurophilosophy of emotions and self-representation the computational theory in the field was less developed so that we can now make some additional observations about the role of the AIC.

To do so we first reiterate the distinction between being able to sense body state, which is the phenomenon baptized by Damasio *interoception*, (to distinguish it from exteroception [perception of the external world]), and sensing states of a self. The distinction is a subtle one of course but we can approach it intuitively by noting that there is a crucial difference between being able to sense heart rate, blood pressure or temperature as part of an illness and as part of an emotional episode. We observed earlier that the second kind of awareness is the one we describe as self-awareness in virtue of the fact that

it reflects affective processing rather than pure bodily regulation. There is a difference in feeling state caused by raises in blood pressure generated by walking up stairs and by heated argument. This is so even though heart rate is heart rate, however caused. But the point of affective processing, as we saw, is to assess the self-relevance of unpredicted changes in things like heart rate and to indicate to the subject how and why they might matter in the cognitive context.

The experiential differences between heart rate *per se* and heart rate consequent on affective processing can be explained in terms of the principle of hierarchical computational organization, reflected in cortical organization (Craig 2009, 2010; Dunn et al. 2010). The insular cortex is hierarchically organized to map body state at different levels of abstraction and integration. Posterior sections map body state directly and integrate those representations to coordinate *reflexive* regulatory functions. Thus the Posterior Insular Cortex (PIC), for example, represents things like blood pressure and departures from homeostasis and integrates that information to enable reflexive regulatory processing. More anterior regions re-represent and integrate this information in formats available for higher levels of cognitive control. If we sense raised blood pressure the PIC is primary in the representation of that information. When, however, we are deciding how to respond, we need to integrate that information with current and long term goals, representations of contextual information, memory, planning and inference. We may have to inhibit or reprogram automatic behavioral tendencies (not punch the boss) and perhaps reappraise the situation. Thus we need a way, not just to feel raised blood pressure, but to *feel its significance* in order to program a suitable response. This is the role of the AIC.

This explains a recent finding which seems paradoxical on the "somatic" James-Lange view of emotions revived by Damasio. On that view emotions are representations of body state *simpliciter*. The feeling of fear is the feeling of being primed to take avoidance action, for example. Michal and collaborators compared the "interoceptive accuracy", that is ability of patients to judge body state (using heart rate as a proxy), of patients with DPD to normal patients. Strikingly they found that "[there] was no correlation of the severity of 'anomalous body experiences' and depersonalization with measures of interoceptive accuracy." They explained this finding as follows: "[The] findings highlight a striking discrepancy of normal interoception with overwhelming experiences of disembodiment in DPD. *This may reflect difficulties of DPD patients to integrate their visceral and bodily perceptions into a sense of their selves*" (Michal & Reuchlein 2014, p. 1; my emphasis).

The AIC can only integrate currently available bodily feeling. As Craig says, it "represents the sentient self at one moment of time [and] provides the basis for the continuity of subjective emotional awareness in a finite present" (2009, p. 67). However we can extend the temporal range of information represented by those feelings by integrating them with representation of past and future episodes of experience and/or semantic knowledge. Simulations involved in planning and episodic memory are associated with activation of the AIC to provide sense of extended self. In other words it is the integration of the metarepresentations of body state produced by the AIC with representations of episodes of a temporally extended autobiography that produces the feeling that we are a self with a past and future, rather than a series of disconnected selves, moment to moment.

Nothing in what I have said refutes skepticism about the self, or episodic theories of first person experience (Strawson 2004). It is in fact consistent with the idea that experience is a series of episodes. Whether we feel those episodes are ours depends on how they are integrated. There is no suggestion that everyone integrates them the same way or that integration evokes an equally strong sense of presence in each person. All I have suggested is that there are mechanisms which can create self-awareness moment to moment and mechanisms which integrate those moments of self-awareness with higher level forms of cognitive control that represent past and future actions and outcomes in order for the organism to assess the self-relevance of actual and potential actions. The ex-

planation of awareness of self-relevance in different contexts is a sufficient explanation of the phenomenon of self-awareness that was our initial quarry.

Craig adds a subtle but important qualification to this account. He (and others) remind us that if the predictive coding account of the mind is correct then we are never directly aware of objects, including the body (Craig 2009, 2010; Seth et al. 2011; Garfinkel & Critchley 2013). Rather representations of objects are computed on the basis of discrepancy between their predicted informational effects on us and actual incoming information. It is fluctuations and discrepancies measured against expectations computed at different levels in the control hierarchy that determine the information that becomes consciously available. "An *expected* event does not need to be explicitly represented or communicated to higher cortical areas which have processed all of its relevant features prior to its occurrence" (Bubic et al. 2010, p. 10; Clark 2013, p. 199; my emphasis.)

The same should be true of neural activation in the AIC, and hence of moments of self-awareness. We are aware of what is relevant to us via unpredicted changes in bodily feeling consequent on affective processing.

This latter feature is the key to understanding the link between "de-affectualisation", as Medford called it, and depersonalization (Medford 2012). It is not the fact that affect is suppressed that matters, but that affect which was predicted to occur does not in virtue of the *involuntary* inhibition of the AIC by the VLPFC. When people engage in voluntary or effortful inhibition of affect they do not feel depersonalized. We noted earlier the role of expectation in post-natal depression, but there the expectation is of affective response to a specific object, a baby. In depersonalization it seems that almost all expected affective feelings are absent because of hyperactivity in the VLPFC.

The predictive coding framework also allows us to finesse explanations of the role of anxiety in the experience of derealisation. We noted that Cotard described anxiety as part of the aetiology of the depersonalization experience in Cotard delusion. Medford, in an early discussion of DPD, also postulated a role for anxiety in order to explain an apparent paradox of DPD: the distress experienced by the patient at the absence of affective response. It is not merely that the patient has no emotions, but, as a patient of Medford's said, "I don't have any emotions—it makes me so unhappy." Medford (2012) pointed out that this is only slightly paradoxical: the distress is at the lack of *internal* affect, the inability to feel rather than at the derealisation of the external world. Medford related this specifically to the anxiety component of the syndrome. The patient expects that the world will induce positive affect but when it does not an expectation is violated and the patient anxiously attends to that absence of affect. On this view highly anxious patients are hyperattentive to their experience and encounter, not the normal bodily experience, which represents how they are faring in the world, but a strange absence of such experience, in combination with intact exteroception which tells them that the world is unchanged (Paulus & Stein 2010; Garfinkel & Critchley 2013; Seth 2013; Terasawa et al. 2013). Their problem is that they no longer feel the relevance of changes in their own bodies and the world to themselves and this inability to feel the world increases their anxiety. Medford quotes an earlier theorist (Ackner 1954) who noted "increased responsiveness for anxiety of internal origin, whereas that of external origin [is] reduced" (Medford 2012, p. 141).

This perhaps explains the differences in casual aetiology between depersonalisation arising in the Cotard syndrome and in DPD. In the Cotard syndrome something is amiss with the mechanisms that appraise perceptual and interoceptive information for self-relevance. The AIC is not getting any information from affective systems to integrate and relay to higher order cognition. Thus felt significance disappears. When the depressive patient then focuses on her experience she feels alienated from the world and depersonalised. In the case of DPD it appears that the AIC is

hypoactive for another reason: its activity is inhibited by the VLPFC.

In both cases the patient attends to her experience and tries to interpret it in order to respond. This is consistent with the role postulated by predictive coding theories for attention: the attempt to interpret and sharpen the informational content of a signal by improving the signal to noise ratio. Unfortunately an increase in attention does not provide any increase in precision, it only makes the absence of predicted response more salient. Since those predictions are, in effect, representations of expected self-relevance that normally provide the experience of self-awareness, the patient concludes that the self does not exist. After all, the information necessary to generate self-awareness is still in place. The body, the world and first order representations of their interaction are all represented in experience. What is lost is a sense of the significance of those interactions for the body that mediates them.

The explanation has become complicated so at this point it is useful to situate it in terms of the conceptual architecture (points (i)-(iii) below) outlined in the introduction. On this view DPD arises in the following way as a personal level response to the absence of predicted affective experience.

i. Appraisal systems normally represent the significance of information for the organism. The primary way of experiencing the result of those appraisals is via activation in the AIC. This is because the AIC is specialised for informing the subject, via bodily experience, of the affective significance of its encounters with the world. These experiences are not the same as experience of body state *per se but the emotional significance of that body state.*

ii. Those experiences can be rehearsed offline in planning and deliberation to extend the temporal horizon of affective experience. We feel like temporally integrated selves because memory and prospection have affective significance.

iii. Predictive coding architecture has the effect of making discrepancy between anticipated and actual affective feeling highly salient.

iv. In DPD activity in the AIC is inhibited most likely as a result of the involuntary activity of the VLPFC.

v. Consequently the patient has normal perceptual and sensory responses to the world but those responses are not integrated into a bodily representation which informs her of their significance. The world feels derealised or as Medford puts it de-affectualised

vi. However, given the way predictive coding works, the patient actually has a model of the world that predicts activity in the AIC as a result of her perceptual encounters. Thus absence of AIC-produced experience is a prediction error that drives metacognitive responses.

vii. Those responses include increased attention (driven by sub personal mechanisms of resource allocation) to the experience itself as the patient tries to extract further information from it. However, being produced by subpersonal mechanisms the experience is both intractable and inscrutable.

viii. Highly anxious people cannot divert attention from the experience, since anxiety is driven by the need to resolve uncertainty. But the experience is inexplicable and irresolvable.

ix. The patient's personal level interpretation of the experience is of depersonalisation "it feels like it is not happening to me". The interpretation is not a direct report of the experience, which I have argued is more like a total deaffectualisation. It amplifies it.

x. However the form that amplification takes, depersonalisation, is explained by the role such experiences have in creating the normal sense of being a self. We feel we are selves precisely because the significance of the world for our organismic goals is normally computed by appraisal systems and represented in characteristic forms of bodily experience.

8 Anatomy of an avatar

Thomas Metzinger has argued that the persisting self is neither an illusion (in the sense of a perceptual experience whose content is incorrect) nor a genuine entity in the sense of an object existing outside the mind like a body or a neural state. Instead the self is a creature of experience itself, a phenomenal representation constructed by the brain to control the body. This representation is in effect an avatar that unifies experiences of ownership (the sense of the integrity of bodily boundaries), perspective on experience (which I have not talked about in this essay), and selfhood ("a single coherent and temporally stable phenomenal subject".) An especially attractive aspect of Metzinger's view is that he treats the nature of the avatar as an empirical matter so that our understanding of its properties can be refined in the face of further discoveries.

Metzinger's view nicely captures what is right and wrong in the illusory view of the self. The illusory view is correct that the self is not an object to be experienced in the same way as we experience perceptual or somatic objects. The self is a way of experiencing the interaction of the body and the world. It is a creature of experience, constructed by the brain to navigate the organism through the world. The self exists as a virtual phenomenal entity in virtue of the integrative processes that create and sustain it.

The Fat Controller view of the self also has some of the picture correct. Self-awareness is needed for higher order cognitive control to integrate and organise experience moment to moment and to assimilate those experiences to an ongoing autobiography for longer-term cognitive control. However there is no single cognitive process with an identifiable neural substrate that represents an organiser/narrator. Also, and this is where Metzinger is correct, there is a genuine experience of being a person in control, but this experience is the experience of integration itself, which suggests that it is a process which can disintegrate and degrade in different ways and to different degrees. It also suggests, although I have not discussed it here, that experience of the self is a prefrontal achievement since prefrontal structures are specialised for "large world" integrative processing (the orchestration of synchronised activity across widely distributed brain areas).

The Embodied Self view is of course very close to the one I have discussed here. I have argued that a particular type of bodily feeling is what goes awry in depersonalisation and hence that those feelings produce the experience of the self. While this is correct, we need to recall that Damasio distinguished between the "core self", which is very close to the phenomenon I have described, and the autobiographical self. Sometimes he treats the autobiographical self as a more abstract or narrative construct. I have tried to show that the integration of the core self with the autobiographical self comes, as it were, for free, given the automatic links between affective processing and the processes which construct the autobiographical self. It is impossible to rehearse episodes of one's autobiography without a sense of presence–unless, of course, one has DPD or the Cotard delusion. But those cases demonstrate the component structure of the avatar.

Finally, the narrative view captures the crucial role of temporal integration in the experience of the self. But the self is not *just* a fictional protagonist in the brain's stories (though it is that). The specialised simulation mechanisms that generate the actual and potential autobiographies automatically integrate each episode with affective feeling. That feeling allows us to experience in the process of recollection, imagination or narration the significance of each episode to our unique organismic trajectory. That, and the ability to incorporate and act on those feelings, is all the selfhood anyone needs.

References

Ackner, B. (1954). Depersonalization: I. Aetiology and phenomenology. *British Journal of Psychiatry*, *100* (421), 838-853. 10.1192/bjp.100.421.838

Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, *1191* (1), 42-61. 10.1111/j.1749-6632.2010.05445.x

Adolphs, R., Baron-Cohen, S. & Tranel, D. (2002). Impaired recognition of social emotions following amygdala damage. *Journal of Cognitive Neuroscience*, *14* (8), 1264-1274. 10.1162/089892902760807258

Allport, G. W. (1961). *Pattern and growth in personality.* New York, NY: Holt, Rinehart & Winston.

American Psychiatric Association, (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR.* Washington DC: American Psychiatric Association.

Bargh, J. A. & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54* (7), 462-479. 10.1037/0003-066X.54.7.462

Bechara, A. & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, *52* (2), 336-372. 10.1016/j.geb.2004.06.010

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

Breen, N., Coltheart, M. & Caine, D. (2001). A two-way window on face recognition. *Trends in Cognitive Sciences*, *5* (6), 234-235. 10.1016/S1364-6613(00)01659-4

Brighetti, G., Bonifacci, P., Borlimi, R. & Ottaviani, C. (2007). "Far from the heart far from the eye": Evidence from the Capgras delusion. *Cognitive Neuropsychiatry*, *189* (197), 12-3. 10.1080/13546800600892183

Brockington, I. F. & Kumar, R. (1982). *Motherhood and mental illness.* London, UK: Academic Press.

Broyd, S. J., Demanuele, C. & , (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neuroscience and Biobehavioral Reviews*, *33* (3), 279-296. 10.1016/j.neubiorev.2008.09.002

Bubic, A., Von Cramon, D. Y. & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4* (25), 1-15.

Buckner, R. L., Andrews-Hanna, J. R. & , (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*, 1-38. 10.1196/annals.1440.011

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, *121* (483), 753-771. 10.1093/mind/fzs106

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

Cotard, J. (1880). Du délire hypocondriaque dans une forme grave de la mélancolie anxieuse. *Annales Medico-Psychologiques*, *38*, 168-170.

——— (1882). Du délire des négations. *Archives de Neurologie*, *4*, 282-295.

——— (1884). Perte de la vision mentale dans le mélancolie anxieuse. *Archives de Neurologie*, *7*, 289-295.

——— (1891). *Études sur les maladies cérébrales et mentales.* Paris, FR: Baillière.

Craig, A. D. (2009). How do you feel now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10* (1), 59-70. 10.1038/nrn2555

——— (2010). The sentient self. *Brain Structure and Function*, *214* (5), 563-577. 10.1007/s00429-010-0248-y

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *Journal of Comparative Neurology*, *493* (1), 154-166. 10.1002/cne.20749

Currie, G. & Jureidini, J. (2004). Narrative and coherence. *Mind and Language*, *19* (4), 409-427. 10.1111/j.0268-1064.2004.00266.x

Damasio, A. R. (1994). *Descares' error: Emotion, reason, and the human brain.* New York, UK: Putnam.

Debruyne, H., Portzky, M., van den Eynde, F. & Audenaert, K. (2009). Cotard's syndrome: A review. *Current Psychiatry Reports*, *11* (3), 197-202. 10.1007/s11920-009-0031-z

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., Cusack, R., Lawrence, A. D. & Dalgleish, T. (2010). Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21* (12), 1835-1844. 10.1177/0956797610389191

Ellis, H. D. & Lewis, M. B. (2001). Capgras delusion: A window on face recognition. *Trends in Cognitive Sciences*, *5* (4), 149-156. 10.1016/S1364-6613(00)01620-X

Enoch, M. D. & Trethowan, W. H. (1991). *Uncommon psychiatric syndromes.* Oxford, UK: Butterworth-Heinemann.

Fair, D. A., Cohen, A. L. & , (2008). The maturing architecture of the brain's default network. *Proceedings of the National Academy of Sciences of the United States of America*, *105* (10), 4028-4032. 10.1073/pnas.0800376105

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16* (9), 1325-1352. 10.1016/j.neunet.2003.06.005

Füstös, J., Gramann, K., Herbert, B. M. & Pollatos, O. (2013). On the embodiment of emotion regulation: Interoceptive awareness facilitates reappraisal. *Social Cognitive and Affective Neuroscience*, *8* (8), 911-917. 10.1093/scan/nss089

Garfinkel, S. N. & Critchley, H. D. (2013). Interoception, emotion and brain: New insights link internal physiology to social behavior. *Social Cognitive and Affective Neuroscience*, *8* (3), 231-234. 10.1093/scan/nss140

Gerrans, P. (2000). A one-stage explanation of the Cotard delusion. *Philosophy, Psychiatry, and Psychology*, *9* (1), 47-53. 10.1353/ppp.2003.0007

——— (2001). Delusions as performance failures. *Cognitive Neuropsychiatry*, *6* (3), 161-173.

——— (2014). *The measure of madness: Philosophy of mind, cognitive neuroscience, and delusional thought.* Cambridge, MA: MIT Press.

Gerrans, P. & Kennett, J. (2010). Neurosentimentalism and moral agency. *Mind*, *119* (475), 585-614. 10.1093/mind/fzq037

Gilboa, A. (2004). Autobiographical and episodic memory one and the same? Evidence from prefrontal activation in neuroimaging studies. *Neuropsychologia*, *42* (10), 1336-1349. 10.1016/j.neuropsychologia.2004.02.014

Goldie, P. (2011). Life, fiction, and narrative. In N. Carroll & J. Gibson (Eds.) *Narrative, emotion, and insight* (pp. 8-22). University Park, PA: Pennsylvania State University Press.

Gusnard, D. A., Akbudak, E., Shulman, G. L. & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *98* (7), 4259-4264. 10.1073/pnas.071043098

Halligan, P. W. & Marshall, J. C. (1996). *Method in madness: Case studies in cognitive neuropsychiatry.* Hove, UK: Psychology Press.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3* (96), 1-14. 10.3389/fpsyg.2012.00096

——— (2013). *The predictive mind.* Oxford, UK: Oxford University Press.

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

Hunter, E. C., Sierra, M. & David, A. S. (2004). The epidemiology of depersonalisation and derealisation. *Social Psychiatry and Psychiatric Epidemiology*, *39* (1), 9-18. 10.1007/s00127-004-0701-4

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34* (4), 169-188. 10.1017/S0140525X10003134

Jureidini, J. (2012). Explanations and unexplanations: Restoring meaning to psychiatry. *Australia and New Zealand Journal of Psychiatry*, *46* (3), 188-191. 10.1177/0004867412437347

Kenny, A. (1963). *Action, emotion & will.* London, UK: Routledge & Kegan Paul.

Koenigs, M. & Grafman, J. (2009). The functional neuroanatomy of depression: Distinct roles for ventromedial and dorsolateral prefrontal cortex. *Behavioral Brain Research*, *201* (2), 239-243. 10.1016/j.bbr.2009.03.004

Medford, N. (2012). Emotion and the unreal self: Depersonalization disorder and de-affectualization. *Emotion Review*, *4* (2), 139-144. 10.1177/1754073911430135

Metzinger, T. (2003). *Being no one: The self-odel theory of subjectivity.* Cambridge, MA: MIT Press.

——— (2011). The no-self alternative. In S. Gallagher (Ed.) *The Oxford Handbook of the Self* (pp. 279-296). Oxford, UK: Oxford University Press.

Michal, M. & Reuchlein, B. (2014). Striking discrepancy of anomalous body experiences with normal interoceptive accuracy in depersonalization-derealization disorder. *PLoS One*, *9* (2), e89823-e89823. 10.1371/journal.pone.0089823

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J. & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and cognition*, *25* (100), 67-76. 10.1016/j.concog.2014.01.009

Murphy, D. (2006). *Psychiatry in the scientific image.* Cambridge, UK: MIT Press.

N'Diaye, K., Sander, D. & Vuilleumier, P. (2009). Self-relevance processing in the human amygdala: Gaze direction, facial expression, and emotion intensity. *Emotion*, *9* (6), 798. 10.1037/a0017845

Ochsner, K. N., Bunge, S. A. & , (2002). Rethinking feelings: An fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience*, *14* (8), 1215-1229. 10.1162/089892902760807212

O'Connor, A. R. & Moulin, C. J. (2010). Recognition without identification, erroneous familiarity, and déjà

vu. *Current Psychiatry Reports*, *12* (3), 165-173. 10.1007/s11920-010-0119-5

O'Reilly, R. C. & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* Cambridge, MA: MIT Press.

Park, H. J. & Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science*, *342* (6158), 1238411-1238411. 10.1126/science.1238411

Paulus, M. P. & Stein, M. B. (2010). Interoception in anxiety and depression. *Brain Structure and Function*, *214* (5-6), 451-463. 10.1007/s00429-010-0258-9

Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion.* Oxford, UK: Oxford University Press.

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness.* Oxford, UK: Oxford University Press.

Sander, D., Grafman, J. & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences*, *14* (4), 303-316.

Sander, D., Grandjean, D. & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, *18* (4), 317-352. 10.1016/j.neunet.2005.03.001

Scherer, K. R. (2004). Feelings integrate the central representation of appraisal-driven response organization in emotion. In A. S. R. Manstead, N. Frijda & A. Fischer (Eds.) *Feelings and emotions: The Amsterdam Symposium* (pp. 136-157). Cambridge, UK: Cambridge University Press.

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

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, *2* (395), 1-16. 10.3389/fpsyg.2011.00395

Sierra, M., Lopera, F., Lambert, M. V., Phillips, M. L. & David, A. S. (2002). Separating depersonalisation and derealisation: The relevance of the "lesion method". *Journal of Neurology, Neurosurgery & Psychiatry*, *72* (4), 530-532. 10.1136/jnnp.72.4.530

Solomon, R. C. (1976). *The passions: Emotions and the meaning of life.* Indianapolis, ID: Hackett.

——— (1993). The philosophy of emotions. In J. M. Haviland & M. Lewis (Eds.) *Handbook of emotions* (pp. 3-15). New York, NY: Guildford Press.

Spinelli, M. (2009). Postpartum psychosis: Detection of risk and management. *American Journal of Psychiatry*, *166* (4), 405-408. 10.1176/appi.ajp.2008.08121899

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience*, *2* (4), 1-8. 10.3389/neuro.10.004.2008

Strawson, G. (2004). Against narrativity. *Ratio*, *17* (4), 428-452. 10.1111/j.1467-9329.2004.00264.x

Terasawa, Y., Shibata, M., Moriguchi, Y. & Umeda, S. (2013). Anterior insular cortex mediates bodily sensibility and social anxiety. *Social Cognitive and Affective Neuroscience*, *8* (3), 259-266. 10.1093/scan/nss108

Tomkins, S. S. (1962). *Affect, imagery, consciousness vol. 1: The positive affects.* New York, NY: Springer.

——— (1991). *Affect, imagery, consciousness vol. 3: The negative affects: Anger and fear.* New York, NY: Springer.

Young, A. W., Leafhead, K. M. & Szulecka, T. K. (1994). The Capgras and Cotard delusions. *Psychopathology*, *27* (3-5), 226-231. 10.1159/000284874