Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-018-0415-2

Depersonalization Disorder, Affective Processing and Predictive Coding

Philip Gerrans[1]

Published online: 05 September 2018

Abstract
A flood of new multidisciplinary work on the causes of depersonalization disorder (DPD) provides a new way to think about the feeling that experiences "belong" to the self. In this paper I argue that this feeling, baptized "mineness"(Billon 2013, 2018a, b) or "subjective presence" (Seth, *Trends in Cognitive Sciences* 17(11): 565–573, 2013) emerges from a multilevel interaction between emotional, affective and cognitive processing. The "self" to which experience is attributed is a predictive model made by the mind to explain the modulation of affect as the organism progresses through the world. When the world no longer produces predicted affect, the organism needs to explain or explain this unpredicted absence of feeling away. It is important to this account that cognition and perception are otherwise intact. Consequently the mind's representation of the world and its emotionally salient properties are unchanged, leading the mind to predict a characteristic affective response. When that prediction is not fulfilled the organisms feels as if she is no longer present in experience. This is reported at the as feeling of depersonalization.

1 Pathologies of Self-Awareness

A flood of new multidisciplinary work on the causes of depersonalization disorder (DPD) provides a new way to think about the feeling that experiences "belong" to the self. In this paper I argue that this feeling, baptized "mineness"(Billon 2013, 2018a, b) or "subjective presence" (Seth 2013) emerges from a multilevel interaction between emotional, affective[1] and cognitive processing. The "self" to which experience is attributed is a predictive model made by the mind to explain the modulation of affect

[1]I am distinguishing affective from emotional processing on the basis that affects are felt. Affective processes are involved in the generation of such feelings. Emotional processes coordinate the components of an emotional episode (including feelings) by representing the goal-relevance of information.

✉ Philip Gerrans
Philip.gerrans@adelaide.edu.au

[1] Department of Philosophy, University of Adelaide, Adelaide, SA 5000, Australia

as the organism progresses through the world. When the world no longer produces predicted affect, the organism needs to explain the consequent unexpected feeling. It is important to this account that cognition and perception are otherwise intact. Consequently the mind's representation of the world and its emotionally salient properties are unchanged, leading the mind to predict a characteristic affective response. When that prediction is not fulfilled the organisms feels as if she is no longer present in experience. This is reported at the as feeling of depersonalization.

This account has three elements. The first is a conception of the nature of affective experience, which draws from and integrates elements of different theories of emotion: neoJamesian bodily theories, core affect theories (and their close cousins conceptual act theories) and the appraisal theory. Emotional processes, known as appraisals, detect and represent this relevance and coordinate a suite of appropriate responses including autonomic physiology, action tendencies, behavior and, crucially, the affective feeling characteristic of an emotional episode. Affect is a form of bodily feeling which informs the subject about the relevance of information to her goals and interests determined by appraisal processes.

The second element derives form the first: it is crucial to this account that there be some mechanism which allows representations of body state (e.g. raised blood pressure) to *also* carry information about the emotional salience of situations which evoked them. This mechanism is the encoding of bodily information in the anterior insula cortex.

The third is drawn from the predictive coding approach to cognition. The mind infers the existence of entities as part of a predictive coding strategy of "feature binding". In the case of perceived objects attributing perceptible properties to stable underlying objects makes the world predictable and tractable. As a result we experience a world as a collection of objects not features. Mental representations of objects are predictive models made by the mind as part of a general strategy of inferring hidden causes.

The experience of self-awareness or subjective presence is the result of the same process applied to affective experience. The fluctuations of affect evoked by encounters with the world as appraised by emotional processes are attributed to changes in an underling entity, the self. The self is a model inferred by the human mind to predict and explain the way the world makes the organism feel. The term inference as used in predictive coding refers to the construction and application of models at all levels of cognition, not just explicit propositional inference. So in this generous sense of inference perceptual systems perform inferences. When they misfire the resultant aberrant experience becomes the object of reflective cognition. For example there are rare disorders or illusions of feature binding for objects in which features, trajectories and boundaries are not integrated into a coherent object, but such failures cannot be resolved by reflective cognition because they are produced by perceptual and quasi-perceptual systems. Just as feature binding for objects is inaccessible to reflective cognition so are the inferential processes that attribute the flow of affective experience to an enduring entity. When they misfire, however the resultant experience becomes a focus of higher-level cognition. Just as we are unaware in normal cognition that our perceptual experience of objects is the result of a modeling process, so we are unaware that our feelings of being the subject of experience are the result of a similar process of "affective feature binding".

DPD arises when the world unpredictably and intractably ceases to evoke affective processes, even though it appears in every other respect unchanged. The mind makes the inference that the entity that sustains affect, the self, is no longer present. It is important to the account that loss of affect alone is insufficient to produce loss of a sense of self. It is loss of predicted affect. For example a new mother may predict (at all levels of cognition from implicit to explicit) that her baby will evoke positive affect. If however she suffers from post natal expression the smiles and hugs of her baby and the support and pleasure of family leave her feeling numb or despairing. She no longer feels present in experience.

I do not claim that the theory presented here is a complete theory of subjectivity or captures every form of self-awareness. For example a feeling of being present in experience is produced by being embodied, by having a perspectival orientation on experience and in the experience of agency. What is striking about DPD, however, is that it arises in subjects in whom these other forms of subjectivity are intact. Yet they still report feeling detached from their experience as if it is not happening to them. So my aim here is to explain how that aspect of subjectivity or mineness of experience is normally produced and is lost in DPD.

An advantage of this account is that it explains the neural correlates of DPD in terms of a larger theoretical picture that integrates the explanation of DPD with the role of those neural correlates in self-awareness, affective feeling and bodily awareness. The core neural correlate is hypoactivity in the anterior insula cortex (AIC), which has led to many proposals that the AIC is a neural correlate of self-awareness. These proposals however do not really explain why and how it is that the AIC should play this role. I propose that it does so in virtue of its role at the intersection of emotional processing, bodily representation and the generation of affective experience. This role allows it to be an essential component of the predictive self-model.

The structure of the paper is as follows: I briefly introduce DPD and Alexandre Billon's argument that "mineness" cannot be straightforwardly explained in terms of a psychological process. This is argument in effect shows that all other aspects of cognition except affective responses are intact in DPD. I explain the relationship between emotional processing and affective experience. I show how the AIC is specialized for sensing not just body state per se but the emotional significance of body state. I also present evidence that the AIC functions according to the predictive principle: modulating its activity according to higher-level predictive models. Pain asymbolia is an example. Colin Klein has described this experience as a form of DPD and thus it serves as an introduction to DPD proper, in which the subject feels detached, not only from particular episodes of experience, such as pain, but globally. I conclude that convergent evidence about the cognitive and neural correlates of DPD suggests that mineness results from the predictable modulation of affective experience signaled by activity in the AIC, attributed to an underlying entity: the self.

I then consider two objections to this account: one suggesting that activity in the AIC is neither necessary nor sufficient for the experience of mineness. The other suggests that mineness cannot be essentially an affective phenomenon because there are cases of depersonalization with intact affect.

2 Depersonalisation Disorder and Subjective Presence

The American Psychiatric Association describes Depersonalisation Disorder (DPD) this way: "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, DSM 5 2013. my italics). An example

> I feel some degree of 'out of it' all the time (...) I can sit looking at my foot or my hand and not feel like they are mine. This can happen when I am writing, my hand is just writing, but I'm not telling it to. It almost feels like I have died, but no one has thought to tell me. So, I'm left living in a shell that I don't recognize any more (Sierra 2008, 27).

Psychiatrists who first described DPD called the type of experience lost in DPD "personalization" or "self awareness". Seth et al. (2012, Seth 2013) uses the term "subjective presence" to refer to the feeling that experiences belong to oneself. Billon (2018a, b) has coined the useful term "mineness" to capture the same phenomenon.

Billon has given an exhaustive psychological and phenomenological account of mineness. He argues that mineness is not a matter of perception, interoception, emotion or cognition. He also argues that mineness is not an epistemic phenomenon, a matter of knowing or believing that one's experiences are one's own. His argument against psychological and epistemic accounts of mineness is that DPD does not essentially involve psychological or epistemic deficits. People with DPD can perceive and cognize their world, including their bodies, and know that they are having experiences. Nonetheless they do not feel as if they are the person having those experiences.

Billon's account does raise a puzzle, If cognition and perception including interoception are basically intact in DPD what in what does the sense of "mineness" consist? What neurocognitive process sustains it? And how does it relate to its neural correlate, hypoactivity in the AIC?

The first step in resolving this puzzle is to note that while cognition is intact, affective responses are flattened or absent. This suggests that one area to explore is the relationship between affective processing, feelings of mineness and activity in the AIC. As noted above there is a subtle relationship between these phenomena produced by the predictive architecture of emotional processing.

Anil Seth proposed that the association of affective processes with ongoing experience of the world (including the body) creates a feeling of "sense of presence", the feeling of being the subject of experience. For the rest of the paper I will use the terms subjective presence, mineness and self-awareness interchangeably to refer to the aspect of experience lost in DPD. As a way of explaining subjective presence Seth contrasts it with what he calls a sense of objective presence, the feeling that the world is "real". The phenomenon of derealisation, a feeling that the world is unreal, arises when this sense of objective presence is disturbed.

The current explanation of derealisation is that it a phenomenon of flattened or absent affect (Medford 2012). When experience becomes drained of affect, the world, or aspects of it, feel strange and unreal precisely because the mind "expects" the world to be affectively salient. Derealisation experience can be local, particular and episodic

(for objects, places or people, lasting moments, hours or days). For example we are all familiar with non-pathological cases of fleeting *jamais vu,* which are momentary episodes of derealisation for a familiar place, and with fleeting episodes of misrecognition in which familiar people feel like strangers (Christodoulou 1986; Brighetti et al. 2007). Pathological derealisation results from a global and sustained loss of predicted affective response to the world (Medford 2012).

Seth calls derealisation a loss of "objective presence" because the predicted affective response is to the external world. The concept of objective presence has a subjective correlate: subjective presence. Subjective presence is a sense of presence for one's own body.

On this way of looking at things there are two classes of *representanda*, external to the body and internal to the body. Exteroception tells us about external objects. Interoception carries information about internal states of our bodies such as levels of autonomic arousal, heart rate, blood pressure and so on.

This explanation is elegant, symmetrical (internal-external, exteroception-interoception, objective presence- subjective presence) and uses resources that have identifiable neural correlates implicated in relevant pathologies. We do have exteroceptive, interoceptive and affective capacities and their interactions are implicated in pathologies of derealisation and depersonalisation.

A question Seth's account raises is why patients who lose a sense of subjective presence do not just report derealisation for their own body, or those aspects of bodily functioning which have ceased to evoke an affective response. Why not just say "my body feels strange and unreal" rather than "I feel as though I am not present", or "I feel as though things are not happening to me"? In fact reports of derealisation for the body comprise part of the spectrum of symptoms of depersonalization disorder but derealisation for the body and depersonalization are not identical (Baker et al. 2003, Simeon, Kozin et al. 2008, Sierra and David 2011).

The answer to this question requires an account of the relationship between emotional processing, affective and bodily feelings.

3 Emotion and Affect

The experience of emotion, affect and bodily feeling are very closely related, so much so that some theories of emotion identity them. For example some theoretical descendants of William James identify the emotional state of fear with the bodily feeling evoked by perceiving a dangerous stimulus (Damasio 2006; Prinz 2004). For such views the essence of emotion is bodily feeling. There are two problems for such views and theories of emotion divide in their approach to these problems.

The first is that bodily feelings and the feelings associated with an emotional episode are not identical. One can have one without the other. For example feelings of nausea and lethargy can arise as the result of an illness or in an episode of grief. Emotional feelings are a subclass of bodily state evoked by an emotionally salient event. Call such emotional feelings affects. A question for emotional theory is how to distinguish affect from bodily feeling per se, for example grief from the feelings evoked by some of illnesses or the arousal of anger from that produced by adrenalin.

An answer is that affects are distinguished by the fact that they carry emotional information, the feeling of fear informs me of danger, anger or insult, of grief irrevocable loss. Each emotion has a characteristic "core relational theme" which determines which information is relevant (the CRT of fear for example is danger). But this raises a further question for pure bodily theories. How does bodily feeling acquire this thematic content?

Here there are two common answers. One is Darwinian. It says that the feeling of fear represents danger in virtue of its evolutionary history. We are in Prinz's words "set up to be set off" by dangers. So he calls his theory an "embodied appraisal" theory, where the concept of appraisal refers to the evaluation of information as falling under a CRT.

An alternative view locates content downstream from the affective feeling. Such "core affect" views note that at any given moment we experience an integrated overall readout of the our body state which informs us how we are faring in the world. However for core affect theses the content of such body states is limited to degrees of arousal (high or low) and valence (feelings of aversion or attraction). More specific emotional content depends on the concepts we use to interpret affective states, and those concepts are acquired in developmental and cultural contexts which in effect teach us how to appraise/evaluate the relationship between feelings and eliciting situations (Russell and Feldman Barrett 2009) .

A third view is that affective responses are the result of appraisals. It is because we evaluate a situation as dangerous that we feel fear, as an irrevocable loss that we feel sadness. It is no part of appraisal theory that all such appraisals must be conscious and propositional. Sometimes they are an almost instantaneous consequence of perception. In other cases they can be the product of a long chain of reflective thought (Sander et al. 2005, Grandjean, Sander et al. 2008, Brosch et al. 2010). The main features of appraisal theory I emphasise are that it is

> *Hierarchical*. Appraisal takes place at different levels of cognitive processing from quasi perceptual to reflective and conceptual. The lowest levels are coordinated by the amygdala (now conceptualized as a "relevance detector"). At higher levels of explicit cognition the coordinating role of the amygdala is recapitulated by the ventromedial prefrontal cortex.
> *Integrative*. Emotional processes work by coordinating other cognitive systems to the detection and evaluation of relevant information. They work by biasing cognition to the processing of information relevant to CRTs
> *Componential*. Emotions have components: cognitive, bodily, behavioural and affective. The role of appraisal is to coordinate these components adaptively in an episode of emotion. For example to bias perception and cognition to represent the nature of a loss, to produce the bodily responses and action tendencies and characteristic feelings of sadness once a loss is determined as irrevocable.

What we feel at a particular moment is the emotional significance of bodily state produced by the context-sensitive appraisal that commandeers most processing resources. On this view affective states are the result of appraisals that take place at different levels of cognition. For example sadness can be the result of implicit quasi-perceptual processes, or of sustained explicit reflection.

The relationship between affective processing and self-modeling also hierarchical. At low levels of implicit processing the self-model is implicit, entailed by the fact that the goals and needs which create CRTs belong to the organism. It is *my* survival needs that determine the nature and extent of perceived dangers. This is why the experience of sadness evoked perceptually is automatically and implicitly self attributed. The resultant feelings then inform higher levels of cognition at which the processes of appraisal are recapitulated at an explicit level. We may for example reflect on a loss in the overall context of our life. Here the attribution of affect to a self is explicit involving the so called "narrative I" (Schechtman 2011). But it is precisely the fact that feelings, which hover at the border of implicit and explicit cognition are part of the process of reflection which gives it its personal dimension.

4 Mechanisms of Mineness

The ability to feel the significance of such appraisals depends on activity in the anterior insula cortex (AIC). Sections of the posterior insula cortex (PIC) map body state directly and integrate those representations to co-ordinate basic regulatory functions. The PIC, for example, represents things like blood pressure, nociception (bodily damage), pressure and departures from homeostasis and integrates that information to enable reflexive regulatory processing (Bennett and Baird 2009, Craig 2009a, b, Craig 2010). Thus the PIC essentially plays a role in basic bodily monitoring and regulation. The PIC allows us to be aware of body state per se, as a state of the world.

The AIC is specialized to re-represent and integrate information about body state to allow us to feel significance of body states as affects. Critchley summarises the idea that the AIC allows us to feel emotional significance.

> 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 (Critchley 2005) (Critchley 2005).

The role of affect is to enable is to consciously experience the nature of an emotion rather than for example simply perceiving or believing that a family member has died. The feeling makes the information particularly salient focusing attention and higher "The anterior insula ensures that salient stimuli such as painful stimuli will have preferential access to the brain's attentional and working memory resources" (Wiech and Tracey 2013).

Thus we can say that the AIC represents, not just body state, but the emotional significance of body state in context. As Craig puts it "activity in the AIC "provides the basis for the continuity of subjective emotional awareness" (Craig 2009b).

The account so far suggests that activity in the AIC allows us to feel the emotional salience of information in the form of affect. To extend the account to explain mineness and depersonalization we need to add the idea that the AIC functions according to a predictive coding principle.

Predictive coding is a familiar concept in the cognitive science of perception and motor control and, more recently, interoception (Hohwy, Roepstorff et al. 2008, Seth et al. 2011, Clark 2013, Hohwy 2013, Seth 2013, Barrett and Simmons 2015). It can also explain integrative cognitive processing such as that which links affective to cognitive states.

Predictive coding is an hierarchical process in which higher-level predictive models generate predictions and resolve error signals. The flow of information up and down the hierarchy is driven by the degree of fit between predictive models and information produced by action. On this view conscious reflection and deliberation is a form of active inference, a way of sampling outcomes of simulated actions to deal with discrepancy signaled by lower level systems. Exploratory behavior, cognitive search for predictive models and mental simulation are ways of sampling outcomes until the optimal fit between action and model is achieved. This type of search and sampling is called "active inference" Adams et al. 2013; Howhy 2102).

In the case of interoception this means that bodily regulation employs predictive models against which fluctuations in body state are evaluated. Importantly these models include mot merely models of bodily state per se but models of the emotional significance of body state.

By this I mean that how the body interprets a rise in blood pressure or nociceptive input depends on whether that input is predicted, not merely cognitively and perceptually but affectively. An increase in arousal can be fitted to a predictive model of bodily and/or emotional functioning in context. Whether it is felt as a purely physiological response, as fearful or exciting depends on whether and how the situation is appraised at different levels in the cognitive hierarchy. Consequently we update predictive models: how the AIC "interprets" what is going on in the body depends on prior models of the bodily/affective consequences of action.

Recent work in affective neuroscience treats interoceptive processing in the insula within the predictive framework (Paulus and Stein 2010, Garfinkel and Critchley 2013). Feldman Barrett and Simmons, for example, argue that interoception, like other cognitive functions, is not simply signal detection but is rather a matter of using a model of interoceptive activity to predict, interpret and explain signals about changing body state.

> Rather than interoceptive perceptions being solely the representation of afferent sensory input from the body, they can be thought of as inferences about the sensory consequences of homeostatic budgeting that are implemented as upcoming visceromotor commands; these inferences are constrained by error signals that result from the failure of previous predictions to accurately account for incoming interoceptive sensations (Barrett and Simmons 2015)

The predictive principle applies, not only to the production of affect but to its interpretation at the level of conscious reflection. Predictive coding theories of cognitive function treat the objects of perception as inferred constructs that make the flow of information predictable and intelligible. This framework provides an excellent way to explain why mineness is interpreted as self-awareness. We can treat the self as an inferred entity that explains the flow of subjective presence.

The constant modulation of affect produces changes in the quality of feeling but, as long as feeling state is being modulated, subjective presence is constant, providing the subject with a sense of how she is faring in the world. The idea that this flow represents changes to a persisting entity, a subject, is a natural "inferred representation of the self" (Moutoussis et al. 2014). "agents model the self as a hierarchy of hidden, endogenous causes, and further…the self is identical to these causes" (Hohwy and Michael 2017, in press).

So the idea is that as the world evokes affective states in us the mind builds a model of the unobservable entity which sustains those states, attributing them to a persisting object, the self.

Seth puts it this way:

> …emotion and embodied selfhood are grounded in active inference of those signals most likely to be 'me' across interoceptive and exteroceptive domain (Seth 2013).

The self model is distributed across multiple levels of hierarchical processing from implicit to explicit. As Hohwy and Michael put it "agents model the self as a hierarchy of hidden, endogenous causes, and further, that the self is identical to these causes" Hohwy and Michael in press.

Hohwy and Michael have a more expansive version of self modeling in mind than the one defended here: their model includes all aspect of subjectivity (agency, forms of embodiment, perception) where as I am only concerned with affective self modeling. The reason is that I am concerned to explain the loss of the sense of mineness in DPD.

5 Pain Asymbolia and Placebo Analgesia

The relationship between emotional processing, body state, affect and the sense of mineness is illustrated by the functioning of the so–called neurocognitive "pain matrix" in pain asymbolia. The experience of pain does not reduce to nociception (the representation of bodily damage). Normally we experience a combination of signals representing bodily damage, the significance of that bodily damage and the bodily, behavioural and regulatory responses (Garland 2012; Wiech and Tracey 2013)). These elements can dissociate. Pain asymbolics feel the pain, but feel detached from it as thought it is not theirs.

Their experience resembles reports of pain experience under opioid analgesia, in which patients report that the pain is not extinguished but "no longer matters". A key finding here is that that opioids target, not only the PIC, as one might expect, but the AIC and related limbic structures involved in emotional processing. The AIC in fact is even more responsive than PIC to low doses of opioids. This is presumably an adapation. It is easier for an organism to regulate initial emotional response to bodily damage than to repair bodily damage. Thus in contexts where the organism cannot devote resources to repair it inhibits the system that produces negative affect in response to bodily damage and thereby stops pain from drawing attention away from other relevant activities. Opioids exploit or mimic this adaptation, down regulating the

AIC, reducing the felt significance of pain. "the FMRI data suggest that opioid analgesics can directly influence emotional responses at low doses that do not alter sensory aspects of pain" (Lee, Wanigasekera et al. 2014).

Pain asymbolics no longer assign emotional significance to bodily damage in virtue of hypoactivity in their AIC. For this reason pain asymbolia has been compared to DPD.

> If this is right, the phenomenology of asymbolia might resemble a kind of depersonalization syndrome. … The asymbolic, and the depersonalized more generally, feel sensations that they are estranged from — that they do not take to be theirs in the sense that we normally do. … [This]does show that there is another sense in which our sensations may be unified: as sensations over which we have a feeling of ownership. Asymbolia, and depersonalization more generally, shows that this sort of unity may fail. Its failure comes not from a change in the sensations we feel, but in the *sort of agents* we are. (Klein 2015) my italics.

Another way to put this is to say that pain asymbolics have lost the sense of mineness for pain.[2]

Klein connects these findings to his thesis about the nature of pain asymbolia (pain has lost its characteristic "imperative" quality) in terms of compromised self-representation ("the sort of agents we are"). This allows him to draw a connection between DPD and pain asymbolia. In pain asymbolia patients experience bodily damage, but do not experience that damage as significant to them. But patients with asymbolia do not have DPD, even if DPD patients are sometimes indifferent to pain. What this shows is that mineness can be lost locally, for aspects of bodily functioning, such as pain.

Depersonalisation arises when the loss of mineness is more global not restricted to a particular class of experience such as pain.

6 Depersonalisation Disorder

As one might expect from this account, a crucial neural correlate of DPD is *hypoactivity* in the AIC which produces a lack of predicted sense of presence. A current hypothesis is that this hypoactivity is produced by involuntary inhibitory activity in the ventrolateral prefrontal cortex (Baker et al. 2003). The vlpfc is a structure that plays a crucial role in the downregulation of affective feeling, evidenced in studies of voluntary reappraisal (Ochsner and Gross 2005, Kalisch 2009, Füstös et al. 2013). 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

[2] This idea is inconsistent with some conceptual theses about the nature of pain (that it is a form of immediate unstructured self awareness) that suggest that it is impossible to have pain without motivation or ownership. But apparently simple experiences have complex internal structure that can decompose under the conditions characteristic of certain pathologies.

result of involuntary deactivation of systems that produce the experience of mineness. It is not absence of mineness that generates depersonalization, but that predicted mineness is absent in virtue of the involuntary inhibition of the AIC by the vlpfc. When people engage in voluntary or effortful inhibition of affect while reappraising they do not feel depersonalized because the mind predicts the change in affective feeling as a result of top down signals.

The absence of predicted feeling in DPD could arise in different ways and at different levels. One possibility might be that appraisal processes malfunction as in mood disorders. For example a mother with post-natal depression might predict (including tacit predictions at low levels in the cognitive hierarchy) positive affective reactions to the new baby, and also be surrounded by social expectations that reinforce that prediction at the level of explicit thought. But due to her depression holding the baby produces none of the bodily/affective changes characteristic of positive appraisal. The AIC then has no positive signals to integrate with the mother's reflections on her situation. If her mind is predicting that the AIC should be giving a positive readout for interaction with the baby but she feels nothing then she may well develop feelings of derealisation for the baby, ultimately even depersonalization.

Another possibility, which seems confirmed by some studies of DPD (Sedeño, Couto et al. 2014), is that the PIC is disrupted in its early integrative and synthesizing role, and thus relays distorted or suppressed signals about body state to the AIC. In such a case the AIC's metarepresentation of body state would inherit the first order misrepresentations or lack of representations in the PIC.

Another possibility (see the case described by Schilder below) is that reflexive appraisals are intact, leading to normal autonomic response integrated and detected by the PIC, but downstream problems with the AIC mean that these early affective responses are not in fact treated at as subjectively relevant at higher levels of appraisal. The subject sees an angry face recognizes it as threatening, gets the predicted bodily response, but does not feel that bodily response affectively. As Michal et al. put it:

> Thus, their cognitive evaluation seems to be disconnected from their bodily or autonomic responses, respectively. This may be in line with the observation, that DPD patients have greater difficulties in identifying own feelings as compared to other patients or healthy controls [57]. It also reminds of findings of hypoactive insula in DPD, as the insula is involved in the *conscious representation of autonomic states of the body* [13,17,58] (Michal et al. 2014).

It might seem counter intuitive that the bodily and behavioural components of reflexively-generated emotional episodes could dissociate from affective feeling, but as with the dissociation between affective feeling and pain, the split between experiencing bodily state and affective feeling can be adaptive. It allows us to continue to cognize, appraise and respond in normal bodily and behavioral fashion while inhibiting the consequent affective feelings. For example avoidance behavior and bodily arousal might continue in a threatening situation but it might be adaptive to down regulate affective feeling while planning alternative responses. Surgeons, trauma counsellors and bomb disposal experts benefit from abilities to inhibit affective feeling while appraisal machinery continues to function. Such cases are not experienced as

depersonalisation because affective down regulation is controlled or acquired through habituation. Indeed as noted above opoids exploit this adaptation, down regulating affective response and self attribution of pain.

If DPD is a pathological case of the independent regulation of bodily and affective experience then it is less surprising that Michal and collaborators found cases in which "[there] was no correlation of the severity of "anomalous body experiences" and depersonalization with measures of interoceptive accuracy." (Michal et al. 2014, p. 1). 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 et al. 2014, p. 1; my italics.)

7 Depersonalisation for Affective Responses?

I have argued that mineness is produced by the modulation of affect by cognitive context, and that depersonalization results from loss of the sense of mineness. So the account is essentially an affective one. Billon (2018a, b) has pointed to a problem for such accounts raised by DPD patients with apparently intact affective responses. He quotes the rich clinical summaries of people like Schilder (1928) quoted by Billon, who wrote

> The objective examination of such patients reveals not only an intact sensory apparatus, but also an intact emotional apparatus. All these patients exhibit natural affective reactions in their facial expressions, attitudes, etc.; so that it is impossible to assume that they are incapable of emotional response.

At face value this hardly seems consistent with the idea that patients with DPD have lost affective responses. Interestingly, however, in the same passage Schilder wrote: "The emotions likewise undergo marked alteration. Patients complain that that they are capable of *experiencing* neither pain or pleasure; love and hate have perished with them". (my italics).

The apparent inconsistency here is resolved by the componential nature of emotional processing. As we noted earlier, emotional episodes involve reflexive and reflective appraisal, bodily and regulatory responses, action tendencies, expressive behaviour and *feelings.* The explanation of DPD was that it results from the unpredicted loss of the felt/affective aspect. It is consistent with this account that other components remain intact. The DPD patient continues to recognise the emotional salience of situations: her basic appraisal mechanisms are intact (there is no suggestion that the appraisal hubs, the amygdala and vmpfc are impaired) and initiate the automatic bodily and behavioural responses. However due to hypoactivity in the AIC she does not feel the significance of those changes and it is the dissonance between absence of predicted feeling, cognition and action which makes her feel "as if" she is an automaton performing actions which are not really hers.

This is consonant with the description by a patient of Dugas who first described the disorder. ""I only feel anger from the outside, *by its physiological reactions*" (Dugas and Moutier 1911) my italics.

This is consistent with the idea that the cognitive and behavioural components of an emotional episode are intact while the felt aspect has gone missing. Indeed if all aspects disappeared at once the result would be a disorder of emotion: inability to recognize and respond to emotionally salient information.[3] It is the unpredicted dissociation between affective and other components that is distinctive of the disorder.

Billon himself goes close to endorsing this view when he says that the biological role of mineness is to "help us feel concern for what really matters". This is part of an explanation of DPD as a defensive form of dissociation from distressing experiences. Suppressing the sense of mineness is a way to make sure that distress is not annexed to oneself so that one no longer fully identifies "with the entity concerned by what these affective states represent".

On my view affect allows us to *feel* concern for what really matters (subjective relevance) and hence mineness reduces to predictable modulation of affect. On Billon's view "mineness" provides that feeling. His main reason for rejecting the affective account is the dissociation between affective feeling, cognition, bodily response and behavior. But this dissociation is explicable if these are dissociable components of emotional processing.

8 Roger's Version

Another problem for my account is raised by a patient with extensive damage to the insular as well as the amygdala, ventromedial prefrontal cortex and anterior cingulate cortex. Roger, has a range of cognitive and affective deficits including dense anterograde amnesia but, arguably, he does not have depersonalization.

Philippi et al. tested Roger on a range of standard tests of self awareness (SA) and found "R is a conscious, self-aware, and sentient human being despite the widespread destruction of cortical regions purported to play a critical role in SA, namely the insula, anterior cingulate cortex, and medial prefrontal cortex" (Philippi et al. 2012). The experimenters concluded that that Roger was self-aware in virtue of his ability to integrate very basic bodily signals arising in the brainstem with his declarative and semantic knowledge about his life and personality traits "we find little support for the hypotheses that implicate the insular cortex as critical to all aspects of SA".

Not only that but Roger's aversive response to painful stimuli was not only intact but amplified. Unlike the typical subject he does not seem to habituate to a series of painful stimuli but finds each occasion equally painful.

Thus it is very tempting to conclude that Roger has both intact affective responses and sense of self despite lacking the structures that I have claimed are (in the typical case) required to construct a sense of self from the flow of affective feeling.

An obvious response is to claim that Roger has compensated for his loss, and in fact this is one interpretation of Roger's symptoms

[3] Such disorders are characteristic of problems with appraisal

> his intact affective experience of pain is due to plasticity…the adaptive role of pain affect is so essential that the brain may automatically rewire in service of self-preservation…[suggesting that] … emotional experience of pain can be instantiated by brain structures outside of those traditionally presumed to be critical for pain affect (Feinstein et al. 2016)

Alternatively Roger's may compensate for his deficits by using alternative cognitive strategies. For example his intact ability to sense his heartbeat is very likely not dependent on interoception but via the skin surface of his chest. When this surface was anaesthetized Roger's ability declined, unlike that of controls who presumably exploited the intact interoceptive pathway (Philippi et al. 2012).

If this is the case then the basic structure of my account can be preserved: Roger has has reinstalled the basic cognitive architecture of mineness through a combination of neural plasticity, redeployment and adaptation. However the idea that one's whole emotional/cognitive/affective architecture could be seamlessly reinstalled seems unlikely. More likely is a complex combination of compensation and deficit.

And there is a more complicated interpretation that suggests that Roger in fact has some affective deficits that impair his sense of mineness.

This interpretation starts from the fact that Roger has exaggerated and unmodulated pain responses. He does not habituate to continued painful trials. In this respect he resembles some patients with ventromedial lesions who do not learn from negative experience (unsurprisingly since Roger has ventromedial lesions). His pain responses, behavioural and verbal, are like startle responses.

Recall one important role of the anterior insula is to make affective feelings available in the context of secondary appraisal, to allow us to determine significance of bodily states. Where appropriate we can then redirect or focus attention, inhibit or amplify the affective response. Roger has none of this equipment available: he just has nociception, consequent behavior and higher order cognition. My suggestion is that Roger is not in fact appraising his pain, even at low levels. He is simply experiencing the automatic transition from nociception to behavioral expression. Similarly he cannot use higher-level predictive models to interpret his nociceptive signals. Thus pain is painful for Roger but not emotionally contextualized.

This interpretation is in fact advanced by the experimenters

> data suggest that the limbic structures commonly associated with pain may play a fundamental role in pain regulation. Under this view, the missing regions in Roger's brain would impair his ability to control and downregulate his pain responses.

This suggests that in fact Roger may be missing precisely the affective capacity for which feeling is required. Regulation and control.

This lack of ability to appraise and sense the significance of appraisal should on my account be reflected in the loss of a sense of mineness. His condition is consistent with such a loss despite his intact semantic knowledge about himself.

Unfortunately both hubs of Roger's appraisal system are lesioned and he is densely amnesic as well as having "myopia for the future" in Antoine Becchara's nice phrase. This explains his constant happiness.

> Roger appears remarkably unconcerned by his condition. He hardly ever complains, and in general, shows little worry for anything in life. Both of his parents and his sister fervently claim that "Roger is always happy," an observation that is consistent with our own impression. Moreover, based on his family's report, Roger is paradoxically happier now than he was before his brain damage. (Feinstein et al. 2010)

Roger does not does not use a predictive model of his feeling states to navigate his world. When we focus on that aspect of self-awareness described as subjective presence it seems that Roger not only lacks it, but does not feel the loss of that lack. He has lived in the same unmodulated emotional world for 30 years with one default setting that does not depend on the normal functioning of an appraisal hierarchy.

9 Mineness without me?

The account here is neutral between substantialist or realist views of the self, which treat it as a genuine (neurocognitive) entity,[4] and anti-realist, phenomenalist views like that of Metzinger (2004, 2011, 2016) *.[5] Phenomenalists about the self argue that the experience of being a unified entity does not veridically represent an entity that causes the experience. A problem for the Metzinger* (family of) views is that we do in fact feel like selves and think of ourselves as selves, that is as persisting entities that sustain experiences. So anti-realist views need to explain the origin and persistence of this illusion.

The affective theory provides a solution to this problem consistent with Metzinger's* explanation of the persistence of the illusion. He argues that we need, for purposes of cognitive control and social interaction, not only to be able to represent our organismic trajectory from a first person perspective but to *feel as if* we are selves, consistent unified substrates of experience. However feeling *as if* you are a self is not the same thing as the experience of a being a self. Any more than a visual illusion is the experience of an actual object.

> It is *yourself* who will be punished or rewarded in the future, it is *yourself* who will either enjoy a good reputation in the future or be subjected to retaliation. What we need for that is an intact "narrative self-model", an *illusion of sameness*.

[4] See e.g. Craig's (2009a) explanation of the role of AIC "it generates an image of 'the material me' (or the sentient self) at one moment in time"

[5] the quotation is from an interview, his written work is far more complicated and subtle on the question, so Metzinger* refers to his interview persona).

> Then the "stabs of conscience" can make us even more self-conscious, integrating individual preferences with group preferences (Metzinger 2016)

The self plays the same, anchoring or indexing role as objects do for the binding of perceptual experience. But as Metzinger* puts it, in the case of the self: "There is just no entity there, no individual substance, and scientifically we can predict and explain everything we want to predict and explain in a much more parsimonious way."

The basis of that parsimonious explanation is that we attribute feelings of subjective presence to a continuing entity as a way to predict and interpret the constant flow of affective experience. We have to do that in order to make ourselves behave adaptively over time, especially in social contexts. If we could not recall or imagine the self-relevant aspect of emotional experience (that is to say that part of feeling which is not exhausted by representation of body state per se) social life would have no significance for us.

The course of our life is marked by affective fluctuations as we appraise and reappraise the world and our place in it. When those fluctuations occur in a predictable way the sense of presence reinforces the sense of a continuing self. When they are unpredicted or absent our sense of presence is comprised or disappears. We call this loss of a sense of self or self-awareness but perhaps what is lost is the illusion of self, constantly generated by the persistence of subjective presence.

References

Adams, R.A., S. Shipp, and K.J. Friston. 2013. Predictions not commands: active inference in the motor system. *Brain Structure and Function* 218 (3): 611–643.

American Psychiatric Association. 2013. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub.

Baker, D., E. Hunter, E. Lawrence, N. Medford, M. Patel, C. Senior, M. Sierra, M.V. Lambert, M.L. Phillips, and A.S. David. 2003. Depersonalisation disorder: Clinical features of 204 cases. *The British Journal of Psychiatry* 182 (5): 428–433.

Barrett, L.F., and W.K. Simmons. 2015. Interoceptive predictions in the brain. *Nature Reviews Neuroscience* 16 (7): 419–429.

Bennett, C.M., and A.A. Baird. 2009. The processing of internally-generated interoceptive sensation. *Neuroimage* 47 (1): S84.

Billon, Alexandre. 2013. Does consciousness entail subjectivity? The puzzle of thought insertion. *Philosophical Psychology* 26 (2): 291–314.

Billon, A. (2018a). Making sense of the Cotard syndrome. Insights from the study of depersonalization. Mind and Language.

Billon, A. (2018b). "What is it like to lack "mineness". Depersonalisation as a probe for the scope, nature and role of mineness."

Brighetti, G., P. Bonifacci, et al. 2007. 'Far from the heart far from the eye': Evidence from the Capgras delusion. *Cognitive Neuropsychiatry* 12 (3): 189–197.

Brosch, T., G. Pourtois, and D. Sander. 2010. The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion* 24 (3): 377–400.

Christodoulou, G.N. 1986. Role of depersonalization-derealization phenomena in the delusional misidentification syndromes. *Bibliotheca Psychiatrica* 164: 99–104.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36 (03): 181–204.

Craig, A. 2009a. How do you feel—Now? The anterior insula and human awareness. *Nature Reviews Neuroscience* 10 (1).

Craig, A.D. 2009b. Emotional moments across time: A possible neural basis for time perception in the anterior insula. *Philosophical Transactions: Biological Sciences* 364 (1525): 1933–1942.

Craig, A. 2010. The sentient self. *Brain Structure and Function*: 1–15.

Critchley, H.D. 2005. Neural mechanisms of autonomic, affective, and cognitive integration. *Journal of Comparative Neurology* 493 (1): 154–166.

Damasio, A. R. (2006). *Descartes' error*. Random House.

Dugas, L. and F. Moutier (1911). *La Dépersonnalisation*, Félix Alcan.

Dunn, B.D., T. Dalgleish, and A.D. Lawrence. 2006. The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioural Reviews* 30: 239–271.

Dunn, B.D., H.C. Galton, et al. 2010. Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological Science* 21 (12): 1835–1844.

Ellis, H.D., and M.B. Lewis. 2001. Capgras delusion: A window on face recognition. Trends in Cognitive Sciences 5 (4): 149–156.

Feinstein, J.S., D. Rudrauf, S.S. Khalsa, M.D. Cassell, J. Bruss, T.J. Grabowski, and D. Tranel. 2010. Bilateral limbic system destruction in man. *Journal of Clinical and Experimental Neuropsychology* 32 (1): 88–106.

Feinstein, J.S., S.S. Khalsa, T.V. Salomons, K.M. Prkachin, L.A. Frey-Law, J.E. Lee, et al. 2016. Preserved emotional awareness of pain in a patient with extensive bilateral damage to the insula, anterior cingulate, and amygdala. *Brain Structure and Function* 221 (3): 1499–1511.

Füstös, J.R., K. Gramann, et al. 2013. On the embodiment of emotion regulation: Interoceptive awareness facilitates reappraisal. *Social Cognitive and Affective Neuroscience* 8 (8): 911–917.

Garfinkel, S.N., and H.D. Critchley. 2013. Interoception, emotion and brain: New insights link internal physiology to social behaviour. Commentary on "anterior insular cortex mediates bodily sensibility and social anxiety" by Terasawa et al. (2012). *Social Cognitive and Affective Neuroscience* 8 (3): 231–234.

Garland, E.L. 2012. Pain processing in the human nervous system: A selective review of nociceptive and biobehavioral pathways. *Primary Care: Clinics in Office Practice* 39 (3): 561–571.

Gerrans, P. 2014. *The measure of madness: Philosophy, cognitive neuroscience, and delusional thought*. Cambridge: Mass, MIT Press.

Hohwy, J., A. Roepstorff, and K. Friston. 2008. Predictive coding explains binocular rivalry: An epistemological review. *Cognition* 108 (3): 687–701.

Hohwy, J. 2012. Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology* 3.

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

Hohwy, J., and J. Michael. 2017. Why should any body have a self? In *The Subject's matter: Self-consciousness and the body*, ed. F. Vignemont and A. Alsmith. Cambridge: MIT Press.

Kalisch, R. 2009. The functional neuroanatomy of reappraisal: Time matters. *Neuroscience & Biobehavioral Reviews* 33 (8): 1215–1226.

Kircher, T., and A. David, eds. 2003. *The self in neuroscience and psychiatry*. Cambridge: Cambridge University Press.

Klein, C. 2015. What Pain Asymbolia Really Shows. *Mind* 124 (494): 493–516.

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

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. MIT Press.

Metzinger, T. 2011. The no-self alternative. In *The Oxford handbook of the self*, ed. S. Gallagher, 279–296. Oxford: Oxford University Press.

Metzinger T. 2016. All about the ego tunnel. Interview 3ammagazine. http://www.3ammagazine.com/3am/all-about-the-ego-tunnel/

Michal, M., A. Koechel, et al. 2013. Depersonalization disorder: Disconnection of cognitive evaluation from autonomic responses to emotional stimuli. *PLoS One* 8 (9): e74331.

Michal, M., B. Reuchlein, et al. 2014. Striking discrepancy of anomalous body experiences with normal interoceptive accuracy in depersonalization-derealization disorder. *PLoS One* 9 (2): e89823.

Moutoussis, M., P. Fearon, et al. 2014. Bayesian inferences about the self (and others): A review. *Consciousness and Cognition* 25: 67–76.

Ochsner, K.N., and J.J. Gross. 2005. The cognitive control of emotion. *Trends in Cognitive Sciences 9* (5): 242–249.

Paulus, M.P., and M.B. Stein. 2010. Interoception in anxiety and depression. *Brain Structure and Function* 214 (5–6): 451–463.

Phillips, Mary L., et al. 2001. Depersonalization disorder: thinking without feeling. *Psychiatry Research: Neuroimaging* 108 (3): 145–160.

Philippi, C.L., J.S. Feinstein, S.S. Khalsa, A. Damasio, D. Tranel, G. Landini, et al. 2012. Preserved self-awareness following extensive bilateral brain damage to the insula, anterior cingulate, and medial prefrontal cortices. *PloS one* 7 (8): e38413.

Prinz, J. 2004. *Gut reactions: A perceptual theory of emotion*. New York: Oxford Universty Press.

Russell, J. A., and L. Feldman Barrett. (2009) Core affect. The Oxford companion to emotion and the affective sciences:104.

Sander, D., et al. 2005. A systems approach to appraisal mechamisms in emotion. *Neural Networks* 18: 317–352.

Sander, D., D. Grandjean, and K.R. Scherer. 2005. A systems approach to appraisal mechanisms in emotion. *Neural Networks* 18 (4): 317–352.

Schechtman, Marya. 2011. "*The narrative self*."

Scherer, K. 2004. Feelings integrate the central representation of appraisal-drive response organisation in emotion. In *Feelings and emotions: The Amsterdam symposium*, ed. A.S. Manstead, N. Frijda, and A. Fischer, 136–157. Cambridge: Cambridge University Press.

Sedeño, L., B. Couto, M. Melloni, A. Canales-Johnson, A. Yoris, S. Baez, et al. 2014. How do you feel when you can't feel your body? Interoception, functional connectivity and emotional processing in depersonalization-derealization disorder. *PloS one* 9 (6): e98769.

Seth, A.K. 2013. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences* 17 (11): 565–573.

Seth, A.K., K. Suzuki, et al. 2012. An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology* 2.

Sierra, M. 2008. Depersonalization disorder: Pharmacological approaches. *Expert Review of Neurotherapeutics* 8 (1): 19–26.

Sierra, M., and A.S. David. 2011. Depersonalization: a selective impairment of self-awareness. *Consciousness and Cognition* 20 (1): 99–108.

Simeon, D., D.S. Kozin, K. Segal, B. Lerch, R. Dujour, and T. Giesbrecht. 2008. De-constructing depersonalization: further evidence for symptom clusters. *Psychiatry Research* 157 (1-3): 303–306.

Wiech, K., and I. Tracey. 2013. Pain, decisions, and actions: A motivational perspective. *Frontiers in Neuroscience* 7: 46.