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

Memory for Prediction Error Minimization: From Depersonalization to the Delusion of Non-Existence

A Commentary on Philip Gerrans

Ying-Tung Lin

Depersonalization is an essential step in the development of the Cotard delusion. Based on Philip Gerrans' account (this collection), which is an integration of the appraisal theory, the simulation theory, and the predictive coding framework, this commentary aims to argue that the role of memory systems is to update the knowledge of prior probability required for successful predictions. This view of memory systems under the predictive coding framework provides an explanation of how experience is related to the construction of mental autobiographies, how anomalous experience can lead to delusions, and thus how the Cotard delusion arises from depersonalization.

Keywords
Affective processing | Cotard delusion | Depersonalization | Memory | Narrative | Predictive coding framework | Self-awareness | Simulation model

Commentator

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

Target Author

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

Editors

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

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

1 Introduction

In Le Délire de Négation (1897), Jules Séglas considers depersonalization to be an essential step in the development of the Cotard delusion[1] (CD; as cited in Debruyne 2009; Gerrans 2002), and *prima facie* the two states share a number of characteristics: Patients suffering from the former feel *as if* they are dead or do not exist,

1 According to Berrios & Luque (1995), the English translation of "le délire des negations"—a term first introduced by the French neurologist, Jules Cotard (1840–1889)—only conveys a part of what it means: "Délire is not a state of delirium or organic confusion (in French, *délire aigu* and *confusion mentale*) or a delusion (in French, *idée* or *thème délirante*)—it is more like a syndrome that may include symptoms from the intellectual, emotional, or volitional spheres" (p. 219). The original French concept of "délire" fits better with Gerrans' account of the Cotard delusion, in which the Cotard delusion does not merely concern beliefs of denial, but also anomalous affective processing.

whereas those who suffer from the latter sincerely believe and experience this state. However, the central characteristics of these disorders are distinct. Patients describe the experience of depersonalization as follows:

> It's really weird. It's sort of like I'm here, but I'm really not here and that I kind of stepped out of myself, like a ghost... I feel really light, you know. I feel kind of empty and light, like I'm going to float away... Sometimes I really look at myself that way... It's kind of a cold eerie feeling. I'm just totally numbed by it. (Cited in Steinberg 1995)

> The emotional part of my brain is dead. My feelings are peculiar, I feel dead. Whereas things worried me nothing does now. My husband is there but he is part of the furniture. I don't feel I can worry. All my emotions are blunted. (Shorvon 1946, p. 783)

As illustrated in these subjective descriptions, depersonalization is characterized by a loss of the sense of presence (Critchley 2005) or an increased "sense of detachment"—the "[e]xperience of unreality, detachment, or being an outside observer with respect to one's thoughts, feelings, sensations, body, or actions" (American Psychiatric Association 2013, p. 302). On the other hand, in the Cotard delusion (CD), mental autobiographies are acutely distorted—in such a way that patients are convinced that they are dead or that they do not exist:

> An 88-year-old man with mild cognitive impairment was admitted to our hospital for treatment of a severe depressive episode. He was convinced that he was dead and felt very anxious because he was not yet buried. This delusion caused extreme suffering and made outpatient treatment impossible. (Debruyne et al. 2009, p. 197)

Researchers in the field of delusion studies have debated the way in which anomalous experience leads to false belief. In this commentary I am interested in the following questions: What cognitive architecture could, in principle, explain CD in terms of its development from depersonalization, and what exactly are the underlying differences between patients suffering from the Cotard delusion and those suffering from depersonalization disorder (DPD) but free from the Cotard delusion?

In his target paper, Gerrans explores the cognitive structure of self-awareness—the "awareness of being a unified persisting entity" (this collection, p. 2). To explain the emergence of self-awareness and its loss in DPD and CD, he provides an account that integrates the appraisal theory of emotion, the simulation model of memory and prospection, and the hierarchical predictive coding model. First, based on the appraisal theory, Gerrans shows that the activation of the anterior insular cortex (AIC) allows an organism to experience the emotional significance of a relevant state by experiencing appraisal. According to Gerrans, these reflexive processes are what sustain the self from moment to moment: "An organism that can use that affective information in the process is a self" (this collection, p. 8). Second, the integration of affective processing and simulated episodes allows the organism to experience itself as a persisting entity overtime (see more below). Last, he endorses the predictive coding framework, according to which the human mind can be accounted for by the principle of predictive error minimization. Perception, for instance, is realized by the operation of both top-down prediction and bottom-up predictive error. If the general theoretical model is correct, it will not only apply to perception, but also to affective processing (ibid., p. 9). Gerrans (this collection) applies this framework to explain the phenomenon of depersonalization and CD: Depersonalization occurs due to a failure to attribute emotional relevance to bodily states, which results from hypoactivity of the AIC. The prediction error from the mismatch between the predicted and the actual activation level of the AIC would lead to allocation of attention, the function of which, according to the predictive coding framework, is to disambiguate signals. If the prediction error cannot be cancelled and attention

cannot be diverted, increased attention brings about anxiety in DPD and CD, which is "an adaptive mechanism that primes the organism cognitively and physiologically to solve uncertainty" (ibid., p. 11). This is reflected in the patients' subjective reports concerning the loss of awareness of their bodies. This integrated theory provides an explanation of depersonalization as well as of how self-awareness is constructed through the interaction of different forms of cognitive processing.

In Gerrans' account, the simulation system allows the organism "to *simulate* temporally distant experiences by rehearsing some of the same perceptual and emotional mechanisms activated by the simulated situation" (ibid., p. 6), such that the affective associations result in integrated episodes of experience that lead to the feeling of persisting over time. I argue (1) that the simulation model should not be thought of as independent from other memory systems: without memory systems at lower levels—semantic and procedural memory systems—the simulated episodes cannot be constructed (section 2); and (2), that by considering the role of memory under the predictive coding framework, the simulation model not only plays a role in simulating temporally-distant episodes but also contributes to the knowledge required for the creation of predictive models in the present (section 3). On such a view of the simulation model, delusion can be explained and I will suggest (3) two factors contributing to the development of CD from depersonalization: the compromised decontextualized supervision system and the expectation of high precision from interoceptive signals (section 4); that is, only if these two factors are present in a depersonalized subject may CD develop.

2 The simulation model and the mental autobiography

> [W]e are all virtuoso novelists, who find ourselves engaged in all sorts of behavior, more or less unified, but sometimes disunified, and we always put the best 'faces' on it we can. We try to make all of our material cohere into a single good story. And that story is our autobiography. (Dennett 1992, p. 114)

As persons, our beliefs and desires are structured in a more or less coherent fashion, such that a mental autobiography—an autobiographical framework (Gerrans 2013) or narrative (Schechtman 1996)—can be attributed, which can explain our cognitive structure. Many people have proposed theoretical entities such as the "autobiographical self" (Damasio 1999, 2010), the "conceptual self" (Conway 2005; Conway et al. 2004), and the "narrative self" (Feinberg 2009), etc. to account for how one comprehends and navigates through the world and over time—that is, how one is able to make sense of external or internal signals, to have preferences, to have goals and to values, to know who oneself is, to be a diachronically persisting agent, to recall the past, and to imagine the future. In general, these different versions of the "extended self" (Gallagher 2000) are characterized by the following phenomenal and epistemic properties.

First, phenomenally, we experience ourselves as thinkers of thoughts (e.g., "I think…" or "I believe…") and as beings who recollect the past and plan for the future; while at the same time we have a sense of ownership of relevant beliefs (e.g., "this thought is mine"). Second, subjectively, events and objects are presented in a way that manifests their relevance to the subject. In addition, epistemically, we tend to treat the self-told story as if it were highly reliable: The content is treated as objectively real, and its truthfulness is seldom questioned. This is the way we consciously comprehend the world and our place within it, and it is thought to be reliable. Accompanied by a certain degree of the "feeling of familiarity" and the "sense of pastness" (Russell 2009, p. 208), there is a degree of certainty about the veridicality of a mental autobiography. When inconsistency or non-veridicality is detected and such certainty is lost (e.g., due to introspection or contradiction to external information), the mental autobiography will be modified to re-create a new subjective reality—a new story about ourselves with more or less difference (e.g., self-deception).

Delusional patients have anomalous forms of mental autobiography: Their mental autobiographies are radically distorted, for different reasons. For instance, RZ, a 40-year-old female patient with reverse intermetamorphosis, believed that she was her father (and sometimes believed that she was her grandfather) during her assessment by Breen et al. (2000). When asked to sign her name and answer questions about her life, she signed her father's name and provided her father's personal history. She acted according to her delusional beliefs. Here we see that her mental autobiography constructs her subjective reality. Semantic dementia patients who suffer from an incapability of constructing personal futures (Irish & Piguet 2013) provide examples of the loss of partial subjective reality.[2] It is speculated that this form of futureless mental autobiography accounts for the higher suicide rate in semantic dementia (Hsiao et al. 2013). As we will see, patients suffering from CD also maintain a mental autobiography.[3] They believe that they are dead or no longer exist: They may refuse to eat or visit the graveyard—the place in which they believe they belong. But how are mental autobiographies constructed? The rest of this section considers how memory systems and simulation models lead to the construction of a mental autobiography.

Studies on misrepresentations in memory have suggested that—against the traditional and folk-psychological idea of a "store-house" (Locke 2008), in which memory as a copy of past experience is stored for future use—memory is constructive in nature. It represents different facets of experience, which are distributed across different regions of the brain, where retrieval is realized in a process of pattern completion, which allows a subset of features to comprise a past experience (Schacter et al. 1998). The prevalence of misremembering (episodic memory in particular) and the view of constructive memory have led to the debate over the function of memory: If the proper function of memory is to veridically represent past experiences or events, is our memory system fundamentally defective? Or, does it serve other functions? If there is any adaptive advantage of memory systems, they must serve a function that concerns the *current and/or future* states of the organism (Westbury & Dennett 2000). New findings regarding a default-mode network suggest a "constructive episodic simulation hypothesis" (Schacter & Addis 2007a, 2007b), according to which the constructive nature of episodic memory is partially attributable to its proposed role in mentally simulating our personal futures (e.g., planning a future event). This hypothesis is supported by fMRI evidence showing that remembering the past as well as imagining the possible past and future share correlates with the activities of the default mode network (Addis et al. 2007; De Brigard et al. 2013; Szpunar et al. 2007). Therefore, it is suggested that episodic memory is adaptive in that it allows us to employ past experiences in such a way as to enable simulations of possible future episodes.

However, simulation is not realized by episodic memory alone. Though memory systems (i.e., procedural, semantic, and episodic memory) can be conceptually distinguished, they are considered parts of a "monohierarchical multimemory systems model" (Tulving 1985): Semantic memory is a specialized subsystem of procedural memory that lies at the lowest level of the hierarchy, and semantic memory in turn contains episodic memory as its specialized subsystem. The subsystems at higher levels are dependent on and supported by those at lower levels. That is, our everyday autobiographical memory is realized by multiple memory systems. For instance, a recent study has shown the importance of semantic memory in the construction of autobiographical memory: While episodic memory provides episodic details, semantic memory acts as a schema for integrating them (Irish & Piguet 2013). That is, our mental autobiographies are constructed by the interplay of multiple memory systems (e.g., Tulving's SPI model, see Tulving 1995).

2 If the predictive coding framework and the role of memory for which I argued in section 3 is correct, one should expect to find an anomalous phenomenon in semantic dementia—not only with respect to one's narrative consciousness, but also with respect to one's perception.

3 It might be a contradiction in terms to claim that patients suffering from the Cotard delusion have mental autobiographies, since "auto" means "self, one's own" and "bio" means "life". Here, it can merely be understood as a personal-level response to the system's condition.

This applies to prospection as well. Different categorizations of prospection are proposed (e.g., Atance & O'Neill 2001; Szpunar et al. 2014). In this commentary, I adopt a distinction offered by Suddendorf & Corballis (2007), who distinguish procedural, semantic, and episodic prospection (p. 301, Figure 1). Suddendorf & Corballis (2007) suggest that the function of the memory and anticipatory systems is to provide behavioral flexibility; and they also examine the phylogenetic development of different memory systems. According to their model, the flexibility of anticipatory behavior supported by different memory systems can offer varies in degree. From the primitive form, procedural memory enables stimulus-driven predictions of regularities and allows behavior to be modulated by experience, such that the resulting behavior is stimulus-bound. Declarative memory provides more flexibility because it can not only be retrieved involuntarily, but can also be voluntarily triggered top-down from the frontal lobe, which enables decoupled representations that are not directly tied to the perceptual system. That is, even though we are still tied to the present in that we recall and imagine the future at the present moment, the content of representation can extend beyond the current immediate environment. Specifically, semantic memory is considered more primitive than episodic memory as it has less scope for flexibility (Suddendorf & Corballis 2007).[4] The former, in allowing learning in one context to be voluntarily transferred to another, provides the basis for reasoning. However, this is about regularities and not particularities. Episodic memory supplements this weakness: A scenario can be simulated and pre-experienced. Through mental reconstruction or memory construction, episodic memory not only recreates past events, it also allows the learned elements to be incorporated and arranged in a particular way in order to simulate possible futures. It thereby provides greater flexibility in novel situations and provides for the possibility of making long-term plans, extending even beyond the life-span of the individual.

To sum up, our mental autobiography is constructed through the interaction of multiple memory systems at different levels. The simulation model should not only be associated with the episodic memory system; rather, it should be understood as a hierarchical model of multiple memory systems—i.e., procedural, semantic, and episodic memory as well as procedural, semantic, and episodic prospection. In the next section I will consider the role of memory systems within the predictive coding framework.

3 Memory under the predictive coding framework

Recent development of the predictive coding framework (Clark 2013b, this collection; Friston 2003; Hohwy this collection) provides an integrated conceptual framework for perception and action. According to the framework, the brain constantly attempts to minimize the discrepancy between sensory inputs (including exteroceptive and interoceptive signals) and the internal models of the causes of those inputs via reciprocal interactions between hierarchical levels. Each cortical level employs a generative model to predict representations of the subordinate level, to which the prediction is sent via top-down projections—the bottom-up signal is the prediction error. Prediction error minimization can be achieved in a number of ways (Clark this collection; Hohwy this collection); but in general, errors can be minimized either by updating generative models to fit the input or by carrying out actions to change the world to fit the model. In the target paper, Gerrans integrates appraisal processing into the predictive coding framework; however, he treats only the simulation model as a mechanism for simulating temporally distant experiences (this collection, pp. 6–8). In this section, I propose that under the predictive coding framework, the simulation model serves the function of updating

4 Tulving (2005) and Suddendorf & Corballis (2007) argue that episodic memory emerges later in the course of evolution and belongs uniquely to human beings. Even if there is evidence suggesting the existence of episodic-like memory—memory encoding "what", "where", and "when" information—in non-human creatures (e.g., Western scrub jays; Clayton 2003; Clayton & Dickinson 1998), Tulving (2005) argues that these phenomena can be explained merely by semantic memory. In a recent paper, Corballis (2013) changes the claim he makes in the earlier article (Suddendorf & Corballis 2007) and argues that mental time travel also exists in rats, and that the difference between this and human mental time travel is simply the degree of complexity.

the knowledge required for successful prediction, which constitutes perception and affective experience.

How can we understand the role of memory or the simulation system under the predictive coding framework?[5] Here I examine how memory systems can be incorporated into the framework. According to the predictive coding framework, perceiving is distinct from the traditional model of perception; instead, it is:

> to use whatever stored knowledge is available to guide a set of guesses about [...external causes], and then to compare those guesses to the incoming signal, using residual errors to decide between competing guesses and (where necessary) to reject one set of guesses and replace it with another. (Clark 2013a, p. 743)

That is, perception is knowledge-driven and top-down, rather than stimulus-driven and bottom-up. "Stored knowledge" refers to a repertoire of prior beliefs or knowledge—the belief of the likelihood of a hypothesis or guess irrespective of sensory input. It is acquired or shaped by learning from past experience—or, in other words, it is a modification of parameters in order to minimize prediction error.[6]

Moshe Bar (2009) suggests that "our perception of the environment relies on memory as much as it does on incoming information" (p. 1235). Since we seldom encounter completely novel objects or events, our systems rely on representations stored in memory systems to generate predictions. According to Bar's "analogy-as-association-prediction" framework (Bar & Neta 2008), once there is a sensory input, the brain actively generates top-down guesses in order to figure out what that input looks like (analogy); the match triggers activation of associated representations (association), which allows predictions of what is likely to happen in the relevant context and environment (prediction). Thus, instead of aiming to answer the question "what is this?", perception studies should answer the question "what is this *like*?" or "what does this resemble?": Brains proactively compare incoming signals with existing information gained in the past (see Bar 2009, Figure 1 & Figure 2). Bar (2009) suggests that predictions also influence memory encoding. Memory systems primarily encode that which differs from memory-based prediction, and if sensory information meets the prediction, the information is less likely to encode (Bar 2009, p. 1240).

This account provides a new view of the concepts of encoding, retrieval, and reconsolidation. The older view describes encoding as the process by which incoming information is stored for later retrieval, and retrieval as a process involved in utilizing encoded information in reviving past events. Nevertheless, under the predictive coding framework, when discrepancy between prediction and perceptual information occurs, encoding is the process of minimizing prediction error—the adaptation of the model to reduce discrepancy based on the forward-feeding, bottom-up input from its subordinate level. Retrieval is then regarded as the process of utilizing this knowledge for predictive model construction.

Accordingly, I suggest that the role of memory systems is to update the knowledge required for successful predictions of the organism's current (and future) informational state. That is, under the predictive coding framework, our perception is knowledge-driven, and knowledge is experience-based. The mechanisms of our memory systems allow the knowledge required for the construction of predictive models to be updated based on experience. Prediction error can trigger encoding that modifies our knowledge, which then optimizes the predictive model to achieve prediction error minimization. In addition, as we will see later in this section, the development of episodic memory and mind-wandering allows us to generate new knowledge.

This knowledge-driven perception is realized by a multi-layer hierarchical structure in

5 Felipe De Brigard (2012) considers how the predictive coding framework can predict remembering. He modifies Anderson's Adaptive Control of Thought-Rational model (Anderson & Schooler 2000); here the probability of a memory retrieval can be calculated based on how well memory retrieval will minimize prediction error given the cost of the retrieval and the current context. Here, however, I shall not consider the retrieval of individual memories; instead I focus on the role of the memory systems within the framework.

6 See Clark (2013a) for the problem of the acquisition of the very first prior knowledge.

which "each layer is trying to build knowledge structures that will enable it to generate the patterns of activity occurring at the level below" (Clark 2013a, p. 483). The information encoded at each level is distinct: At higher hierarchical levels, the representations become more abstract and involve a larger spatial and temporal dimension: The predictive models generated not only represent the immediate state of the system or environment but also the system in relation to the spatially and temporally-extended environment. Moreover, the higher-level knowledge also supports predicting how sensory signals will change and evolve over time. It allows one to predict the future and execute long-term plans involving multiple steps. The hierarchical structure is crucial to our capacity to comprehend the world, which is highly structured, with regularity and patterns at multiple spatial and temporal scales and interacting and complexly-nested causes (Clark 2013a).

I suggest that each level of knowledge has an updating mechanism, which is consistent with Tulving's (1985) monohierarchical multimemory systems model and Suddendorf & Corballis' (2007) model of memory and prospection. Procedural memory at the lowest level is involved in the sensori-motor predictive function: It updates the procedural knowledge required for predicting the states in which given actions are executed. Whereas implicit memory is mainly involved in immediate responses to current stimuli, declarative or explicit memory (episodic memory in particular) contributes to the construction of a model of the system itself and its environment with spatial and temporal dimensions. It supplements higher-level knowledge structures for the construction of a generative model, which explains actual states and predicts possible changes and actions for reaching desired states. Under the predictive coding framework, the semantic memory system, which allows learning in one context to be transferred to another, supports semantic knowledge, which in turn provides regularities in the construction of predictive models (e.g., during reading). And episodic memory, together with semantic memory, supports the knowledge required to construct a model of one's autobiography—a model of one's own relevant past and potential future. However, it is worth noting that our mental autobiography is not realized by knowledge at a single hierarchical level; instead, it is constructed through the interplay of the mechanisms at multiple levels.

In addition to its contribution to an autobiographical-scale model, episodic memory, along with other memory systems, also generates new knowledge by simulation. Bar (2007) proposes that:

> [the] primary role [of mental time travel] is to create new 'memories'. We simulate, plan and combine past and future in our thoughts, and the result might be 'written' in memory for future use. These simulated memories are different from real memories in that they have not happened in reality, but both real and simulated memories could be helpful later in the future by providing approximated scripts for thought and action. (p. 286)

This is supported by the evidence that mind-wandering—that is, having thoughts that are unrelated to the current demands of the external environment (Schooler et al. 2011)—is beneficial to autobiographical planning and creative problem solving (Mooneyham & Schooler 2013).[7]

The role of memory systems under the hierarchical predictive coding framework is consistent with the function of memory and the concept of a memory system proposed by De Brigard (2013). Following Carl F. Craver's idea of a mechanistic role function (2001), De Brigard argues for a larger cognitive system of "episodic hypothetical thinking", which includes

7 This is related to the philosophical debate on whether one can gain new knowledge from imagination or a purely mental activity, as was famously denied by Sartre (1972) and Wittgenstein (1980) (for a general discussion, see Stock 2007). It is worth noting that if the predictive coding framework is correct, the concept of "knowledge" may be revised: Knowledge may depart from veridicality; instead, it is close to information that can provide successful predictions. Thus, under the predictive coding framework, the only kind of knowledge Sartre recognizes (as cited in Stock 2007, p. 176)—observational knowledge—is not substantially different from other kinds of knowledge, because the knowledge gained through perception cannot be conceptually distinguished from those that are not: Gaining knowledge at each level is all about optimizing the predictions of lower levels.

future simulation and past counterfactual simulations: To determine the mechanistic function of memory we require an investigation into the way that its components contribute to the system, and then of how memory contributes to the functioning of the organism, helping it to reach goals at higher levels. It is worth noting that these concepts of memory function and malfunction are different to traditional ones: The distinction between memory function and malfunction is not equivalent to the distinction between remembering and misremembering or veridical representation and misrepresentation. Under the predictive coding framework, memory function can be regarded as updating knowledge for predictive model construction. Likewise, memory function and malfunction are independent from the generation of a predictive model that succeeds or fails in representing the world. That is, certain misrepresentations can lead to error minimization; furthermore, it is possible for misrepresentation rather than veridical representation to lead to a generative model.

4 From depersonalization to Cotard delusion

If the predictive coding framework is correct, it provides a new view not only on memory function but also on how we think about memory systems and the relation between memory and other cognitive systems. This framework provides a theory about the role of simulation models in the relationship between reflexive forms of self-consciousness and the narrative self (Hohwy 2007). It provides a theoretical explanation of the finding that memory systems are also involved in perception[8] and interoception. This implies that we not only simulate offline (e.g., mental time travel, mind-wandering), but also simulate online. The simulated model provides us with a subjective reality through which we see the external world and ourselves. It is transparent and immediate: We experience it as objectively real and we directly interact with what is represented.

8 This is consistent with the evidence that memory influences perception (e.g., Summerfield et al. 2011).

However, this characteristic is absent in patients suffering from depersonalization. Depersonalization is an example of how one can become detached from one's simulated model of oneself: One's mental autobiography is no longer direct, and one experiences a sense of distance from the model.[9] Gerrans (this collection) suggests that the loss of sense of presence in depersonalized patients results from a failure to minimize prediction error from the hypoactivity of the AIC—the activation of which informs us of the significance of external or internal information. Gerrans' theory is based on Seth et al.'s idea of interoceptive inference (or interoceptive predictive coding; see also Seth this collection), according to which predictive coding not only applies to exteroception but also to interoception, and emotional states, including the sense of presence, arise from interoceptive prediction successfully matched to actual interoceptive signals (Seth 2013; Seth et al. 2011). As it is suggested that the AIC is suggested to be the correlate of the integration of exteroceptive and interoceptive signals and that it plays a role in maintaining a salience network for the relevant states, the hypoactivity of the AIC leads to the failure to associate affective significance with bodily states. As Gerrans suggests, "not all higher level control systems can and do smoothly cancel prediction errors generated at lower levels" (this collection, p. 9). Because the coding formats at each level are distinct, the coding format of low-level processing is opaque to introspection (p. 9). The problems faced by depersonalized patients can be accounted for by the prediction error based on persisting, unexpected hypoactivity. Attention is then directed towards resolving the prediction error. Gerrans' proposal is that an inability to explain away the surprisal and this increased attention causes anxiety in DPD. Here, CD can be seen as a strategy for some systems to react to anxiety in order to minimize the prediction error.

As Gerrans suggests, "[d]elusions are best conceptualized as higher-level responses to pre-

9 In contrast to depersonalization, derealization refers to the "[e]xperiences of unreality of detachment with respect to surroundings" (American Psychiatric Association 2013, p. 302)—patients suffer from detachment from the simulated model of the environment.

diction error which, however, cannot cancel those errors" (this collection, p. 10). That is, even though not all prediction error can be successfully cancelled, the brain—the organ that constantly minimizes prediction error, according to predictive coding framework—still tries to modify its model in order to decrease surprisal, though unsuccessfully. If what I have suggested in the last section is correct, the function of memory systems is to update knowledge contributing to the construction of predictive models in order to minimize prediction error. The anomalous model of CD is thus one constructed by the hierarchical simulation model to match the hypoactivity of the AIC—the loss of appraisal that represents the significance of self-related information. To construct a model in which oneself is dead or does not exist cannot successfully explain away the prediction error—since one still has the experience of a bodily state—it may nevertheless be the best solution the given system can come up with in order to cope with the increased anxiety resulting from increased attention.

However, this still leaves us with the question of why some depersonalized patients develop CD, whereas most of them do not develop this delusion. Gerrans (2014) suggests that the difference between delusional and non-delusional minds lies in differences in the default mode network, which include information that triggers activity, hyperactivity, and hyperconnectivity, interaction with the salience system, and absent or impaired "decontextualized supervision" (pp. 73–74). Decontextualized supervision allows one to "reason about *oneself* using impersonal, objective rules of inference" (p. 76).[10] The activity of its circuit is anti-correlated with the activity of the default mode network (pp. 83–84) because of the limited cognitive resources for high-level metacognitive processes. Gerrans suggests that delusional thoughts arise from the system's failure to balance this allocation; thus they slip through the supervision system.

Nevertheless, the existence of decontextualized supervision explains how anomalous forms of predictive models—which would be suppressed in non-delusional subjects—could emerge, but it does not account for the model's relation to anomalous experience or to the way in which the content of delusion is constructed (e.g., Cotard delusion). I therefore propose that a delusional mind does not only result from a compromised decontextualized supervision; it also results from an aberrant precision expectation[11] of exteroceptive or interoceptive signals. Jakob Hohwy (2013) proposes the notion of uncertainty expectations: We predict the causal structure of the world (and of one's own bodily state), as well as the level of uncertainty in the environment, which allows us to respond to the external environment under various levels of uncertainty. The strength of prediction error is proportional to the expected certainty: When the uncertainty level is expected to be higher (due to external or internal noise), the prior model is weighted higher, whereas expected low uncertainty gives more weight to bottom-up prediction error. According to Hohwy (2013), delusion arises when precision expectation is either too high or too low, and those in between would report only the anomalous experience, without forming a delusion. In the case of Cotard delusion developed from depersonalization, when one has the expectation of high precision, the system tends to be driven by the bottom-up predictive error of unexpected hypoactivity of the AIC, rather than the prior model. One is, therefore, more likely to revise the model in order to explain away the surprisal resulting from the mismatch between the actual and predicted activation level of the AIC; that is, the systems of patients suffering from CD are driven by an urge to modify their top-down predictive models in order to conform to the loss of AIC activity. The construction of the model in CD is considered an attempt to minimize prediction error.

Finally, explaining delusion under the predictive coding framework provides new understanding to the debate between one- and two-

10 The system of decontextualized supervision is distinct from the semantic memory system discussed in the last section: The latter provides objective elements for the construction of a contextualized autobiographical episode, while the former supervises autobiographical episodes by utilizing decontextualized reasoning.

11 "Precision" is also used to refer to the precision of inferences about hidden causal structures (e.g., in Friston et al. 2013). Here and in Hohwy (2013) it indicates the precision of incoming signals.

stage models of delusion. The one-stage model holds that anomalous experience only suffices to explain the occurrence of delusion (Gerrans 2002; Maher 1974, 1988); according to two-stage model, however, other cognitive disruption is required to explain the content of the delusion in particular (Young & De Pauw 2002). However, if the predictive coding framework is correct, the clear distinction between experience and rationalization assumed in the traditional discussion does not exist: Perception, cognition, and action are now considered continuous and highly integrated (Clark 2013b; Hohwy & Rajan 2012). Experience and rationalization are different layers of abstraction within the very same process of prediction error minimization under the predictive coding framework.

5 Conclusion

In his target paper, Philip Gerrans proposes a theory of self-awareness that integrates the predictive coding framework, the appraisal theory, and the simulation model. It accounts for the loss of self-awareness in DPD and CD, and provides a new understanding of patients' anxiety. In this commentary, I have proposed (1) that the simulation model should be considered a hierarchical model involving multiple memory systems—namely, it is constituted by procedural, semantic, and episodic memory and prospection (section 2); and (2) that the function of memory systems or simulation models, under the predictive coding framework, is to update the knowledge required for successful prediction (section 3). This implies that memory function and malfunction are independent from the generation of a predictive model that succeeds or fails in representing the world, since it is possible that misrepresentation rather than veridical representation leads to a generative model that minimizes prediction error. Based on such view of the simulation model, CD can be regarded as the modification of top-down prediction in an attempt to explain away the prediction error resulting from unexpected hypoactivity of the AIC. I also suggested (3) that a combination of two factors is necessary for the occurrence of CD from depersonalization: the compromised decontextualized supervision system and the expectation of high precision of interoceptive signals (section 4).

If both the general framework and my suggestions are correct, there are a number of issues worthy of further investigation: First, if the model that explains the symptoms of CD is created by the system in order to minimize prediction error from hypoactivity of the AIC, with the aim of affording relief from anxiety, it is expected that the change of prediction may be accompanied by minimized prediction error or/and prediction error from other unpredicted activities. In the case of Cotard delusion, the new model—the model of the organism's death or non-existence—would encounter new kinds of prediction error due to information about bodily states, instead of a lack of emotional significance. This may as well be the kind of prediction error that cannot be cancelled top-down and which can be expected to lead to anxiety based on Gerrans' theory. Therefore, the anxiety characteristic of the Cotard delusion is speculated to be the result of different prediction errors from patients suffering from Cotard syndrome. Studies on the difference between the anxiety present in DPD and that in CD can the support or refutation of the framework proposed. Furthermore, it is worth noting that not all patients with the CD suffer from anxiety. For example, in Berrios & Luque's (1995) analysis of 100 cases, anxiety is reported in only 65% of subjects, and patients were categorized: Cotard type I patients showed no affective component, whereas type II patients showed depression and anxiety. Can the proposed framework account for both types of patients?

Another interesting question for future research is whether we can better understand the relation between the simulation model and affective processing within the predictive coding framework, and whether an explanation of this would be consistent with the existing evidence relating to emotional memory (e.g., LaBar & Cabeza 2006). Affective processing can influence encoding and retrieval of memories, whereas simulating possible episodes is thought to help rehearse affective responses. One possible avenue might be the investigation of the influence

of different forms of simulation on affective processing (e.g., memory retrieval from a field or an observer perspective; Berntsen & Rubin 2006), and further on one's awareness of one's future and past (Wilson & Ross 2003): How can this be accounted for by the principle of prediction error minimization? Does the simulation of potential affective responses optimize prediction and reduce potential error in the future? The simulation and integration of future potential changes into the model of one's autobiography is thought to potentially contribute to the prevention of dramatic changes in one's model at higher levels, and to maintain mental autobiographies that are more consistent across time.

Acknowledgments

I am grateful to Thomas Metzinger and Jennifer M. Windt, as well as two reviewers, for their critical and constructive comments.

References

Addis, D. R., Wong, A. T. & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45* (7), 1363-1377. 10.1016/j.neuropsychologia.2006.10.016

American Psychiatric Association, (2013). *The diagnostic and statistical manual of mental disorders.* Arlington, VA: American Psychiatric Publishing.

Anderson, J. R. & Schooler, L. J. (2000). The adaptive nature of memory. In E. Tulving & F. I. M. Craik (Eds.) *The Oxford handbook of memory* (pp. 557-570). New York, NY: Oxford University Press.

Atance, C. M. & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, *5* (12), 533-539. 10.1016/s1364-6613(00)01804-0

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11* (7), 280-289. 10.1016/j.tics.2007.05.005

——— (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364* (1521), 1235-1243. 10.1098/rstb.2008.0310

Bar, M. & Neta, M. (2008). The proactive brain: Using rudimentary information to make predictive judgments. *Journal of Consumer Behaviour*, *7* (4-5), 319-330. 10.1002/cb.254

Berntsen, D. & Rubin, D. C. (2006). Emotion and vantage point in autobiographical. *Cognition and Emotion*, *20* (8), 1193-1215. 10.1080/02699930500371190

Berrios, G. E. & Luque, R. (1995). Cotard's syndrome: Analysis of 100 cases. *Acta Psychiatrica Scandinavica*, *91* (3), 185-188. 10.1111/j.1600-0447.1995.tb09764.x

Breen, N., Caine, D., Coltheart, M., Hendy, J. & Roberts, C. (2000). Towards an understanding of delusions of misidentification: Four case studies. *Mind & Language*, *15* (1), 74-110. 10.1111/1468-0017.00124

Clark, A. (2013a). Expecting the world: Perception, prediction, and the origins of human knowledge. *Journal of Philosophy*, *110* (9), 469-496.

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

——— (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Clayton, N. S. & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, *395* (6699), 272-274. 10.1038/26216

Clayton, N. S., Bussey, T. J. & Dickinson, A. (2003). Can animals recall the past and plan for the future? *Nature Reviews Neuroscience*, *4* (8), 685-691. 10.1038/nrn1180

Conway, M. A. (2005). Memory and the self. *Journal of memory and language*, *53* (4), 594-628. 10.1016/j.jml.2005.08.005

Conway, M. A., Meares, K. & Standart, S. (2004). Images and goals. *Memory*, *12* (4), 525-531. 10.1080/09658210444000151

Corballis, M. C. (2013). Mental time travel: A case for evolutionary continuity. *Trends in Cognitive Sciences*, *17* (1), 5-6. 10.1016/j.tics.2012.10.009

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, *68* (1), 53-74. 10.1086/392866

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

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness.* New York, NY: Harcourt Brace.

——— (2010). *Self comes to mind: Constructing the conscious brain.* New York, NY: Pantheon.

De Brigard, F. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*, *3* (420). 10.3389/fpsyg.2012.00420

——— (2013). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191* (2), 1-31. 10.1007/s11229-013-0247-7

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L. & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, *51* (12), 2401-2414. 10.1016/j.neuropsychologia.2013.01.015

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

Dennett, D. C. (1992). The self as a center of narrative gravity. In F. S. Kessel, P. M. Cole & D. L. Johnson (Eds.) *Self and consciousness: Multiple perspectives* (pp. 103-115). Hillsdale, NJ: Lawrence Erlbaum Associates.

Feinberg, T. E. (2009). *From axons to identity: Neurological explorations of the nature of the self.* New York, NY: WW Norton & Company.

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

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. [Hypothesis & Theory]. *Frontiers in Human Neuroscience*, *7* (598). 10.3389/fnhum.2013.00598

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, *4* (1), 14-21. 10.1016/S1364-6613(99)01417-5

Gerrans, P. (2002). A one-stage explanation of the Cotard delusion. *Philosophy, Psychiatry, & Psychology*, *9* (1), 47-53. 10.1353/ppp.2003.0007

——— (2013). Delusional attitudes and default thinking. *Mind & Language*, *28* (1), 83-102. 10.1111/mila.12010

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

——— (2015). All the self we need. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche*, *13* (1), 1-20.

——— (2013). Delusions, illusions and inference under uncertainty. *Mind & Language*, *28* (1), 57-71. 10.1111/mila.12008

——— (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Hohwy, J. & Rajan, V. (2012). Delusions as forensically disturbing perceptual inferences. *Neuroethics*, *5* (1), 5-11. 10.1007/s12152-011-9124-6

Hsiao, J. J., Kaiser, N., Fong, S. & Mendez, M. F. (2013). Suicidal behavior and loss of the future self in semantic dementia. *Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology*, *26* (2), 85-92. 10.1097/WNN.0b013e31829c671d

Irish, M. & Piguet, O. (2013). The pivotal role of semantic memory in remembering the past and imagining the future. *Frontiers in Behavioral Neuroscience*, *7* (27). 10.3389/fnbeh.2013.00027

LaBar, K. S. & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7* (1), 54-64. 10.1038/nrn1825

Locke, J. (2008). *An essay concerning human understanding.* Oxford, UK: Oxford University Press.

Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology*, *30* (1), 98-113.

——— (1988). Anomalous experience and delusional thinking: The logic of explanations. In T. F. Oltmanns & B. A. Maher (Eds.) *Delusional beliefs* (pp. 15-33). Oxford, UK: John Wiley & Sons.

Mooneyham, B. W. & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67* (1), 11-18. 10.1037/a0031569

Russell, B. (2009). *The analysis of mind.* Auckland, NZ: The Floating Press.

Sartre, J.-P. (1972). *The psychology of imagination.* Oxford, UK: Blackwell.

Schacter, D. L. & Addis, D. R. (2007a). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362* (1481), 773-786. 10.1098/rstb.2007.2087

——— (2007b). Constructive memory: The ghosts of past and future. *Nature*, *445* (7123), 27-27. 10.1038/445027a

Schacter, D. L., Norman, K. A. & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49* (1), 289-318. 10.1146/annurev.psych.49.1.289

Schechtman, M. (1996). *The constitution of selves.* New York, NY: Cornell University Press.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, *15* (7), 319-326. 10.1016/j.tics.2011.05.006

Séglas, J. (1897). *Le délire des négations: séméiologie et diagnostic.* Paris, FR: Masson, Gauthier-Villars.

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

——— (2015). The cybernetic Bayesian brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

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

Shorvon, H. (1946). The depersonalization syndrome. *Proceedings of the Royal Society of Medicine*, *39* (12), 779-792.

Steinberg, M. (1995). *Handbook for the assessment of dissociation: A clinical guide.* Washington, DC: American Psychiatric Press.

Stock, K. (2007). Sartre, Wittgenstein and learning from imagination. In P. Goldie & E. Schellekens (Eds.) *Philosophy and conceptual art* (pp. 171-194). Oxford, UK: Oxford University Press.

Suddendorf, T. & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, *30* (3), 299-313. 10.1017/S0140525X07001975

Summerfield, J. J., Rao, A., Garside, N. & Nobre, A. C. (2011). Biasing perception by spatial long-term memory. *The Journal of Neuroscience*, *31* (42), 14952-14960. 10.1523/jneurosci.5541-10.2011

Szpunar, K. K., Watson, J. M. & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences*, *104* (2), 642-647. 10.1073/pnas.0610082104

Szpunar, K. K., Spreng, R. N. & Schacter, D. L. (2014). A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition. *Proceedings of the National Academy of Sciences*

Tulving, E. (1985). How many memory systems are there? *American Psychologist*, *40* (4), 385-398. 10.1037/0003-066X.40.4.385

——— (1995). Organization of memory: Quo vadis. *The Cognitive Neurosciences*, 839-847.

——— (2005). Episodic memory and autonoesis: Uniquely human. In H. S. Terrace & J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3-56). Oxford, UK: Oxford University Press.

Westbury, C. & Dennett, D. C. (2000). Mining the past to construct the future: Memory and belief as forms of knowledge. In D. L. Schacter & E. Scarry (Eds.) *Memory, brain, and belief* (pp. 11-32). Cambridge, MA: Harvard University Press.

Wilson, A. & Ross, M. (2003). The identity function of autobiographical memory: Time is on our side. *Memory*, *11* (2), 137-149. 10.1080/741938210

Wittgenstein, L. (1980). *Remarks on the philosophy of psychology.* Oxford, UK: Blackwell.

Young, A. W. & De Pauw, K. W. (2002). One stage is not enough. *Philosophy, Psychiatry, & Psychology*, *9* (1), 55-59. 10.1353/ppp.2003.0019

Metamisery and Bodily Inexistence

A Reply to Ying-Tung Lin

Philip Gerrans

The difference between the Cotard Depersonalisation and Depersonalisation Disorder may consist, not only in the fact that the Cotard delusion is a response to prediction error affective/bodily information, but the level in the predictive processing hierarchy at which predictions about bodily information are violated.

Keywords
Cotard delusion | Depersonalisation disorder | Interoception | Predictive coding | Self awareness

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

1 Prediction error and veridicality

My explanation of Depersonalisation Disorder (DPD) argued that the characteristic experience is shared by people who suffer from the Cotard Delusion (CD). The difference between the two conditions is that the person with DPD does not develop a delusional response to her experience of de-affectualisation. She simply reports is as it is: "I *feel as if* my experiences do not belong to me". The person with Cotard, however develops an explanation of that feeling and identifies with it "I no longer exist". In commenting on this proposal Ying-Tung Lin opens up a range of new possibilities for cognitive theorizing. The first is that the predictive coding approach provides a new framework for cognitive theorizing which improves on "second factor" approaches to delusion. The second is that attention to the predictive nature of the processes which generate experience might suggest an important difference between the two conditions: namely the role of the Anterior Insular Cortex (AIC).

One way to approach the phenomenon would be to ask why the person with DPD seems to be able to understand that her experience is not veridical while the person with CD

does not (*modulo* all the *caveats* about the epistemic status of delusional attitudes). The CD patient for example does not say "It feels as if I don't exist" she says "I don't exist". This way of approaching the problem fits with a now standard approach to delusion, that argues that there are (at least) two stages of cognitive processing involved in delusion formation. The first generates an anomalous experience and the second generates a delusional response to that experience.

Ying-Tung Lin however, following Hohwy and Clark, explains delusion in terms of the attempt by higher order control systems to account for surprisal in a predictive coding hierarchy. The radical aspect of these ideas is that neither the precipitating experience nor the delusional response need be conceived of as the result of cognitive malfunction. Because there is no intrinsic connection between error minimization and malfunction "certain misrepresentations can lead to error minimization; furthermore, it is possible for misrepresentation rather than veridical representation to lead to a generative model "(Lin this collection, p. 8).

Ying-Tung Lin's commentary applies these ideas to the Cotard delusion, arguing that it is a model that mimimises the prediction error represented by depersonalisation experience. Her target is to describe a

> cognitive architecture [that] could, in principle, explain CD in terms of its development from depersonalization, and what exactly are the underlying differences between patients suffering from the Cotard delusion and those suffering from depersonalization disorder (DPD) but free from the Cotard delusion? (Lin this collection, p. 2)

2 The sense of presence

Before I make some comments, I want to highlight the original aspects of her account and show how it can explain how experience acquires a quality of "mineness" or "sense of presence", that is of belonging to a self. We can then use the predictive coding framework to explain how the sense of presence can go missing. Loss of the sense of presence signals a prediction error which then requires a higher-level system to build a predictive model that fits that error.

The first point to note is that on the most radical interpretation of predictive coding ideas the veridicality of representation is a corollary of cognition not its primary goal. The primary goal of a cognitive system is to predict its own informational states consequent on its actions (broadly construed to include internal regulatory actions). The point is not just that the objects of experience are constructed and hence may be illusory or misrepresented. Rather veridicality of experience is secondary to the accuracy with which cognitive process predicts the flow of information in sensory systems. As she says in the case of perception this means that "instead of aiming to answer the question 'what is this?' perception studies should answer the question '... what does this resemble?' "(Lin this collection, p. 6). This formulation captures the idea that the visual system, for example, is not passively registering retinal information and constructing a representation of the external world, but using a model which predicts the flow of information coming from the retina.

The first step is to apply the same idea to interoception. We see that the mind is not passively registering changes in body state and constructing a model of the body accordingly but predicting the flow of bodily information in cognitive context. Those contexts range from maintenance of homeostasis to the use of affective experience to inform decision-making and reflective cognition. Thus when I think about the past or future these episodes of retrospection or prospection are infused with affective significance.

The radical import for the understanding of pathologies of self-representation is very elegantly brought out by her discussion. Ying-Tung Lin in effect argues that the experience of the self in autobiographical episodes is no more direct than experience of the world in perception or of past events in memory. In each case no object is directly represented or experienced. Rather the relevant object in each case (object

of perception, remembered event, or self in the case of first person awareness) is *inferred* as a part of a process of optimizing predictive accuracy in specific cognitive contexts.

As many have argued the role of the Anterior Insular Cortex (AIC) is to integrate and represent affective information: i.e., those bodily states, which tell the organism how it is faring in the world, actual, imagine or remembered. The point to recall from Ying-Tung Lin's account is that the AIC is not representing a self but constructing and optimizing a model that predicts the flow of affectively-charged bodily information.

This is why when AIC is hypoactive the subject feels a loss of subjective presence, reported as depersonalization. In particular the patient has a loss of subjective presence for her own body: she registers changes in body state but they do not feel affectively significant for her. Because that lack of feeling is not predicted she then reports it in the vocabulary of DPD.

Why does the DPD patient not proceed to something like the Cotard delusion? According to Ying-Tung Lin whether a delusion is formed depends on the degree of precision assigned to the information produced by hypoactivity in the AIC.

In the case of Cotard delusion developed from depersonalization, when one has the expectation of high precision, the system tends to be driven by the bottom-up predictive error of unexpected hypoactivity of the AIC, rather than the prior model. One is, therefore, more likely to revise the model in order to explain away the surprisal resulting from the mismatch between the actual and predicted activation level of the AIC; that is, the systems of patients suffering from CD are driven by an urge to modify their top-down predictive models in order to con- form to the loss of AIC activity. The construction of the model in CD is considered an attempt to minimize prediction error.

3 Conclusion

Reading over this account I wonder if there is an alternative interpretation available consistent with the predictive coding account. It is consistent with the view that patterns of activity in the AIC are abnormal in CD, but unlike DPD those patterns are not the result of VLPFC-induced hypoactivity.

Ex hypothesi the CD patient is extremely depressed. Evidence suggests that circuitry centred on the amygdala is affected, which means that online affective responses are flattened.

The role of the AIC is to monitor for changes driven by affective processing. It thus predicts for example that a typically positive event would be processed as positive. Thus, when that event is processed as negative or neutral, the AIC detects an error, signaled in the form of an anomalous experience. The patient is in the position being able to detect and signal changes in her affective responses, which take the form of unpredicted absences in bodily response. Thus *her lack of felt bodily response is processed as affectively significant* in the Cotard delusion with the result that she experiences it. Thus she does not feel neutral she feels miserable. Or as we might put it *she feels metamisery* because the role the AIC is to enable the person to feel the affective significance of bodily changes *including the absence of predicted changes*. In Cotard delusion the patient feels the affective significance the unpredicted absence of positive changes.

In DPD, by contrast, the patient does not feel the significance of bodily information because her AIC is inhibited and hypoactive.

Thus the difference between the two conditions may consist, not only in the fact that the Cotard delusion is a response to lower level prediction error, but the level in the predictive processing hierarchy at which predictions about bodily information are violated.

References

Lin, Y.-T. (2015). Memory for prediction error minimization: From depersonalization to the delusion of non-existence—A Commentary on Philip Gerrans. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.