CAN PREDICTIVE PROCESSING EXPLAIN SELF-DECEPTION?

Marko Jurjako
mjurjako@ffri.uniri.hr

Department of Philosophy and Division of Cognitive Science
Faculty of Humanities and Social Sciences
University of Rijeka
Rijeka, Croatia, European Union

Abstract

The prediction error minimization framework (PEM) denotes a family of views that aim at providing a unified theory of perception, cognition, and action. In this paper, I discuss some of the theoretical limitations of PEM. It appears that PEM cannot provide a satisfactory explanation of motivated reasoning, as instantiated in phenomena such as self-deception, because its cognitive ontology does not have a separate category for motivational states such as desires. However, it might be thought that this objection confuses levels of explanation. Self-deception is a personal level phenomenon, while PEM offers subpersonal explanations of psychological abilities. Thus, the paper examines how subpersonal explanations couched in the PEM framework can be thought of as related to personal level explanations underlying self-deception. In this regard, three views on the relation between personal and subpersonal explanations are investigated: the autonomist, the functionalist, and the co-evolutionary perspective. I argue that, depending on which view of the relation between the personal and subpersonal is adopted, the PEM paradigm faces a dilemma: either its explanatory ambitions should be reduced to the subpersonal domain, or it cannot provide a satisfactory account of motivated reasoning as instantiated in self-deception.

Keywords: Desires; motivated reasoning; predictive processing; prediction error minimization; personal/subpersonal explanations; self-deception

Introduction

The prediction error minimization framework (PEM) denotes a family of views that aim at providing a unified theory of perception, cognition, and action (see, e.g. Clark, 2016, 2013; Hohwy, 2013; Friston, 2010). According to this framework, the brain enables adaptive behavior by minimizing prediction error, that is, the mismatch between its hierarchical generative model and environmental stimuli. Following Jakob Hohwy (2013), I will refer to this family of views as the prediction error minimization (PEM) framework.

According to its main proponents, PEM provides an overarching framework for building theories in cognitive science (Clark, 2013, 2016; Hohwy, 2013; for discussion, see Litwin & Miłkowski, 2020). Its roots go at least back to Herman von Helmholtz's idea that the brain enables perception by constructing hypotheses about what it perceives and testing them against incoming environmental stimuli (Clark, 2013; Hohwy, 2013). More recently, this pattern of explanation has been generalized to account for other neurocognitive abilities and phenomena, including attention, emotion, action, and psychiatric disorders (e.g. Clark, 2016; Gadsby & Hohwy, 2021; Mathys, 2016). Indeed, according to its main proponents, PEM's successes in accounting for the various cognitive processes and the capacities of the brain, strongly support the idea that the single principle of prediction error minimization describes the brain's overarching computational task that underpins all of its subpersonal cognitive processes (Clark, 2013; Hohwy, 2013; see, also Sprevak, 2021a).

In addition to various phenomena studied by cognitive neuroscience, PEM also promises to account for recognizably personal level phenomena, such as our conscious experiences (see, e.g. Clark, 2020; Dołęga & Dewhurst, 2021; Hohwy, 2013) and to enrich our self-conceptions (see, e.g. Clark, 2016, p. 82; Colombo & Fabry, 2021). In this regard, Hohwy emphasizes the broad ambitions of PEM, asserting that it has the capacity to account for "perception and action and everything mental in between" (Hohwy, 2013, p. 1). According to Hohwy:

> The prediction error minimization framework (…) is extremely ambitious—it gives the organizing principle for brain function as such. It should then encompass and illuminate all aspects of perception and action, including aspects that cognitive science and philosophy of mind view as problematic or poorly understood. (Hohwy, 2013, p. 101; see also Clark, 2013, 2016; Friston, 2010)

Insofar PEM is construed as offering a unified framework for theorizing about the mind/brain that spans across the personal/subpersonal divide, its explanatory aims might be set too high (Colombo & Wright, 2017). As noted by Adina Roskies and Charles Wood "[t]he compactness of PP is elegant, but its aesthetic appeal would be inconsequential if it were unable to save the [recognizable psychological] phenomena" (Roskies & Wood, 2017, p. 850). In this regard, I will argue that because PEM does not have a functionally separate category for motivational states such as desires (Colombo, 2017; Dewhurst, 2017), it will have problems accounting for the garden variety cases of motivated reasoning.[1] As a case study, I will use self-deception, construed as a type of motivationally biased belief-forming process (Mele, 2001; see, also Marchi & Newen, 2022).

It should be noted that whether PEM is in the business of accounting for personal level phenomena, such as self-deception, depends on how we think about the relation between personal and subpersonal levels of explanation.[2] It is likely that not all theorists working within the PEM framework will be of the same mind about this issue. In fact, this problem is rarely explicitly discussed in the PEM literature (for an exception, see Colombo & Fabry, 2021). Thus,

[1] For other recent critical discussions of PEM's ability to offer unified accounts of psychological phenomena, see, e.g. Klein (2018), Ransom et al. (2020), Williams (2020), and Litwin and Miłkowski (2020).

[2] Thanks to an anonymous reviewer for advising me to foreground this issue.

one of the main goals of this paper is to investigate how the relation between personal and subpersonal explanations could be construed and what are its implications for the theoretical limitations of the PEM paradigm. In this regard, the argument of this paper can be read as a dilemma. Either the explanatory ambitions of PEM should be reduced to subpersonal phenomena, or PEM will be open to objections that it cannot account for garden variety forms of motivated reasoning.

The paper is structured as follows. Section 1 provides an overview of the main concepts of the PEM framework. Section 2 examines ways in which PEM can capture typical cases of self-deception and its limitations in explaining the characteristic ways that motivation can bias belief-forming processes. Section 3 explores a response to this objection based on the distinction between personal and subpersonal levels of explanation. It could be argued that self-deception does not present a problem because PEM offers subpersonal explanations, while self-deception is a personal level phenomenon. To evaluate this response, following José Bermúdez (2005), I introduce three views on the relation between personal and subpersonal explanations: autonomism, functionalism, and the co-evolutionary model. In subsection 3.1., I show that adopting the autonomist view would insulate PEM from the objection, but at the cost of excluding personal level phenomena from its explanatory purview. In subsection 3.2., I argue that the functionalist view is not consistent with PEM, because PEM's reconstruction of motivational and cognitive processes in terms of prediction error minimization significantly deviates from how they are standardly construed at the personal level. In subsection 3.3., I argue that even on the co-evolutionary construal, PEM is vulnerable to the motivation-based objections, thereby indicating that its explanatory scope may be more limited than advertised by its most vocal proponents. I conclude the paper by noting the limitations of the present discussion.

1. The conceptual toolbox of the PEM framework

The unique mark of the PEM framework is that it provides a *computational* level description of the tasks that the brain faces, which, as will be explained below, are reduced to sensory prediction error minimization (for illuminating discussions, see Sprevak, 2021a, 2021b). However, on a more general level, PEM is based on the idea that the brain is an inductive machine that implements optimal Bayesian inference through predictive coding (Clark, 2013; Hohwy, 2013).

According to PEM, the brain successfully regulates its environment by having hierarchical generative models that are implemented across cortical and subcortical layers. The purpose of these generative models is to represent the environment and the causes of the incoming sensory signals at different spatio-temporal scales (Clark, 2013).[3] The generative model is composed of prior beliefs and likelihoods. Here, prior beliefs are construed as the brain's predictions or hypotheses about the environmental conditions causing the incoming signals. Likelihood is the evidence the brain receives in the form of sensory signals. Once a generative model is in place, its values can be updated in accordance with Bayes' theorem. However, Bayesian inference is often computationally intractable (Hohwy, 2013; Mathys, 2016). Thus, to successfully regulate its environment the brain needs to rely on more biologically plausible computational procedures that can approximate Bayesian inference.

Here enters the predictive coding part of PEM, according to which, the brain updates its generative model by minimizing prediction error (Hohwy, 2020; Mathys et al., 2011). The idea is that different neuronal populations encode predictions of incoming signals and the actual signal. Based on their difference, a prediction error (PE) is produced, which the brain uses to adjust its underlying hierarchical model of the external causes. Prior beliefs serve as top-down

[3] Formally, a generative model is a joint probability relating variables and state parameters. These variables and parameters are typically called "beliefs" because they represent different aspects of internal and external environments. However, given that they are probability distributions representing subconscious states, they should not be confused with beliefs construed as propositional attitudes (Dewhurst, 2017).

predictions of the signals received at lower levels of the hierarchy. On each level, if there is a mismatch, an error signal is sent up the hierarchy and the prior values are updated. Maintaining PE minimal in the long run guarantees that the updating of hierarchical models will approximate optimal Bayesian inference (Hohwy, 2013; Mathys, 2016).

Importantly, not every PE is equally reliable. According to PEM, the influence of a PE on belief updates will depend on their precision. Formally, "precision" refers to the inverse of the variance of a variable. It can be construed as a measure of noise in a system. The noisier the signals, the less precise they are, and consequently, belief updates should rely less on them. The idea is that the brain's belief updates depend on its prior beliefs about the relevant causes and their *precision* weighted prediction errors. Moreover, the brain can have prior beliefs about the precisions of PEs in order to anticipate and optimize their influence in accordance with situational demands (Clark, 2013; Hohwy, 2013). For instance, when driving on a foggy day, my brain can optimize the impact of PEs by anticipating their lower precisions in those circumstances. Thus, it can attenuate their impact on belief updates and enable me to rely more on prior expectations of the conditions on the road.

Crucially, PE can be minimized either by perceptual or active inference. Updating the model corresponds to perceptual inference. For instance, if I expect a glass in front of me, but the glass is placed beside me, then the PE can be minimized by updating the belief about the glass' position. Alternatively, the PE can be minimized by actively placing the glass in front of me, and thereby adjusting the world to cause incoming stimuli that fit my prior predictions. This corresponds to active inference.

The notion of active inference is important because of its potential to provide a unitary account of action and cognition (Clark, 2013; Friston, 2010; Hohwy, 2013). According to PEM, cognitive and motivational states emerge from the same computational task that the brain is performing, namely the precision weighted prediction error minimization (e.g. Clark, 2020;

Miller Tate, 2021; Pezzulo et al., 2018). The only difference between the mental processes consists in how they minimize prediction error. Perceptual and cognitive processes minimize PE by updating prior expectations, while motivational states are those prior expectations that minimize PE by engendering action that "updates" the world (Clark, 2020; Hohwy, 2013, p. 89).

This ability to offer a parsimonious and unified picture of perception, cognition, and behavior seems to be a theoretical virtue of PEM (cf. Litwin & Miłkowski, 2020). However, I will argue that PEM's structural features present an obstacle to accounting for personal level phenomena that involve the characteristic interaction between motivational and cognitive states as displayed in self-deception. To pave the way for the argument, let us first examine how self-deception could be modeled within PEM.

2. Can self-deception be explained within the PEM framework?

For the purposes of this paper, I will rely on a familiar type of case that can be called wishful self-deception (see, also Krstić, 2021; Van Leeuwen, 2007). Here, a person wants some proposition to be true, and this desire, despite evidence to the contrary, biases their belief forming processes towards accepting it (Marchi & Newen, 2022; Mele, 2001). There are other types of self-deception (for recent discussions, see, e.g. Funkhouser, 2019; Krstić, 2021), but to keep things simple, I will refer to wishful self-deception simply as self-deception. An influential account of this type of self-deception is provided by Alfred Mele. According to him, person S forms a self-deceptive belief that p if:

1. The belief that p which S acquires is false.

2. S treats data relevant, or at least seemingly relevant, to the truth value of p in a motivationally biased way.

3. This biased treatment is a nondeviant cause of S's acquiring the belief that p.

4. The body of data possessed by S at the time provides greater warrant for ~p than for p.

As an example, consider the following case:

> (The parents case) A son is on trial for murder. There is overwhelming evidence showing that the son is guilty. His parents are present at the trial, and they are aware of all the evidence presented by the prosecution indicating that the son is guilty. Based on this evidence, we suppose that an optimal Bayesian agent would find greater support for the judgment that the son is guilty than that he is innocent. However, the thought that he is guilty is unbearable for the parents. This affective state creates an incentive for the parents to think that the son is innocent, which in turn biases their reasoning. It leads them to form and maintain the belief that their son is innocent. It also frames how they think about the available evidence, what kind of information they will find rewarding, and how it impacts their belief updates. Consequently, these motivational and affective states ground and maintain the *epistemically unreasonable* belief that the son is not guilty on all charges.[4]

Examples of self-deception are usually devised for the purposes of philosophical discussion. However, wishful self-deception is instantiated in empirically well researched phenomena involving motivated reasoning (Mele, 2001; see, also Chance & Norton, 2015). One such example is optimism bias, according to which people tend to asymmetrically update their beliefs, depending on the desirability of the contemplated information (see, e.g. Jefferson et al., 2017). For instance, studies show that when presented with evidence that they belong to a socioeconomic group where the prevalence of some disease is higher than subjectively expected, people rely less on this information when updating beliefs about the likelihood of

[4] To keep things simple, I will refer to these motivational states as desires, which should be understood as an umbrella term for various pro-attitudes that play the functional role of goals or ground goals in commonsense psychological explanations (see, e.g. Smith, 1987).

suffering from such disease in the future (e.g. Sharot, 2011). In contrast, belief-updating patterns are opposite when the evidence points to more desirable prospects than initially expected (for instance, indicating that the prevalence of some disease in one's socioeconomic group is lower than subjectively estimated). Many cases of optimism bias can be construed as self-deception because they typically involve *biased* or *unreasonable* belief updating that is causally explained by the subjective (un)desirability of the available information (Kuzmanovic et al., 2018; Sharot & Garrett, 2016; see also Sullivan-Bissett 2022, sec. 4).

Within the PEM framework, this type of biased reasoning might be explained by varying the precision parameter. To see how this might be done, let us consider how precision in general influences belief updating. Precision can be unpacked into the precision of the incoming signal and the precision of the prior belief. If the brain receives a PE, the direction and amount of belief update will depend on the relative precision of the prior belief and the incoming signal. For instance, if under normal circumstances, I have a high-level expectation to see a dog on a doormat, while at the same time I am confronted with a flow of lower-level unexpected stimuli, such as visual stimuli of something fury, but cat-like, producing meowing sounds, and does not respond to my calls, this produces a precise PE that updates my expectation about what is on the doormat. However, if a prior belief about the cause of a signal is more precise, i.e., more reliable, or less noisy than the precision of the incoming signal, then the update will be more influenced by the prior belief.

Self-deception could be explained as a case of inappropriate[5] overweighting or underweighting of the precision of priors, or overweighting or underweighting the precision of the incoming stimuli. For instance, in case of the self-deceived parents, having overly precise

[5] Here the "inappropriateness" of weighting is determined with respect to how the optimal Bayesian agent would weigh the precisions of priors and incoming stimuli (see, e.g. Kuzmanovic & Rigoux, 2017). From the perspective of the PEM framework, belief updating always approximates optimality. The only way for the PEM framework to capture irrational or inadequate belief updating is in terms of the suboptimal generative models (see, e.g. Gadsby & Hohwy, 2021; Schwartenbeck et al., 2015). The significance of this claim will be discussed in subsections 3.2. and 3.3.1.

expectations about their son’s innocence could explain why the parents disregard PE (i.e., disconfirming evidence) as irrelevant. Alternatively, they may overweight the precision of evidence indicating that the son is innocent thereby reducing the probability associated with the belief that the son is guilty. Moreover, in terms of active inference we can explain how misinterpretation of evidence and selective focusing of attention might influence belief updating. For instance, given their strong and precise priors that the son is innocent, the parents’ attention, via active inference, could be selectively focused on the evidence that favors their prior beliefs. In turn, these precise priors could also motivate them to selectively sample the available evidence until they find some confirmation for them (for a more detailed account of self-deception within PEM, see Marchi & Newen, 2022).

Despite its capacity to illuminate important aspects of self-deception, the balancing of precisions as such is not sufficient to explain self-deception. The missing ingredient is the desire for the son to be innocent. This desire *explains* the differences in the precisions of the priors between the parents and the jury, and why the parents’ pattern of belief updating leads to self-deception while the jury’s does not. Indeed, in typical cases of motivated reasoning, it is thought that the *subjective desirability* of information modulates belief updating (for discussion, see Williams, 2021, sec. 4.3). However, PEM radically departs from the desire-belief psychology because it eschews desires as functionally separable states (e.g. Colombo, 2017; Dewhurst, 2017). Consequently, PEM seems to lack the conceptual resources for explaining self-deception and similar phenomena in which desires cause deviations from normal belief-forming processes.

Moreover, without the commonsensical notion of desire, it seems that PEM cannot distinguish motivationally based phenomena, such as self-deception and optimism bias, from other similar, but non-motivationally-valenced belief updating. For instance, both confirmation bias (in the sense of a general human tendency to overweigh prior beliefs regardless of the desirability of their contents) and hallucinations can be explained as consequences of having overly precise prior beliefs in comparison to underweighted prediction errors (Powers et al., 2017; Sharot & Garrett, 2016). Similarly, delusional beliefs can be explained as aberrant encoding of the relative precision of priors and prediction errors (Sterzer et al., 2018). However, what distinguishes self-deception and optimism bias from these other types of aberrant encodings of priors are the specific ways in which motivational considerations hijack belief-forming processes (see, e.g. Mele, 2006; Sharot & Garrett, 2016). In other words, the involvement of desires explains why a certain instance of aberrant PE minimization should count as motivationally biased belief formation instead of a hallucination or a general case of confirmation bias (see Chance & Norton, 2015).

These objections, however, might be artifacts of a failure to properly distinguish between the personal and subpersonal descriptions of cognitive processes. At the personal level we explain people's actions in terms of their desires and beliefs that provide reasons for action (e.g. Dennett, 2010; Drayson, 2012). Accordingly, when explaining that a person forms a belief that p because they want p to be true, we talk about their belief forming processes at the personal level of explanation. In contrast, PEM provides subpersonal level explanations that involve computational mechanisms underpinning various cognitive capacities (e.g. Clark, 2020; Colombo & Fabry, 2021; Dewhurst, 2017; Pezzulo et al., 2018; Sprevak, 2021a; Yon et al., 2020). At that level we should be able to explain the underlying mechanics of motivated reasoning in terms of various patterns of prediction error minimization. Thus, it could be claimed that desires individuate and explain self-deception at the personal level of explanation. In contrast, PEM provides the subpersonal computational mechanism that explains how the personal level processes are implemented in the brain.

Whether this response protects PEM from the objection remains yet to be shown. The problem is that there is more than one way of thinking about the relation between personal and subpersonal explanations, and not all of them will protect PEM from the self-deception challenge. Another problem is that PEM theorists are rarely explicit about how they construe the relation between personal and subpersonal explanations (cf. Colombo & Fabry, 2021). To tackle this problem, in the next section, I will distinguish between three views of the personal/subpersonal explanations and examine their implications for PEM's capacity to respond to the self-deception challenge.

3. PEM, self-deception, and personal/subpersonal explanations

Following Bermúdez (2005), we can distinguish between the autonomist, functionalist, and co-evolutionary views regarding the relation between personal and subpersonal explanations.[6] I will start with the autonomist view.

3.1. PEM and the autonomist view

For some autonomists the distinction between personal and subpersonal levels captures the distinction between different *kinds* of explanations (see Dennett, 2010; Drayson, 2014). Usually, they construe personal level explanations as normative explanations that make people's behavior intelligible under the supposition that they approximate some ideal of rationality (see, e.g. Davidson, 2001; McDowell, 1994). The idea is that at this level we can explain why people behave in certain ways by invoking their reasons, that is, desires and beliefs that would be satisfied by this behavior. For instance, in the parents example, we can say that although their belief forming processes are epistemically irrational, still their belief that the son

[6] In the only discussion so far on the relation between personal and subpersonal explanations within the context of PEM, Colombo and Fabry (2021) opt for the co-evolutionary model on independent grounds. However, in the present paper, my goal is not to adjudicate the plausibility of the different views of the relation between personal and subpersonal levels. Granted their initial plausibility, my goal is instead to investigate whether any of these views can be used to defend the PEM paradigm from the criticism that it cannot explain crucial features of self-deception.

is innocent makes *sense* from the perspective of their desire that the son be innocent. In contrast, subpersonal explanations are typically understood as referring to nonrational causal mechanisms that are used to describe, for instance, physiological processes in the brain. Sometimes, the distinction is made in terms of part-whole relations, where personal explanations refer to the whole person, while subpersonal explanations refer to the functioning of their parts, in our case the brain (Bennett & Hacker, 2003, ch. 3; cf. Hornsby, 2000).

In both cases, the autonomist view is that that there is a categorical distinction between personal and subpersonal levels, where each level forms an autonomous explanatory domain (Hornsby, 1997). In this regard, providing explanations that cut across the divide between the personal and subpersonal levels would amount to conceptual confusion or a change of topic (Bennett & Hacker, 2003). In general, according to the autonomists, the most that we can say about the relation between the personal and subpersonal levels of explanation is that the latter provides causally enabling preconditions for the former (see, e.g. Hornsby, 1997, p. 166).

Sam Wilkinson (2014a) seems to endorse this view. He does not explicitly discuss the autonomist view in the context of the PEM framework, nonetheless, in other papers he argues that personal and subpersonal explanations respond to different explanatory concerns and thus offer different *kinds* of explanations (see Wilkinson, 2014b, 2015). More specifically, according to Wilkinson, personal explanations respond to "why" questions, while the neuroscientific explanations respond to "how" questions. For instance, the activation of dopamine neurons can explain *how* the brain reacts to unexpected rewarding stimuli. However, it cannot explain *why* we judge certain things to be desirable. For this we need to think about the person's beliefs and rational capacities that underlie their value judgments.

It seems clear from the description of the autonomist view that adopting it would limit the explanatory scope of PEM to subpersonal level phenomena. Given that self-deception is essentially individuated and explained as a desire-based personal level phenomenon, according to the autonomist view it would follow that PEM is not in the business of explaining self-deception (or any other kind of personal level phenomena). Thus, adopting this view can dispel the self-deception objection, but at the expense of greatly reducing PEM's capacity for providing unified explanations of action and cognition.

Adopting the autonomist perspective seems at odds with the theoretical ambitions of the most vocal proponents of PEM. In particular, some find PEM explanations of personal level phenomena illuminating, to the point that they might even revise how we think about ourselves at the personal level of functioning (e.g. Hohwy, 2013; Clark, 2016, 2020; Colombo & Fabry, 2021). This plausibly indicates that for many the autonomist view does not really capture the relation between personal and subpersonal explanations that the proponents of the PEM presuppose in their discussions. Thus, it is worth considering whether adopting the functionalist or the co-evolutionary model might help to meet the self-deception challenge.

3.2. PEM and the functionalist perspective

Functionalist approaches construe the relation between personal and subpersonal explanations in terms of functional roles and their underlying realizers (Drayson, 2012). Accordingly, personal level states are defined by the casual roles they play in commonsense-psychological explanations, while subpersonal states can be seen as realizers of those roles at different subpersonal levels of explanation. Some functionalist approaches construe personal explanations as involving causal law-like generalizations that with subpersonal explanations form a reductive hierarchy (Kim, 1993). Others, due to the multiple realizability of mental states, see personal level law-like generalizations as irreducible to law-like generalizations of the lower-level sciences (Fodor, 1974; for discussion, see Ross & Spurrett, 2004). Still others think that the main goal of psychological explanations is to offer a functional analysis of psychological capacities. Functional analysis aims at explaining, at different levels of cognitive and neuroscientific description, how psychological capacities operate in terms of their underlying subcapacities and their functional organization (for discussion, see Piccinini & Craver, 2011; Roth & Cummins, 2018). Despite these differences, the unifying theme in the functionalist approaches is their top-down methodology, where personal level explananda are implemented and explained by subpersonal states and events (Bermúdez, 2005).

PEM models can be thought of as providing neurocomputational mechanisms that connect the (neuro)physiological levels of description with everyday experience as captured by ramified commonsense psychology (see, e.g. Clark, 2016; Miller Tate, 2021). In this regard, precision weighted PE minimization can be considered as a model of the subpersonal neurocomputational processes that, depending on the context, implement personal level perceptual, cognitive, and motivational processes (Hohwy, 2020). More specifically, desires and other motivational states could be thought of as subpersonally implemented or realized by prior beliefs that engender action via active inference (Clark, 2020; Miller Tate, 2021).

In some places, Andy Clark seems to presuppose such a functionalist view when stating that predictions might be thought of as realizing/implementing desires:

> Consider, once again, my desire to go to see a certain movie tonight. [PEM] *realizes* this desire as a high-level prediction that (when estimated as sufficiently precise) entrains apt actions at many time-scales. (Clark, 2020, p. 7, emphasis added)

Similarly, Hohwy, in some places, hints at a functionalist view. For instance, he claims that "desiring a muffin *is* having an expectation of a certain flow of sensory input that centrally involves eating a muffin" (Hohwy, 2013, p. 89, emphasis added). If the "is" from the example is read in a reductive fashion, then we get a functionalist view according to which prior beliefs realize desires. Alex James Miller Tate also seems to presuppose a functionalist view of PEM when stating that:

> The central claim of the predictive processing account of motivation I will be putting forward is that some kind of dependence relation holds between [the] predictions of the generative model and our motivational mental states. (Miller Tate, 2021, p. 4512)

He leaves it open whether the "dependence relation" should be read as identity, supervenience, or grounding (see Miller Tate, 2021, footnote 18). Nonetheless, all the interpretations are consistent with the functionalist view according to which PEM implements or realizes commonsensically construed motivational states.

There is reason to think that PEM on the functionalist construal cannot offer a sufficient explanation of self-deception. The problem is that prior beliefs cannot capture the distinctive functional roles desires and other affective states play in self-deception, that is, in causing epistemically irrational belief-forming processes.

To elaborate on this point, consider how PEM explains false inference and maladaptive behavior characterizing various psychological disorders. As an example we can take a person who suffers from persecutory delusions and has false beliefs about their surroundings, such as, that people are out there to get them, that somebody is sabotaging their work, and so on. Within PEM, such delusions can be explained in terms of aberrant prediction error weighing (see, e.g. Powers et al., 2017). For instance, the person may have a very precise prior belief that their co-workers are sabotaging their work and based on this precise prior dismiss prediction errors indicating that this is not the case.

Crucially, however, PEM cannot explain distorted belief and other pathologies as causal effects of irrational or suboptimal inferences. This is because prediction error minimization is supposed to provide a biologically plausible mechanism for approximating optimal Bayesian inference (Clark, 2013; Hohwy, 2013), which in turn represents our ideal of epistemic rationality (see, e.g. Mathys, 2016; Williams, 2021). The only option for PEM to account for distorted belief formation and other pathological phenomena is in terms of aberrant prior beliefs, that is “generative models that ‘suboptimally’ approximate the true causal structure of the world” (Schwartenbeck et al., 2015, p. 111). Indeed, the proponents of PEM emphasize that this option is guaranteed to work by “the complete class theorems (…) that state that any behavior, no matter how apparently pathological, is Bayes optimal under the right set of prior beliefs (Parr & Friston, 2021, p. 172). Moreover, they consider this explanatory strategy as advantageous, because explaining things in terms of generative models and prior beliefs should enable us, among other things, “to identify the origins of aberrant inference in the cerebral hierarchy” and “to differentiate between different mechanisms that might cause (…) deviation” that characterizes various pathological conditions (Schwartenbeck et al., 2015, p. 111; see, also Gadsby & Hohwy, 2021).

But, for this same reason PEM cannot be thought of as providing a subpersonal computational implementation of self-deception in accordance with the functionalist model. The problem is that if the functional role of prior beliefs is determined in relation to the process of minimization of prediction error, then prior beliefs cannot account for the role that desires are supposed to play in causing irrational belief-forming processes that underpin self-deception. Whatever the content of prior beliefs and the generative models they are part of, the inferences they ground will approximate optimality. However, when it comes to self-deception, desires are supposed to be exactly those states that cause normal belief-forming processes to deviate from optimality, regardless of whether the initial generative model is optimally or suboptimally representing the causal structure of the world.

To illustrate this last point, we can modify the self-deceiving parents example. Imagine that there is a juror who does not have any specific feelings or interests in the son’s being innocent, whose posterior beliefs are, nonetheless, skewed towards the prior that the son is not guilty. Let us suppose that it just so happens that in this respect the juror’s priors are identical to the parents’ priors and that their subpersonal belief updates exhibit the same pattern. This could be because the juror holds a very strong belief that the police framed the son for murder. More specifically, this prior is based on the juror's evidentially ungrounded impression that people belonging to the son's socioeconomic class tend to be systematically disadvantaged by the social institutions. Thus, for different reasons the juror and the parents will share the same generative model of the trial and update their beliefs in a similar fashion. This means that in terms of PEM, the juror and the parents would both exhibit a sort of confirmation bias that is based on the overweighted (suboptimal) priors instead of objective strength of the available evidence. However, reducing self-deception to confirmation bias in this way obscures the crucial difference between the juror and the parents from the example. The parents' belief-forming processes are epistemically irrational because they are motivationally biased. In other words, their priors are suboptimal because they are based on motivationally biased, and thus, suboptimal inferences. In contrast, the juror need not be epistemically irrational because their priors, despite being based on a suboptimal generative model, cohere well with their other beliefs, and thus ground rational inferences.[7]

It might be thought that the objection that PEM cannot capture the biasing role of desires on reasoning processes can be avoided by introducing appropriate distinctions at the level of generative models. In particular, Francesco Marchi and Albert Newen (2022) offer a predictive processing account of self-deception according to which motivational influences on model updating are explained by distinguishing two types of generative models that a brain might implement (Marchi & Newen, 2022, p. 12). According to Marchi and Newen, intelligent organisms have a "world-model" whose functional role is to provide "an accurate representation of the world, such that the biological system is able to survive in the world" (2022, p. 11). However, given that "humans are hyper-social beings" they also benefit from having a "self-

[7] Thanks to an anonymous referee for pressing me to be clearer about the purpose of this example.

model". The self-model has two functional roles. The first one is to enable stable representations of oneself. This includes having representations of

> (…) bodily properties of myself like how far I can reach (peripersonal space), whether I can fit through a certain opening, and many other bodily properties which are also characterized as constituting a body schema (…). (Marchi & Newen, 2022, p. 11)

The second function of a self-model is to enable "an adequate understanding of others", with the ultimate aim of enabling "a social living being to be accepted in the relevant social groups" (Marchi & Newen, 2022, p. 11).

To fulfil these functions, the world-model and the self-model will need to exhibit different updating patterns. On the one hand, given the world-model's function of accurately representing the external environment it will tend to minimize prediction error by updating the model. On the other hand, for the self-model to perform its function properly it will need to represent stable information about oneself:

> (…) we need a rather coherent and stable self-model to act in the world and this includes a representation of one's social rank in a group and also one's assessment of the social expectations a group has toward one: it makes one predictable for others, allows one to plan the future and to express or communicate this self-model in behavior and especially in telling stories about oneself to others (…). (Marchi & Newen, 2022, p. 12)

Thus, given that the self-model is supposed to represent stable information about oneself it will tend to minimize prediction error by attenuating the precision of the PE in favor of the prior beliefs. In this regard, self-deception might be construed as the personal level expression of the system's pattern of updating the self-model. When applying this distinction to the previous example, we can imagine that the parents are motivationally biased because their self-model includes precise priors related to how they feel about their son, while the juror's self-model will not have such expectations because they are not emotionally invested in the case.

Although Marchi and Newen offer an interesting PEM account of self-deception, it is not clear whether it can capture the *biasing* role of desires in cognitive processes. In addition, it is not clear that this account is best captured by the functionalist model of the personal/subpersonal distinction. There are several reasons for this.

First, it is not clear at what level of explanation Marchi and Newen's account is supposed to work. In fact, their account seems to crosscut the distinction between personal and subpersonal explanations. On the one hand, when Marchi and Newen explain that the purpose of the world-model is to maximize the accuracy of representations of the world, while the goal of the self-model is, among other things, to maximize success in social interactions, they seem to be using "model" as a personal level construct. In that case, however, they are simply redescribing in commonsense terms what they think self-deception consists in, rather than providing a subpersonal account of what realizes motivationally biased reasoning. On the other hand, if these descriptions of the world-model and the self-model are supposed to be read as depicting the subpersonal level, then it is not clear that they can be accounted for in PEM terms. This is because from the perspective of PEM the role of all models is to minimize prediction error. Indeed, Marchi and Newen seem to admit this much when writing the following:

> The way in which we have presented the two models relies heavily on the concepts and language of *folk-psychology*, but it is not clear whether predictive processing has the tools to capture the conceptual richness of such a language. In fact, a truly generalized version of predictive processing may need to dispense with any talk of beliefs, desires, and such, and only rely on the language of Bayesianism (prior, posterior, likelihood, precision, etc.). (Marchi & Newen, 2022, p. 16, emphasis in the original)

Thus, it remains unclear how the distinction between the world-model and the self-model is supposed to be captured within the PEM framework and help us to think about the subpersonal implementation of self-deception under the functionalist model.

Second, this account does not solve the original objection that PEM cannot capture the biasing effect of motivational states on belief-forming processes. Marchi and Newen's view seems to be that the biasing processes underpinning self-deception are implemented in the updating patterns of the self-model. In this regard, they write:

> In order to prevent encountering counter-evidence to the contents of the self-model or in order to deal with any counter-evidence encountered, the system operates *biased prediction-error minimization*, in the form of biased active inference. In this way the system may be conceived as to believe the contents of the world model, which update normally, and to desire the contents of the self-model, which resist update and engender biased active inference. (Marchi & Newen, 2022, p. 13, emphasis in the original)

However, it is not clear in what sense the system in question is biased. Of course, we think of it as biased at the personal level. Our problem, however, is how to capture the biasing influence of desires at the subpersonal level. The problem is that at the PEM level of description, the update patterns of the self-model and the world-model will both approximate optimal inference. In this regard, neither the priors from the world-model, nor the priors from the self-model can implement the role of motivational states that cause suboptimal inference.

Of course, we might think that the concept of bias has a different interpretation at the subpersonal level. For instance, it might be suggested that PEM offers a view according to which biased belief-forming processes are those that are based on suboptimal priors. However, this would involve abandoning the functionalist model of the personal/subpersonal because, instead of offering an account of subpersonal realization of biased reasoning processes, we would be revising what it means to be a biased reasoning process. Such an approach would be more congenial to the co-evolutionary model of the personal/subpersonal distinction. This motivates us then to consider whether a co-evolutionary model can be used to account for self-deception within PEM.

3.3. PEM and the co-evolutionary perspective

The proponents of the co-evolutionary model, instead of top-down analysis, emphasize the importance of bottom-up neuroscientific constraints for developing plausible theories of the mind (P. S. Churchland, 1986; see also Colombo, 2013). According to this model, there is no sharp distinction between personal and subpersonal explanations. What was once thought of as exclusively subpersonal might be used to characterize people at the personal level. For instance, if people were to start to describe and explain their mental lives in terms of neuron firings and neurotransmitter activity, then that would indicate that we had devised new ways of thinking about ourselves at the personal level. Given that the co-evolutionary model stresses the importance of bottom-up constraints, in principle it might allow for radical consequences in the form of eliminativism about commonsense-psychology (P. M. Churchland, 1981; Dewhurst, 2021).

Many proponents of PEM often construe its explanations as offering revisionist views on various psychological phenomena. Clark, for instance, claims that the reconceptualization of beliefs and desires in terms of prior probabilistic beliefs offers a "revisionary […] picture of minds" (Clark, 2020, p. 12). Hohwy also adopts the revisionary view when claiming that according to PEM "[b]eing an agent (…) reduces to a matter of optimizing expected precisions of proprioception, which is a far cry from our commonsense idea of what makes an agent" (Hohwy, 2013, pp. 83–84; see also Colombo, 2017). In general, the revisionary implications of PEM regarding motivation and agency are often expressed with the idea that PEM, in contrast to the traditional/commonsensical view of cognitive ontology, offers desert landscapes, in the sense that it depicts "a world in which value functions, costs, reward signals, and perhaps even desires have been replaced by complex interacting expectations that inform perception and entrain action" (Clark, 2016, p. 129; see, also Hohwy, 2013, p. 89; Pezzulo et al., 2018; Yon et al., 2020). Moreover, Matteo Colombo and Regina Fabry (2021) explicitly argue that because of the revisionary implications associated with adopting the PEM framework, it is best to construe it along the lines of, what we are calling, the co-evolutionary model of the personal/subpersonal distinction.

The co-evolutionary model offers a seemingly neat solution to the problem of accounting for phenomena, such as self-deception, that violate PEM's structural/optimality constraints. Since the co-evolutionary model does not primarily aim to preserve psychological phenomena as currently construed at the personal level, it reduces the pressure to find the subpersonal functional equivalents of personal level states. Indeed, it could be argued that PEM *reconceptualizes* self-deception in terms of prior beliefs that play a role in active inference. Moreover, by applying the co-evolutionary model, we could avoid the objection that PEM cannot explain how desires cause belief-forming processes to be epistemically irrational. If the role of desires in self-deception is reconceptualized in terms of overweighted priors, for example, then, similarly, we may reconceptualize the irrationality of having such priors in terms of suboptimal generative models that are comprised of them (e.g. Schwartenbeck et al., 2015). Accordingly, self-deception could be reconstrued as active inference that is based on suboptimal prior beliefs.

Under the co-evolutionary interpretation, Marchi and Newen's (2022) account could also be understood as offering a reconceptualization of self-deception in terms of an active inference that is induced by overweighted priors comprising the *self-model*. Moreover, under this interpretation, Marchi and Newen's account would avoid the objection that it illegitimately switches between the personal and subpersonal levels when explaining the role of the self-model in self-deception. This is because the co-evolutionary model, to a certain extent, presupposes reciprocal relations between the personal and subpersonal explanations. Thus, if subpersonal PEM explanations can inform how we think about ourselves at the personal level, then the personal level should also be allowed to inform our subpersonal accounts (see, e.g. Colombo & Fabry, 2021).

Nevertheless, Marchi and Newen's account might still have problems capturing what would specifically be *suboptimal* about priors that underpin self-deception. According to their account, self-deception can be construed as a form of confirmation bias that is specifically tied to the self-model (see, e.g. Marchi & Newen, 2022, pp. 13–15). However, given their own account of the self-model, it is not clear in what sense active inference could be biased. More specifically, if the function of the self-model is to immunize its priors from disconfirming PEs, then it becomes unclear in what sense active inference engendered by the self-model is supposed to be biased or why we would consider the priors that ground it as *suboptimal*. If the goal of the self-model is to maximize social efficiency and this is done via self-deception, then the priors that engender motivationally "biased" belief updating should be considered as optimal. Thus, it remains to be seen how the notion of motivational bias is supposed to be captured within Marchi and Newen's PEM account of self-deception.

However, even setting aside these open issues for Marchi and Newen's account, there are more general problems with explaining motivated belief-forming processes in terms of PEM. Reconceptualizing self-deception or more generally motivated reasoning in terms of PEM might be a solution countenanced by the co-evolutionary model, but it is not without costs. On the one hand, the problem is that PEM's reconceptualization of motivated reasoning might be too revisionistic, to the point of eliminating the phenomenon instead of explaining it. On the other hand, preserving the existence of motivated cognition may involve conceding to the objection that PEM's explanatory capacities are more limited than advocated by its most ardent proponents. In the last subsection of the paper, I consider this tension between PEM's capacity to offer unifying explanations of cognitive phenomena and its tendency to revise or explain them away.

3.3.1 Reconceptualizing or eliminating motivated reasoning? The explanatory limits of PEM

The problem with reconceptualizing motivated reasoning in terms of inferential processes that are based on suboptimal priors, is that, in an important sense, it amounts to claiming that there are no motivational biases that underpin irrational belief-forming processes. Or if there are, PEM changes them into something *very* different from how they are normally construed and modeled. The standard view is that motivated reasoning involves desires that cause deviations from optimality in cognitive processes (see, e.g. Kuzmanovic et al., 2018; Kuzmanovic & Rigoux, 2017; Sharot & Garrett, 2016; Williams, 2021). As we saw earlier, PEM, given its structural features, does not countenance cognitive processes that systematically deviate from optimal inferences. From this it follows that there are no desires that could cause suboptimal inferences. This means that PEM, in its co-evolutionary guise, might be more properly understood as eliminating self-deception and similar motivationally biased belief-forming processes, rather than explaining them. As noted by Daniel Williams in a closely related context,

> (…) this response is quite extreme: both commonsense and everyday observation attest to the biasing influence of an individual’s motives on thought, and there is an enormous empirical literature documenting examples of motivated cognition (for a review, see Kunda, 1990)” (Williams, 2021, p. 928).

By “motivated cognition”, Williams of course refers to cognitive processes that due to motivational factors *deviate* from optimal Bayesian inference. Here, I have been emphasizing wishful self-deception as one such commonsensical example (Chance & Norton, 2015; Marchi & Newen, 2022; Mele, 2001). Given that PEM inherits the Bayesian optimality constraints, it seems that instead of illuminating the many facets of the mind/brain, adopting it would entail eliminating or revising beyond recognition many phenomena that seem to have good commonsensical and empirical standing.

This leads us to the other point regarding the theoretical and explanatory ambitions of the PEM paradigm. PEM was supposed to give us a unified conception of cognition and illuminate “perception and action and everything mental in between” (Hohwy, 2013, p. 1). Now, we see that PEM has problems capturing garden variety cases of motivated reasoning, at least when such phenomena are taken at face value as epistemically suboptimal desire-based belief forming processes. Although, at first blush, it seemed that PEM might offer a unified picture of the mind, cognition, and agency, and connect personal and subpersonal levels of explanations, now it seems it can do this only at the expense of radically reconstructing what it means to have a mind and be an agent. Of course, this price might be worth paying if reconstructing cognitive phenomena in terms of PEM’s conceptual toolbox would bring more methodological, theoretical, and empirical benefits compared to other models that purport to account for the same empirical phenomena (for a classic discussion of this issue, see P. M. Churchland, 1981; see, also Fink & Zednik, 2017). In that case, however, the proponents of PEM should be more upfront about PEM’s potential for eliminating garden variety psychological phenomena, and not gloss over it with proclamations about its capacity to offer unified accounts of the mind/brain (cf. Colombo, 2017).

4. Conclusion

The PEM paradigm provides a very ambitious framework for theorizing about mental phenomena in that it purports to capture all interesting aspects of our minds and brains (Clark, 2013; Friston, 2010; Hohwy, 2013). Thus, it is not surprising that there are critical voices discussing the explanatory limits of the PEM paradigm (Litwin & Miłkowski, 2020). I have joined the critical camp by arguing that PEM faces challenges when it comes to explaining self-deception. More generally, the problem is that PEM seems to lack the conceptual resources to explain how motivations could cause suboptimal belief-forming processes.

The limitations of the present discussion should be noted. The present discussion suggests that PEM cannot provide a unitary account of all interesting aspects of cognition and motivation, and their interrelations. However, this claim is compatible with the fact that particular process models and theories derived from the general principles of the PEM paradigm can successfully account for different features of perception, cognition, and behavior (see, e.g. Friston et al., 2017). Moreover, the present discussion is not meant to provide a final blow to the theoretical aspirations of PEM. Instead, it should be seen as an invitation to think harder about the conceptual limitations of the PEM paradigm and how they can be overcome to capture characteristic interactions between motivational and cognitive states that characterize important psychological phenomena.

Acknowledgments

I would like to thank two anonymous reviewers for Synthese who provided critical and constructive comments that helped to improve the paper. Special thanks go to Luca Malatesti for reading and commenting on previous versions of the manuscript. Different versions of the paper were discussed and presented at numerous places. Most notable are the EPSA 2019 conference (Geneva) and Contemporary Philosophical Issues PhD 2021 (Rijeka). Thanks to Joe Dewhurst, Niccolò Aimone Pisano, Zdenka Brzović, and Matteo Colombo for giving valuable comments on those occasions. Part of the paper was written at the Institute of Philosophy of the University of Graz, where I was a visiting researcher supported by a scholarship from the Austrian Academy of Sciences (the JESH program). I would like to thank Norbert Paulo, Thomas Pölzler, and Lucas Meyer for supporting my grant application and for providing an ideal working environment in Graz.

Funding

This paper is an outcome of project Harm, intentions, and responsibility (HIRe) that is financed by the Croatian Science Foundation (grant number UIP- 2017-05-4308). Work on this paper is also supported by project KUBIM (University of Rijeka, grant Uniri-human-18-265).

References

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Oxford: Blackwell Publishing.

Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. Abingdon: Routledge.

Chance, Z., & Norton, M. I. (2015). The what and why of self-deception. *Current Opinion in Psychology*, *6*, 104–107. https://doi.org/10.1016/j.copsyc.2015.07.008

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78*(2), 67–90.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, Mass.: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, *98*, 1–15. https://doi.org/10.1080/00048402.2019.1602661

Colombo, M. (2013). Constitutive relevance and the personal/subpersonal distinction. *Philosophical Psychology*, *26*(4), 547–570. https://doi.org/10.1080/09515089.2012.667623

Colombo, M. (2017). Social motivation in computational neuroscience: Or if brains are prediction machines, then the Humean theory of motivation is false. In J. Kiverstein (Ed.), *Routledge Handbook of Philosophy of the Social Mind* (pp. 320-340). New York: Routledge.

Colombo, M., & Fabry, R. E. (2021). Underlying delusion: Predictive processing, looping effects, and the personal/sub-personal distinction. *Philosophical Psychology*, *34*(6): 829–855. https://doi.org/10.1080/09515089.2021.1914828

Colombo, M., & Wright, C. (2017). Explanatory pluralism: An unrewarding prediction error for free energy theorists. *Brain and Cognition*, *112*, 3–12.

Davidson, D. (2001). *Essays on actions and events*. Oxford: Oxford University Press. https://doi.org/10.1093/0199246270.001.0001

Dennett, D. C. (2010). *Content and consciousness*. Abingdon: Routledge.

Dewhurst, J. (2017). Folk psychology and the Bayesian brain. In Thomas Metzinger & Wanja Wiese (Eds.), *Philosophy and Predictive Processing* (pp. 148–160). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573109

Dewhurst, J. (2021). Folk psychological and neurocognitive ontologies. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms* (pp. 311–334). Cham: Springer. https://doi.org/10.1007/978-3-030-54092-0_14

Dołęga, K., & Dewhurst, J. E. (2021). Fame in the predictive brain: A deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*, *198*(8), 7781–7806. https://doi.org/10.1007/s11229-020-02548-9

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, *26*(1), 1–18. https://doi.org/10.1111/phpe.12014

Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass*, *9*(5), 338–346. https://doi.org/10.1111/phc3.12124

Fink, S. B., & Zednik, C. (2017). Meeting in the dark room: Bayesian rational analysis and hierarchical predictive coding. In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573154

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97–115. https://doi.org/10.1007/BF00485230

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Funkhouser, E. (2019). *Self-deception*. Abingdon: Routledge.

Gadsby, S., & Hohwy, J. (2021). Why use predictive processing to explain psychopathology? The case of anorexia nervosa. In D. Mendonça, M. Curado, & S. S. Gouveia (Eds.), *The philosophy and science of predictive processing* (pp. 209–226). Bloomsbury Academic.

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

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language, 35*(2), 209–223. https://doi.org/10.1111/mila.12281

Hornsby, J. (1997). *Simple mindedness: In defense of naive naturalism in the philosophy of mind*. Cambridge, Mass.: Harvard University Press.

Hornsby, J. (2000). Personal and sub-personal: A defence of Dennett’s early distinction. *Philosophical Explorations, 3*(1), 6–24. https://doi.org/10.1080/13869790008520978

Jefferson, A., Bortolotti, L., & Kuzmanovic, B. (2017). What is unrealistic optimism? *Consciousness and Cognition, 50,* 3–11. https://doi.org/10.1016/j.concog.2016.10.005

Kim, J. (1993). Multiple realization and the metaphysics of reduction. In J. Kim, *Supervenience and Mind* (pp. 309–335). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511625220.017

Klein, C. (2018). What do predictive coders want? *Synthese*, *195*(6), 2541–2557. https://doi.org/10.1007/s11229-016-1250-6

Krstić, V. (2021). On the function of self-deception. *European Journal of Philosophy*, *29*(4): 846-863. https://doi.org/10.1111/ejop.12608

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kuzmanovic, B., & Rigoux, L. (2017). Valence-dependent belief updating: Computational validation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01087

Kuzmanovic, B., Rigoux, L., & Tittgemeyer, M. (2018). Influence of vMPFC on dMPFC predicts valence-guided belief formation. *The Journal of Neuroscience*, *38*(37), 7996–8010. https://doi.org/10.1523/JNEUROSCI.0266-18.2018

Litwin, P., & Miłkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science*, *44*(7). https://doi.org/10.1111/cogs.12867

Marchi, F., & Newen, A. (2022). Self-deception in the predictive mind: Cognitive strategies and a challenge from motivation. *Philosophical Psychology*, 1–20. https://doi.org/10.1080/09515089.2021.2019693

Mathys, C. D. (2016). How could we get nosology from computation? In A. D. Redish & J. A. Gordon (Eds.), *Computational Psychiatry: New Perspectives on Mental Illness* (pp. 121–135). Cambridge, Mass: MIT Press.

Mathys, C. D., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00039

McDowell, J. H. (1994). The content of perceptual experience. *The Philosophical Quarterly*, *44*(175), 190. https://doi.org/10.2307/2219740

Mele, A. R. (2001). *Self-deception unmasked*. Princeton: Princeton University Press.

Mele, A. R. (2006). Self-deception and delusions. *European Journal of Analytic Philosophy*, *2*(1), 109–124.

Miller Tate, A. J. (2021). A predictive processing theory of motivation. *Synthese*, *198*, 4493–4521. https://doi.org/10.1007/s11229-019-02354-y

Parr, T., & Friston, K. J. (2021). Disconnection and diaschisis: Active inference in neuropsychology. In D. Mendonça, M. Curado, & S. S. Gouveia (Eds.), *The philosophy and science of predictive processing* (pp. 171–185). Bloomsbury Academic.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *22*(4), 294–306. https://doi.org/10.1016/j.tics.2018.01.009

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4

Powers, A. R., Mathys, C. D., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600.

Ransom, M., Fazelpour, S., Markovic, J., Kryklywy, J., Thompson, E. T., & Todd, R. M. (2020). Affect-biased attention and predictive processing. *Cognition*, *203*, 104370. https://doi.org/10.1016/j.cognition.2020.104370

Roskies, A., & Wood, C. (2017). Catching the prediction wave in brain science. *Analysis*, *77*(4), 848–857. https://doi.org/10.1093/analys/anx083

Ross, D., & Spurrett, D. (2004). What to say to a skeptical metaphysician: A defense manual for cognitive and behavioral scientists. *Behavioral and Brain Sciences*, *27*(5), 603–627. https://doi.org/10.1017/S0140525X04000147

Roth, M., & Cummins, R. (2018). Neuroscience, psychology, reduction, and functional analysis. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science* (Vol. 1). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199685509.003.0002

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. (2015). Optimal inference with suboptimal models: Addiction and active Bayesian inference. *Medical Hypotheses*, *84*(2), 109–117.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941–R945. https://doi.org/10.1016/j.cub.2011.10.030

Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. https://doi.org/10.1016/j.tics.2015.11.002

Smith, M. (1987). The Humean theory of motivation. *Mind*, *96*(381), 36–61.

Sprevak, M. (2021a). Predictive coding I: Introduction. *TBC*. http://philsci-archive.pitt.edu/id/eprint/19365

Sprevak, M. (2021b). Predictive coding II: the computational level. *TBC*. http://philsci-archive.pitt.edu/id/eprint/19366

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Sullivan-Bissett, E. (2022). Debunking doxastic transparency. *European Journal of Analytic Philosophy*, *18*(1), A3(5)-24. https://doi.org/10.31820/ejap.18.1.3

Van Leeuwen, D. S. N. (2007). The product of self-deception. *Erkenntnis*, *67*(3), 419–437. https://doi.org/10.1007/s10670-007-9058-x

Wilkinson, S. (2014a). Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework. *Consciousness and Cognition*, *30,* 142–155. https://doi.org/10.1016/j.concog.2014.09.002

Wilkinson, S. (2014b). Levels and kinds of explanation: Lessons from neuropsychiatry. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00373

Wilkinson, S. (2015). Dennett's personal/subpersonal distinction in the light of cognitive neuropsychiatry. In C. Muñoz-Suárez & F. De Brigard (Eds.), *Content and Consciousness Revisited* (pp. 111–127). Cham: Springer. https://doi.org/10.1007/978-3-319-17374-0_6

Williams, D. (2020). Predictive coding and thought. *Synthese*, *197*(4): 1749-1775. https://doi.org/10.1007/s11229-018-1768-x

Williams, D. (2021). Epistemic irrationality in the Bayesian brain. *The British Journal for the Philosophy of Science*, *72*(4), 913–938. https://doi.org/10.1093/bjps/axz044

Yon, D., Heyes, C., & Press, C. (2020). Beliefs and desires in the predictive brain. *Nature Communications*, *11*(1), 4404. https://doi.org/10.1038/s41467-020-18332-9