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