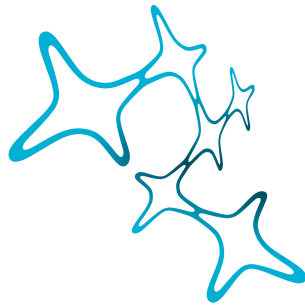

Mental states and cognitive mechanisms in predictive agents

Sofia Rappe



Graduate School of
Systemic Neurosciences

LMU Munich



Dissertation der Graduate School of Systemic Neurosciences
der Ludwig-Maximilians-Universität München

Munich, 1st June 2022

First Supervisor: Prof. Dr. Ophelia Deroy
Chair of Philosophy of Mind
Faculty of Philosophy, Philosophy of Science and the Study of Religion
Ludwig–Maximilians–Universität München

Second Supervisor: Prof. Dr. Stephan Hartmann
Chair and Head of the Munich Center for Mathematical Philosophy
Faculty of Philosophy, Philosophy of Science and the Study of Religion
Ludwig–Maximilians–Universität München

Third Supervisor: Dr. Sam Wilkinson
Senior Lecturer in Philosophy
Department of Sociology, Philosophy and Anthropology
University of Exeter

First Reviewer: Prof. Dr. Ophelia Deroy
Second Reviewer: Prof. Dr. Stephan Sellmaier

Date of Submission: 1st June 2022
Date of Defense: 17th October 2022

Summary

Over the past two decades, predictive processing accounts, which originated in the domain of perception, have developed into a framework of brain functioning that can be applied to explain various processes in perception, cognition, and action and how they relate to each other. A large body of empirical and theoretical literature that takes a predictive processing perspective has developed in different domains—from the lower-level unconscious perceptual processes to subjective experience and deliberate, intentional thought. Such extension is almost unprecedented. Cognitive science has primarily been characterized by specialized explanations. Predictive processing, on the other hand, at least according to some of its proponents, promises to unify perception, action, and cognition, bringing them under a single set of processing principles and integrating different sub-fields in the cognitive sciences.

Nevertheless, some features of our cognition and behavior still resist predictive-processing-based explanations. The cases involving such features are often used to highlight the explanatory limitations of the framework's architecture. However, upon closer inspection, many of these cases are not limited to the explanations of subconscious processing and behavioral responses but involve some conscious mental states. Explaining conscious states adds further complexity, but this complexity offers not only an extra challenge but also a way to re-construe the resistant cases. What if conscious mental states resist predictive processing explanations not because of the limits of this framework but because of the false insights we, as cognitive agents or sometimes as philosophers, draw from these very conscious states? Then, the primary challenge is to reconcile this experience with the theoretical framework of predictive processing. This reconciliation is bidirectional: sometimes it means letting go of the intuitions about our mental lives (*subjective-to-objective direction*), sometimes—augmenting the underlying framework to make room for the experiences themselves

(*objective-to-subjective direction*).¹ In this doctoral thesis, which consists of a collection of individual articles, I examine three specific cases where predictive processing explanations offered for distinct cognitive processes conflict with the first-person experiences we have when we draw on these processes. By doing so, the goal is to contribute to extending the predictive processing framework beyond perception.

The first article concerns the relationship between thought and language and addresses two challenges for predictive processing accounts when it comes to accommodating conceptual thinking—generality and rich compositionality of thought. I note that we do not have access to cognitive processes but only to their conscious manifestations and argue that compositionality is a manifest property of thought rather than a feature of the thinking process (*subjective-to-objective direction*). I also argue that surface compositionality results from the interplay of thinking as a process and language (*objective-to-subjective direction*). Both capacities, constituting parts of a complex cognitive system, can then be explained based on the architectural principles of predictive processing. Under the assumption that language and conceptual thought are distinct, I sketch out one possible way for predictive processing to accommodate both generality and rich compositionality.

The second article discusses the sense of reality characterizing perceptual experiences. Contrary to most of the literature, it argues that the sense of reality is not exhausted by monitoring the internal or external source of one's experience, which must occur in a predictive brain (*subjective-to-objective direction*). The sense of reality can also vary in cases where the source of experience is identified as external. The paper proposes that our sense of reality must then be composed both of a categorical marker, distinguishing internal and external origins of our representations, and other subjective markers, which come to modulate the subjective reality felt for perceptual scenes (*objective-to-subjective direction*). This composite account makes new predictions regarding the possible breakdowns of the sense of reality in perception.

Building on this account, the third article investigates how such breakdowns in one's sense of reality may be relevant to the predictive processing-based explanations of psychosis. It rejects the idea that bizarre predictions

¹This does not mean that predictive processing presents the objectively correct way to think about the mind and brain. Instead, "objective" refers to the cognitive mechanisms in the brain that are objective in the sense of being subpersonal and independent of the agent's perspective. "Subjective," on the contrary, refers to the conscious personal-level experiences/mental states.

manifesting in psychosis are absent in the neurotypical brains (subjective-to-objective direction) and argues that predictive processing accounts must be augmented with what I call counterfactually rich internal models (objective-to-subjective direction).

I conclude the dissertation with a general discussion of the implications of this work for philosophy of mind and cognitive sciences.

Acknowledgements

First, I would like to thank my primary thesis advisor, Ophelia Deroy, for her support, advice, and stimulating intellectual discussions throughout my Ph.D. She has stoically endured my bad early drafts and underdeveloped ideas while providing helpful suggestions and guiding me towards becoming a better philosopher, writer, and career academic. The Cognition, Values, and Behaviour (CVBE) group has offered me a unique and vibrant environment for completing my Ph.D. I am proud to be a part of this group.

I would also like to thank my second advisor, Stephan Hartmann, for his helpful feedback during our thesis advisory meetings, and my third advisor, Sam Wilkinson, for his supportive and encouraging attitude and a tremendously positive experience collaborating on a paper. I am grateful to my official faculty mentor Stephan Sellmaier for his insights, fruitful philosophical discussions, and his philosophy of mind reading group that has greatly expanded my horizons. Special thanks to Paul Taylor for giving me a glimpse of empirical neuroscience and letting a philosopher inside a lab full of expensive equipment. This experience was invaluable for gaining insight into the empirical literature from which I draw in my philosophical research.

Further, I would like to thank the Graduate School of Systemic Neurosciences (GSN) for creating a welcoming interdisciplinary community and for the financial support that made it possible for me to pursue my academic curiosity full-time. My special thanks go out to the GSN administration and, in particular, Lena Bittl, Stefanie Bosse, and Nadine Hamze, for always being there to answer my questions and help me navigate all the administrative and bureaucratic issues I encountered as a Ph.D. student.

My academic work has significantly benefited from the many inspiring discussions with my colleagues at the CVBE, the Research Center for Neurophilosophy, the Crowd Cognition Group, and the larger GSN community. I am particularly grateful (in alphabetical order) to Oriane Armand, Lucas Battich, Mark Wulff Carstensen, Sebastian Drosselmeier, Jurgis Karpus, Chris Kymn,

Slawa Loev, Louis Longin, and Justin Sulik. It has been a pleasure to have you as my colleagues.

I am also immensely grateful to my previous supervisors at the University of Toronto and the University of Edinburgh—John Vervaeke and Andy Clark—for their encouragement at the beginning of my academic journey and for the inspiration to pursue a career in philosophy.

Special thanks to my friends for being present in my life despite the time and distance apart. Finally, thanks to my partner, Michael Clark, for his loving support throughout my Ph.D., political turmoil, pandemic, long-distance relationship, and international relocations. I am fortunate to have you in my life.

This dissertation is dedicated to my parents:

To Dina and Mikhail Rappe

Contents

Summary	iii
Acknowledgements	vii
1 Introduction	1
1 The basics of predictive processing	3
2 The varieties of predictive processing	7
3 The unifying account of the mind	11
4 Mental states and cognitive mechanisms	14
5 Roadmap	17
2 Paper 1. Predictive minds can think: addressing generality and surface compositionality of thought	19
3 Paper 2. The clear and not so clear signatures of perceptual reality in the Bayesian brain	43
4 Paper 3. Counterfactual cognition and psychosis: adding complexity to predictive processing accounts	71
5 Discussion	97
Bibliography (Introduction and Discussion)	101
Eidesstattliche Versicherung / Affidavit	111
Author contributions	113

Chapter 1

Introduction

Predictive processing is an influential framework for explaining cognition, perception, and action, which proposes that the brain’s fundamental purpose is to function as a prediction machine (Clark, 2016). This machine constantly minimizes the discrepancy between its predictions about the state of the world and the incoming sensory signal—either through updating the internal generative model or through bringing the predictions to life through action, as a kind of self-fulfilling prophecy.

Over the past decade, a large body of empirical and theoretical literature that takes a predictive processing perspective has emerged in different domains—from the lower-level unconscious perceptual processes to subjective experience and deliberate, intentional thought (for a great review, see Hohwy, 2020). Such extension is almost unprecedented. Cognitive science has primarily been characterized by specialized explanations, while predictive processing, at least according to some of its proponents, could unify perception, action, and cognition, bringing them under a single set of processing principles and integrating different sub-fields in the cognitive sciences (see, e.g., Clark, 2013; Friston, 2010; Seth, 2015).¹ As Paul Thagard notes, “the value of a unified theory of thinking goes well beyond psychology, neuroscience, and other cognitive sciences” (Thagard, 2019, p. xvi). Philosophically, the mind has also been mostly theorized as a set of faculties, modules, or capacities, which may not only be *functionally* distinct (Deroy, 2015) but also, in some cases, require their own explanatory accounts. Predictive processing does not prohibit functional modularity and localization (see, e.g., Asprem, 2019) but suggests that all these modules may operate by the same computational

¹Not everyone shares such optimism. See, e.g., Litwin and Miłkowski (2020) and the discussion in Section 3 of this chapter.

principles. This proposal aligns well with a controversial but increasingly popular idea that functional localization may be (to a significant extent) a natural consequence of developmental processes rather than an inherent feature of the human brain—a “software,” not a “hardware” property (see, e.g., Dobs et al., 2021). The idea of a framework offering a single explanatory basis for various cognitive capacities as well as cognitive and social sciences, arts, and humanities, is daring and philosophically novel.

However, many researchers remain skeptical that predictive processing can bear such a load, both in terms of the kind of unification it provides (Colombo & Hartmann, 2017; Litwin & Miłkowski, 2020; Poth, 2022) but also when it comes to explaining “offline cognition” (or in some cases, offline perception²)—the processes that at least partially disregard the sensory input (and hence sensory signal-based error minimization) and depart from the immediate perceptual reality. The examples include conceptual and linguistic thinking (Williams, 2020), imagination (Jones & Wilkinson, 2020), and even certain aspects of conscious perception, such as its phenomenal unity (Block, 2018).

Some of these challenges, such as the generality and compositionality of language, are not exclusive to predictive processing but pertain to all connectionist accounts that try to explain mental phenomena using artificial neural networks (the generative hierarchical models postulated by predictive processing may be construed as such). Often, this connectionist heritage of predictive processing is blamed for the framework’s perceived failures: the architecture of predictive agents and the limitations that follow from such architecture cannot accommodate the properties of our cognition (broadly construed). Yet, the explananda themselves are often hard to pin down, especially when they involve conscious mental states. The presence of the subjective aspect, on the one hand, introduces another level of explanatory complexity. On the other, it offers a new way of construing the challenges predictive processing faces in accounting for some features of our mental life.

My stance is that, at least in some cases, the real tension lies not in the inability of predictive processing accounts to accommodate specific features of our cognition (broadly construed) but in reconciling our conscious mental states with the underlying cognitive mechanisms. This claim is not a reductionist one: I do not suggest that there is an identity relationship between our mental states (such as pain) and the brain mechanisms that are responsible for such states. Mine is a much weaker assumption common in neuroscientific

²For example, there is a long-standing debate about whether imagination is a perceptual or a cognitive process (see, e.g., Cavedon Taylor, 2021).

research, namely, that our mental lives relate to the physical processes in our brains and bodies. Hence, to a certain extent, our mental states should be explained by or, at the very least, compatible with our neurocognitive mechanisms. To simplify, the fact that we feel pain means that the internal workings of our bodies must be such as to allow us to have pain experiences. On the other hand, I am cautious about the program of *rational psychology* that aims to use mental states as sources of insight into the principles that underlie the mind and make experience possible. As I discuss in more detail in section 4 of this chapter, our mental states, their contents, and phenomenology may not always be informative and, in fact, are often misleading about the mechanisms that are responsible for them.

This simple idea is the basis of what I call *bidirectional reconciliation*. On the one hand, we should be willing to abandon our intuitions about cognition to break the limiting assumptions (*subjective-to-objective direction*). On the other hand, we should still explain with the help of the chosen cognitive framework why our conscious mental states are as they are (*objective-to-subjective direction*). In the attempt to further the case for predictive processing, I demonstrate how these directions may be realized in two specific examples—conceptual thought (Chapter 2) and the sense of reality and related symptoms in psychosis (Chapters 3 and 4).

1 The basics of predictive processing

Predictive processing (PP) (Clark, 2013, 2016; Friston, 2005; Hohwy, 2013; Rao & Ballard, 1999) is a theoretical framework of brain functioning, according to which the brain is constantly engaged in the process of minimizing the discrepancy between the predictions it generates and the incoming sensory input. Essentially, the framework can be boiled down to two key elements—constantly updated hierarchical generative models and the discrepancy (prediction error) minimization strategy called predictive coding.³

³The terms predictive processing and predictive coding are often used interchangeably. Technically speaking, predictive coding is a system-updating strategy implemented within the broader predictive processing framework.

Hierarchical Generative Model

Predictive processing accounts have roots in Helmholtz’s proposal (1866) that the content of our perception somehow reflects our brain’s causal “hypotheses” about the environment constructed based on the accumulated knowledge about how the world is. In predictive processing, this knowledge is represented in terms of priors specified by the generative model in the brain. This model consists of a hierarchy of (broadly causal) hypotheses, each specifying the causal origin of the hypothesis below. For example, in the case of vision, lower levels may represent sensory signals coming directly from the retina, while higher levels—represent the objects that may cause one to perceive the specific distribution of intensities (Figure 1).

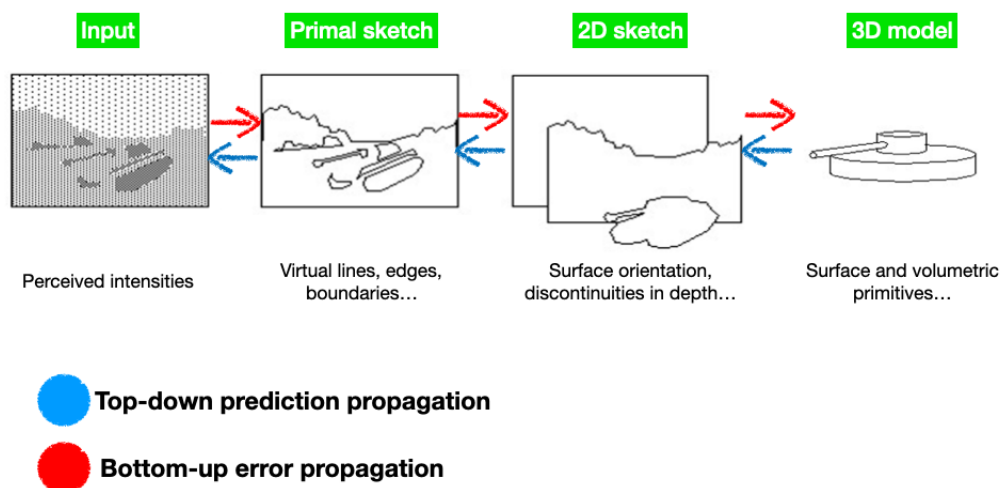


Figure 1. Hierarchical model in visual perception

The hypotheses at each level of the model are taken to be causally dependent on (or affected by) only the adjacent levels. The overall winning hypothesis(/es) associated with the highest overall probability is what ultimately defines perception and drives behavior.

Predictive Coding

The second key feature of predictive processing accounts is the strategy for updating the generative model, i.e., predictive coding. Only the unpredicted elements of an input (called residual error) are fed forward for further stages of information processing. In the classical interpretation, what is propagated

forward is not the residual error itself but the signal indicating the mere presence of the error. Andy Clark compares that to a parlor game—trying to guess something about the environment while blindfolded by asking yes/no questions (2016). Despite being minimal, the residual error signal (or its absence) is richly informative about the state of the world.

On the other hand, the downward flow delivers precision-weighted predictions that reflect the “expectation” of the system based on previous experience. The processing system operates simultaneously both in a top-down and bottom-up fashion, progressively updating predictions to minimize prediction error until the system arrives at a stable hypothesis about the state of the world (Clark, 2016). The influence of the information flow in each direction on processing depends on the situation. A highly variable, imprecise, or unimportant signal, for example, is given less weight, and vice versa. The balance of the system can shift towards a more bottom-up processing approach by assigning less precision to prediction and more to the sensory input. However, the computation that determines the degree of bottom-up influence is top-down because the generative model that encodes world knowledge also encodes the knowledge that underpins precision estimations.

Importantly, not all the contributors to the hypothesis probability estimation on the lower-level branches must align with the winning hypothesis, as long as their weights are not overwhelmingly high. The precision of the associated prediction errors determines which levels drive processing more strongly. To use Karl Friston’s analogy (see Rappe, 2019), appointing a middle manager (i.e., selectively increasing the precision of prediction errors at a certain level) increases that level’s influence on the CEO (higher levels) while at the same time making the voice of the regular employees less heard—higher levels of the model become relatively impervious to (prediction error) messages from lower levels. In many cases, occasional inconsistencies between the different levels of the hierarchy are not perceived by the individual, nor are they corrected, for example, when the system operates in a heavily top-down manner or when the input is assigned a small weight. In that sense, predictive processing accounts are satisficing ones—the brain is not striving for the most accurate model of the world but merely a good enough one. This feature is often involved in predictive processing-based explanations of visual illusions, such as the Hollow Mask illusion (Gregory, 1973). Here, the higher-level prediction of convex faces dominates the lower-level features, which results in the erroneous (but very persistent) perception of the hollow mask as a convex one (see Figure 2). Consistent with this explanation, the illusion decreases when the face is shown upside-down and increases when the presented face is shown in a familiar orientation (Hill & Johnston, 2007).

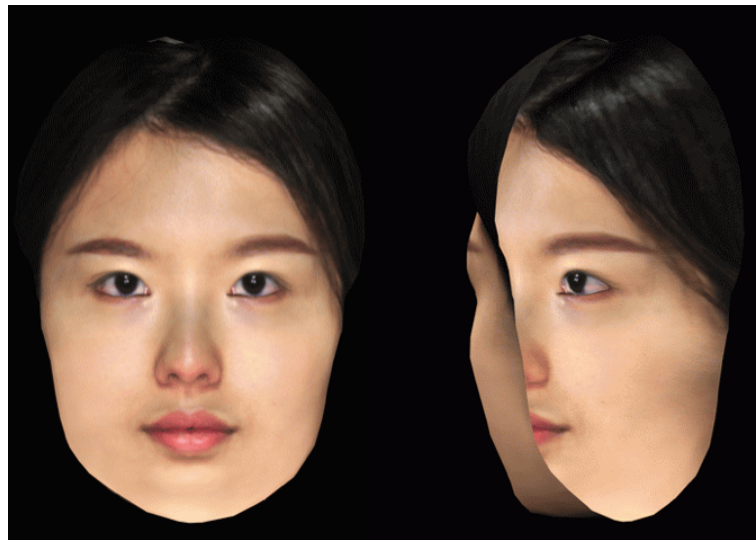


Figure 2. The Hollow Mask Illusion.

Images adapted from the GIF animation by Cmglee, 2017, Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Hollow_face_illusion.gif). CC BY-SA 4.0.

Commonly, the process of error minimization is thought to implement approximated Bayesian inference, presenting predictive processing as a Bayesian framework. This coupling is not obligatory, as the predictive-processing architecture described above does not necessitate operating according to the Bayesian principles (Atchison & Lengyel, 2017). That said, the coupling has proved to be productive. Feedforward Bayesian inference does not present a realistic computational approach to human cognition—it is slow, sometimes computationally intractable, and often requires a separate account of the origin of the relevant priors and likelihood as well as their implementations in the brain (Penny, 2012; Tenenbaum et al., 2011). Predictive processing presents a neurally plausible story of implementing Bayesian inference. Brains come to approximate the results of exact Bayesian inference as a result of generative model updating while computing prediction errors provides the brain with an internally accessible quantity to minimize.

It remains to be explained why a predictive-processing-based generative model would come to approximate Bayesian inference in the first place. As one theory, this may be a natural consequence of utilizing evidence-based updating (Bayesian world leads to Bayesian reasoning). However, the critical point is that it is not predictive processing that motivates Bayesian reasoning but the opposite. Bayesian models of cognition motivate the use of the predictive processing architecture as a realistic way of implementing the relevant statistical calculations. As Hohwy (2020, p. 211) puts it: “it is easier to conceive

of (biologically plausible) systems that minimize prediction error and thereby approximate Bayes, than systems capable of engaging in exact Bayesian inference”.

2 The varieties of predictive processing

Beyond these general principles, different predictive processing accounts are underlined by distinct philosophical views on the nature of cognition. This dissertation is not concerned with such foundational issues. Instead, it simply attempts to extend the current domain of predictive processing as a set of processing principles in ways that are maximally accommodating of a whole range of predictive processing accounts and remain as neutral as possible on the contentious philosophical debates. Yet, to specify the boundaries, it seems prudent to present the points of philosophical contention within predictive accounts.

Perceptual vs. active inference

The first debate relates to the primacy of perceptual vs. active inference. Prediction error minimization may not only matter for generating percepts in accordance with sensory input and inferring causal relationships in the world but, more importantly, for bringing about the changes that help the agent stay alive (Seth, 2015; Wiese & Metzinger, 2017). In short, the error between sensory signals and predictions of sensory signals (derived from internal estimates) can be minimized by changing internal estimates but also, in some cases, by changing sensory signals (through action). Further, the exact representations (predictions) that are active in perception can also be deployed to enable action. This combined mechanism “by which perceptual and motor systems conspire to reduce prediction error using the twin strategies of altering predictions to fit the world, and altering the world to fit the predictions” (Clark, 2016, p. 122) is often referred to as *active inference* (Friston et al., 2017). Interoceptive prediction error minimization is an illustrative example of how perception and action are coupled according to predictive processing. Interoception refers to the perception of the body’s physiological state, “a process associated with the autonomic nervous system and with the generation of subjective feeling states” (Seth et al., 2012, p. 1). Subjective emotions, such as fear, are then determined by predictions about the interoceptive state of the body. In acute distress, internal parameters like blood pressure and heartbeat rise and deviate from the normal, predicted levels. One way to minimize the

difference is to act—fight to resolve the conflict or run away (perhaps, metaphorically).

Some researchers argue that inference through action is the dominant type of inference in the human cognitive system. I do not have a stake in this debate. However, the kinds of cognitive phenomena discussed in this dissertation, such as conceptual thinking, the sense of perceptual reality, and psychosis, do not directly involve active inference, even if they engage in action-oriented prediction error minimization.

The free energy formulation

The idea that prediction error minimization is not (or at least not solely) aimed at making sense of the world but rather at keeping the organism alive is well-aligned with a broader ecological perspective: Our interactions with the world are constrained by the fact that we are living, embodied creatures. “Recent advances in statistical physics and machine learning point to a simple scheme that enables biological systems to comply with these constraints” (Friston et al., 2010, p. 227). The scheme comes down to minimizing free energy—an information-theoretical quantity that refers to a state associated with disorder or uncertainty. Biological systems have a limited range of states in which they can survive. Hence, they must maintain themselves within the “possible range of states by minimizing disorder and uncertainty” (Venter, 2021, p. 2). Predictive processing can serve as an implementation of the Free Energy Principle (FEP) since minimizing free energy can implicitly minimize surprise (Friston, 2010). However, subscribing to predictive processing does not necessarily mean subscribing to the Free Energy Principle and vice versa. The Free Energy Principle is a mathematical, normative theory (or *principles theory*), whereas predictive processing is a mechanistic processing framework (or *process theory*) (Hohwy, 2020). This dissertation is concerned with extending the mechanistic framework rather than adopting an ecological or normative stance.

Cognitivist vs. enactive predictive processing

The free energy formulation encourages an embodied and representation-light approach to predictive processing and stands in contrast to classical cognitivism (Venter, 2021). The cognitivist perspective, which is also sometimes called *conservative predictive processing* (Gładziejewski, 2016; Wiese, 2017), holds that organisms use models of their environments to act. According to this view, the mind is relatively isolated from the external world, and

inference is subservient to both perception and action. The generative model produces predictions that function as representations of what is happening in the external environment, and the representations reflect its causal structure (Gładziejewski, 2016; Kiefer & Hohwy, 2018). However, this representation-ism does not necessarily take a strong shape where “mental states contain real and causally relevant representative content” (Piekarski, 2021, p. 22) but may be purely pragmatic. “In this approach, internal representations do not have real content, but they are only postulated by some researchers who want to explain the cognitive functions of the mind” (Piekarski, 2021, pp. 21-22).

The Free Energy Principle provides a more direct, action-oriented story, which more closely aligns with the *radical predictive processing*. Organisms do not construct models of their environment. Instead, in a sense, they **are** such models. The proponents of radical predictive processing treat “perceptual and active inference are intricately linked rather than one being in the service of the other” (Venter, 2021, p. 3), and the representations are meant to “engage the world, rather than to depict it in some action-neutral fashion” (Clark, 2016, p. 291). Radical predictive processing is much more embodied and is inspired by ecological psychologists such as Gibson (1966, 1979) and later enactivists such as Francisco Varela and Evan Thompson (Varela, Maturana & Uribe, 1974; Varela, Thomson & Rosch, 1992).

Importantly, conservative predictive processing does not deny the importance of the environment. Both radical and conservative predictive processing take the generative model to “recapitulate the causal/statistical structure of the environment” (Piekarski, 2021). The two approaches simply provide different stories regarding how the environment is implicated in cognition and how the mind gets a grip on the world.

My approach is pragmatically representational when it comes to higher-order cognition. Whether all predictive processing-based processes require representations is a debate well beyond the scope of this dissertation.

The extent of cognitive penetrability

As Hohwy puts it, “in any PP scheme, there must be integration of top-down predictions and bottom-up prediction errors” (Hohwy, 2020, p. 216). This feature raises the question about the extent of cognitive penetrability of perception, that is, how much our conscious and unconscious mental states influence the content of perceptual experience.

The standard strategy in PP remains to treat cognitive and perceptual hypotheses as forming a common generative hierarchy, in which the hypotheses

related to higher-order cognitive processes are situated higher than the perceptual ones. This approach is also known as *Cognition-on-Top* (more on this in Chapter 2). As a result, the standard strategy mostly deflects the problem of cognitive penetrability by renouncing the cognition-perception divide altogether (see, e.g., Fletcher & Frith, 2009; Hohwy, 2013).

Leaving aside the question whether meaningful differentiation between cognition and perception can be made in predictive processing, a single hierarchical model for perceptual and cognitive hypotheses (traditionally understood) does not necessarily imply extensive mutual influence (see, e.g., Deroy 2019; Macpherson, 2017). On the one hand, according to predictive processing, perceptual inference results from a combination of the information provided by the current sensory input and prior expectations. For this reason, as long as the prior is assigned a non-zero weighting, there is always a degree of cognitive penetrability in perception. Yet, predictive processing does not imply that all beliefs affect perception “since perceptual inference is subject to precision-weighting extracted in prior learning as well as model selection and complexity considerations” (Hohwy, 2020, p. 216). In fact, strong cases of cognitive penetrability are expected to be rare, as this would require a combination of “the context, the prior learning, and the precision of the sensory input” to result in persistent uncertainty where “there is no action for epistemic value” (Hohwy, 2020, p. 216).

Another essential factor that could affect the degree of cognitive penetrability is the modularity of the cognitive system. At the minimum, the generative network itself imposes some connectivity constraints since it reflects the inferred causal structure of the world and the degree to which higher-level properties are present in perceptual experience is highly debatable (see, e.g., Siegel, 2016). That said, a common assumption in the predictive processing accounts is that different parts of the generative model are deeply interconnected and integrated. In my dissertation, I do not make any specific assumptions about cognitive penetrability or the limits of modularity of our cognitive capacities (beyond that both, to some extent, exist).

Scope and applicability

Another important debate relates to the scope of applicability of the predictive processing framework. On the one hand, there is a pluralistic approach, where various predictive-processing-based processes may work alongside the non-predictive processes. Here, “different domains of perceptual and cognitive processing call for potentially quite disparate PP solutions [on the level of implementation], tailored to the particular perceptual or cognitive task at

hand... Systems learn, develop, or evolve to exploit the benefits of predictive processing, and it is a matter of discovery where PP occurs” (Hohwy, 2020, p. 213). In other approaches, such as the Free Energy Principle, free energy minimization is treated as a single principle underlying all cognitive processes. Therefore, predictive processing takes a more unifying and reductive shape. As Hohwy notes, adopting the reductive perspective may not aid understanding the phenomena explained through the predictive-processing mechanisms (and may put some people off). However, “there is significant discussion of the explanatory benefits from theoretical unification, and there may be foundational issues that are illuminated by FEP” (Hohwy, 2020, p. 213).

In this dissertation, I take the pluralistic approach but set my course to “maximal” predictive processing (Sims, 2017). Not because I necessarily believe that all our mental life is predictive processing-based, but rather to see how far I can take the framework in its most minimal, the least assuming form.

3 The unifying account of the mind

Not everyone shares the optimism regarding predictive processing as a unifying framework of the mind. Most criticisms fall into three broad, interrelated categories: (i) concerns related to the framework’s explanatory power and general utility, (ii) concerns about the lack of solid empirical evidence, and (iii) concerns regarding the architectural limitations of PP when it comes to accommodating certain specific features of our cognition.

General utility and explanatory power

The first concern is that predictive processing is too accommodating and that unification in predictive processing is achieved through shoe-boxing fundamentally different phenomena together, straining the meanings of terms like “prediction” beyond their limits (Litwin & Miłkowski, 2020). Predictive processing, in its most general form, presents a set of high-level computational principles. Although these principles aim to describe the neurocomputational machinery, they may plausibly be realized in a variety of ways that also come with different assumptions about the fundamental aspect of cognition, such as, for example, the degree of involvement and character of mental representations (Venter, 2021, also see the discussion above). Further, the language of prediction and prediction-error minimization may be vague enough to obscure the actual neural mechanisms of the phenomena under investigation. Meaningful differentiation may be more helpful in getting at

the actual processes. It is simply not sufficient to describe “phenomenological data in the terminology of the theoretical framework” (Kogo & Trengove, 2015). In addition, as Matteo Colombo and Stephan Hartmann (2017) noted concerning Bayesian frameworks more broadly, unifying power is often associated with explanatory power. In practice, however, computational-level frameworks (such as PP) only provide rules or principles of operation but do not reveal aspects of the specific mechanisms at play. In that sense, they have limited explanatory value of (broadly) constraining the search space of potentially applicable mechanisms.

Evidence and testability

Accommodativeness of predictive processing also presents a problem with the testability/falsifiability of the framework. So far, when it comes to empirical evidence, some studies show results compatible with predictive processing (see Walsh et al., 2020 for a review), and there is no clear-cut counterevidence. Yet, this is not much of a win for the framework, since the general principles of predictive processing can accommodate contradicting evidence and opposite predictions regarding, for example, the impact of experimental manipulation on expectation vs. error unit activity and, consequently, the global neural response (Kogo & Trengove, 2015; Walsh et al., 2020). This issue arises due to the same very fact that predictive processing merely provides broad computational principles. On the one side, this is a virtue for integrating explanations on multiple levels—from synapse to processing hierarchy. On the other, “it can also pose a challenge as researchers seeking to test the theory’s validity confront the often-murky translation from algorithm to biophysical implementation” (Walsh et al., 2020, p. 261). However, when it comes to testing the evidence for the more detailed predictive-processing models describing biophysical implementations, the lack of evidence partly reflects the methodological challenges in testing the unique predictions of these models (Walsh et al., 2020). Hopefully, as methodology develops and specific models within the framework mature, a combination of neurophysiology and computational modeling methods will provide more decisive evidence.

Both kinds of concerns about unification and empirical evidence may be eased if one treats predictive processing as a way of thinking about diverse factors and forces that respects their diversity while revealing how they manage to work together in a single cognitive economy. This economy operates through the common elements of predictions, prediction errors, and precision estimations, which provide not merely broad restrictions on the possible mechanisms but also units of the functional structure (although perhaps in a

minimal way). The functional structure may be realized in multiple ways and through multiple subsystems while committing to the single computational principle. The good part is that the functional description provided by predictive processing (at the algorithmic level) must in some way map onto neural-level implementations, and many predictive-processing-based accounts of individual phenomena are already quite specific in that respect. In this sense, although predictive processing presents a cognitive framework and not a falsifiable empirical theory, what matters is that it “enable[s] progressive research programs” (Wiese, 2018, p. 229). The specific implementations of predictive processing are not only individually falsifiable (Wiese, 2018) but also, taken together, put additional requirements on the architecture of the system as they must be consistent and compatible with each other. This condition allows to ‘pin down’ predictive processing with enough local implementations.

Architectural features

In this dissertation, I focus on another type of concern. This type relates to the limitations and implications of predictive processing accounts when explaining certain features of our cognition (broadly construed) and behavior. One popular example is our intense sense of curiosity that seems to clash with the goal of minimizing long-term prediction errors. The Dark Room Problem is the most well-known formulation of this challenge (see, e.g., Sims, 2017). If the goal is to minimize prediction error (and hence the unpredicted sensory input), a great way to do that is to seek out a dark room with not much going on and stay there for as long as possible. Yet, this goes against much of what we know about human behavior. Multiple solutions to the Dark Room Problem have already been proposed. For example, some argue that we seek out good learning situations for evolutionary reasons (Friston, Thornton & Clark, 2012; Kiverstein, Miller & Rietveld, 2017; Oudeyer & Smith, 2016), some argue that the answer could be found in interoceptive predictions (Barrett & Simmons, 2015; Seth, 2013), yet others claim that the key is to consider the optimal rate of prediction error minimization (Van de Cruys, 2017).

Other examples relate less to our behavioral patterns and more to innate cognitive abilities. For instance, recently, Williams (2020) argued that conceptual and linguistic reasoning presents a significant challenge for predictive processing due to its compositionality and generality (see Chapter 2). This taps into a larger array of concerns related to the ability of predictive processing to accommodate the kinds of cognitive processes that require “deliberate imaginative exploration of our own mental space” (Clark, 2016), that is at least partially conscious, intentional, offline (in the sense of diverging

from the current sensory input) cognition. Some of these problems are inherited by predictive processing from the earlier versions of connectionism. For example, Hoerl and McCormack (2019) argue that **any** connectionist architecture on its own may have problems with *temporal reasoning*—tasks that require representing specific times and temporal order and locating events in the past, present, or future. Such tasks are hard to accommodate within connectionist architectures because these architectures do not explicitly represent change but rather update representations as new information comes along, significantly limiting the kinds of temporal cognition the cognitive system can support (more on this in Chapter 4).

If the ambition is to generalize predictive processing beyond online perception, at least some of these seemingly problematic cases must be addressed, if not to establish maximal predictive processing, then to signpost the framework's limits more accurately.

4 Mental states and cognitive mechanisms

The “deliberate imaginative exploration of our own mental space” discussed in the previous section (Clark, 2016, p. 273) often involves first-person mental states. We experience thoughts that present themselves as inner speech, imagery that appears when asked to imagine an apple, or a feeling of being immersed in the perceptual reality rather than merely watching it unfold around us (with some notable exceptions).

The nature of the relationship between these first-person mental states and the associated brain mechanisms is a longstanding debate in metaphysics, the philosophy of mind, and neuroscience. An in-depth discussion about this relationship is far beyond this project's scope. As a computational framework, predictive processing is compatible with different stances, including those postulating an identity relationship between mental and physical states. My view is that mental states are unlikely to be mapped in the brain directly and unambiguously (e.g., via neuronal activation strengths or activity localization).

Nevertheless, neuroscience as a scientific endeavor largely shares the assumption that mental life relates to the physical processes in our brain and body in some way. Hence, if a framework of cognitive functioning has strong unification aspirations, it should be able to explain the content and specific phenomenology of such experiences (or at least theoretically allow for such experiences to arise). Importantly, this is not a request to solve the Hard Problem of Consciousness (Viola, 2017) or **reduce** the phenomenology of our sub-

jective experiences to mechanistic explanations. Instead, the task is to explain how and why the machinery we propose gives rise to (or at least is compatible with) the kinds of conscious experiences we have. In his recent book, Anil Seth calls this problem the Real Problem of Consciousness (Seth, 2021).

Recently, predictive processing has increasingly been used to explain mental states and their phenomenology, including in pathological cases such as psychosis (see Chapter 4 of this dissertation for an extended discussion), autism spectrum disorders (Lawson et al., 2014; Palmer et al., 2017; Pellicano & Burr, 2012; Van de Cruys et al., 2014), PTSD (Wilkinson, Dodgson & Meares, 2017), and depression (Badcock et al., 2017; Stephan et al., 2016), as well as consciousnesses and qualia (see, e.g., Clark, 2019; Clark, Friston & Wilkinson, 2019; Deane, 2021; Dolega & Dewhurst, 2021; Kirchhoff & Kiverstein, 2019; Seth, 2021; Solms & Friston, 2018; Williford et al., 2018). The goal of my dissertation is to further advance such explanations for our experience of linguistic thought, the sense of reality, and psychosis.

However, subjective mental states and mental phenomenology are not just something to be explained but also sometimes the source on which we draw to make sense of our cognition. This is the basis of the program of *rational psychology* that uses mental states as sources of insight to deductively arrive at the principles that underlie the mind and make experience possible. However, drawing on mental states to learn about cognitive mechanisms must be done with great caution and may lead to problems such as, for example, the *problem of cognitive ontologies* (Janssen, Klein & Slors, 2017; Klein, 2012; Price & Friston 2005). The problem of cognitive ontologies is often formulated in terms of the tension between behavioral, task-based studies designed based on the scientifically digested psychological/intuitive descriptions of our mental life and mapping these to neuroimaging. The labels or categories of mental states and processes that we create to describe our cognition based on our subjective experiences often cannot be mapped to individual functions and structures in the brain. For example, it does not seem plausible that we might find wishes or blue perceptions in the brain, or even consistently and reliably identifiable patterns of activation, etc., during pain experiences or seeing blue. The same may be said not just about individual experiences but also processes like “thinking” or “imagining,” which may be nothing more than convenient umbrella terms for various distinct cognitive processes that should be investigated separately. In other words, the mind and brain may be independently “carved up” in ways that provide independent explanations at their respective levels (e.g., psychology and neuroscience) but do not allow for successful one-to-one mapping. As Viola (2017) notes, when such discrepancy occurs, “neuroscience is deemed a legitimate arbiter for refining cognitive ontology, i.e., for

choosing the right set of mental entities” (Viola, 2017, p. 164). Similarly, a cognitive ontology may be adjusted to fit the mechanistic elements of a specific theory of cognitive functioning, assuming that the approach is promising and can accommodate the same behavioral outputs through the new, adjusted ontology: “sophisticated mechanistic frameworks involve both a decomposition and a recomposition of phenomena, as well as a detailed characterization of the overall context” (Bechtel, 2009 in Viola, 2017, p. 164).

Another related mapping problem arises when the features of mental states are ascribed to the information-processing strategies that give rise to these states. For example, as I argue in Chapter 2, the way thought appears in our conscious experience, as if we had an inner speech, may have played a role in adopting Language-of-Thought-like approaches to thinking (Fodor, 2008). However, language-like is just how thought appears to us. Such experience does not correspond to the whole of our thinking activity (Heavey & Hurlburt, 2008) and even less reflects the actual mechanisms of thinking (Machery, 2005; Wilkinson & Fernyhough, 2018). Moreover, the language-like appearance of thought may result from the interaction between **multiple** kinds of underlying processes. What cognitive processes come together to produce a single mental phenomenon and what kind of features these processes possess individually are questions that need to be untangled case by case, both philosophically and scientifically.

Notably, the points above do not imply that mental states are uninformative about aspects of cognition. Some states especially seem to serve an explicit cognitive monitoring role. One example discussed in Chapter 3 is the sense of reality that signals to the agent the status of their experience (“am I perceiving or imagining?”). Nevertheless, even in such a case, the experience reflects the **product** of cognitive processes (“yes, I am perceiving”) rather than the **processes** by which the cognitive system arrives at such a product.

It is crucial then to differentiate between the two kinds of explananda. Are we trying to explain how the cognitive processes (as presented by a particular framework of brain functioning) account for the conscious mental states that we experience? Or are we trying to explain the features of the underlying mechanism, an “objective” cognitive capacity itself? The difference lies in how we should treat the intuitions and insights allowed by the first-person perspective. In the former case, the first-person perceptive and experience are in the focus. In contrast, in the latter case, the agent’s perspective may be irrelevant (or worse yet, misleading). Hence, reconciling mental states with cognitive mechanisms may involve one of the two opposite strategies. The first one is to examine carefully and, in some cases, strip away the intuitions about cognition afforded by our conscious mental lives to explain our cognit-

ive capacities without the first-person bias (*subjective-to-objective direction*). The second one is to focus on the mental states and how they can arise from the chosen framework of cognitive functioning. Sometimes, making room for such states may require augmenting or adjusting the framework (*objective-to-subjective direction*).

What follows is my attempt to further the case for predictive processing by applying the above strategies to two specific cases—conceptual thinking and reality monitoring.

5 Roadmap

Chapter 2 (Paper 1) of this dissertation demonstrates the bidirectional reconciliation approach on the example of language and conceptual thinking in predictive processing. Specifically, I tackle the generality and rich compositionality of language — two properties often presented as challenges for connectionism. I argue that compositionality is a surface, manifest property of language and not a property of the thinking process (*subjective-to-objective direction*). I then show how the interaction between perceptuo-conceptual and linguistic hierarchies operating according to predictive processing principles may account for both the generality and the surface compositionality of linguistically expressed thought (*objective-to-subjective direction*).

Chapter 3 (Paper 2) of this dissertation investigates the relationship between the mechanism of reality monitoring and the sense of reality. Often, reality monitoring is understood as a capacity to discriminate whether an experience is perceptual or imagined. The chapter argues that the sense of reality is not exhausted by monitoring the internal or external source of one's experience, which must occur in a predictive brain (*subjective-to-objective direction*). On the one hand, the sense of reality may often be confused even in cases where the source of experience is clearly identified as external, for example, in virtual reality (VR) experiences and derealization. On the other, some experiences not grounded in sensory information may feel very real (e.g., hallucinations). The proposal is that such perceptual confusions require our subjective sense of reality to be a composite of several subjective markers (not limited to a categorical one that identifies an experience as perceptual and connects us to reality) (*objective-to-subjective direction*). This composite account makes new predictions regarding robustness, non-linear development, and the possible breakdowns of the sense of reality in perception.

Chapter 4 (Paper 3) investigates how such breakdowns in reality monitoring and the sense of reality may be relevant to the predictive processing-based explanations of psychosis. Most current approaches come down to the idea of an “atypical” brain generating inaccurate hypotheses that the “typical” brain does not generate, either due to a systematic top-down processing bias or more general precision weighting breakdown. Although strong at explaining common individual symptoms of psychosis, such approaches face some issues when we look at a more general clinical picture. The chapter rejects the assumption that bizarre predictions manifesting in psychosis are absent in the neurotypical brains (subjective-to-objective direction). Instead, what is going on in psychosis is an inability to correctly identify counterfactual hypotheses (constantly generated by a healthy brain), in some cases due to a failure of reality monitoring. This updated view requires predictive processing accounts to be augmented with what I call counterfactually rich internal models (objective-to-subjective direction). The proposal further expands the space of potential cognitive mechanisms involved in different cases of psychosis, casts “accurate” cognition as more fragile and delicate, but also closes the gap between psychosis and typical cognition.

Chapter 5 briefly discusses some future directions and the implications of this work for philosophy of mind and cognitive sciences.

Chapter 2

Paper 1. Predictive minds can think: addressing generality and surface compositionality of thought

This paper has been published under open access (Creative Commons Attribution 4.0 International License) as:

Rappe, S.(2022). Predictive minds can think: addressing generality and surface compositionality of thought. *Synthese*, 200, 13.
<https://doi.org/10.1007/s11229-022-03502-7>.

Author contributions: S.R. is the sole author of the manuscript.



Predictive minds can think: addressing generality and surface compositionality of thought

Sofia Rappe^{1,2} 

Received: 28 August 2020 / Accepted: 10 November 2021
© The Author(s) 2022

Abstract

Predictive processing framework (PP) has found wide applications in cognitive science and philosophy. It is an attractive candidate for a unified account of the mind in which perception, action, and cognition fit together in a single model. However, PP cannot claim this role if it fails to accommodate an essential part of cognition—conceptual thought. Recently, Williams (Synthese 1–27, 2018) argued that PP struggles to address at least two of thought’s core properties—generality and rich compositionality. In this paper, I show that neither necessarily presents a problem for PP. In particular, I argue that because we do not have access to cognitive processes but only to their conscious manifestations, compositionality may be a manifest property of thought, rather than a feature of the thinking process, and result from the interplay of thinking and language. Pace Williams, both of these capacities, constituting parts of a complex and multifarious cognitive system, may be fully based on the architectural principles of PP. Under the assumption that language presents a subsystem separate from conceptual thought, I sketch out one possible way for PP to accommodate both generality and rich compositionality.

Keywords Predictive processing · Conceptual thought · Compositionality · Generality · Unification

1 Introduction

The predictive processing framework¹ (or PP for short, see Clark, 2013; Hohwy, 2013; Friston, 2005; Rao & Ballard, 1999) successfully accounts for a wide variety of

Sofia Rappe
Sofia.Rappe@campus.lmu.de

¹ Faculty of Philosophy, Ludwig-Maximilians-Universität München, Munich, Germany

² Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Munich, Germany

¹ For an accessible introduction on predictive processing see, for example, Wiese and Metzinger (2018). For a detailed recent summary of PP literature, see Hohwy (2020), and particularly the Supplementary

perceptual and cognitive processes in multiple domains. Vision (Hohwy et al., 2008), body-awareness (Palmer et al., 2015), language and communication (Friston & Frith, 2015; Rappe, 2019), emotion (Miller & Clark, 2018; Seth, 2013; Velasco & Loew, 2020), and psychiatric disorders² have all received explanations that appeal to the basic PP architectural principles such as hierarchical generative models, long term prediction error minimization, and precision weightings. This extension is unusual. Cognitive science so far has been characterized by specialized explanations, while predictive processing promises to unify perception, action, and cognition, fitting them into a single model (Clark, 2013; Seth, 2015). A unified framework could offer a single coherent system of how the mind/brain functions and a better integration of sub-fields in the cognitive sciences. As Paul Thagard notes, “the value of a unified theory of thinking goes well beyond psychology, neuroscience, and other cognitive sciences” (Thagard, 2019, p. xvi). Philosophically, the mind has also been mostly theorized as a set of faculties, modules, or capacities all of which require their own account. The idea of a “one size fits all” framework offering a single explanatory basis for cognitive and social sciences, arts, and humanities is daring and philosophically novel. However, it remains controversial whether PP can accommodate an essential part of cognition—conceptual thought. There are other reasons to object to the unifying power of PP (see e.g. Colombo & Hartmann, 2017), but conceptual thought presents one of the biggest challenges: If it cannot be explained as a predictive process, PP cannot pretend to provide an exhaustive account of the mind (Seth, 2015), certainly not for philosophers.

Several have already debated how PP can account for core characteristics of cognition such as consciousness and qualia (Clark, 2019; Clark et al., 2019; Dołęga & Dewhurst, 2020; Hohwy, 2012) but less is done on conceptual thinking. As Daniel Williams (2018) points out, the *standard strategy* in PP remains to treat cognitive and perceptual hypotheses as forming a common generative hierarchy, with the hypotheses related to higher-order cognitive processes situated “higher up” (*cognition-on-top* approach) or more centrally (if the hierarchy is presented as a net, rather than a ladder). As a result, the standard strategy mostly deflects the problem of explaining thought in PP by renouncing the cognition-perception divide (Fletcher & Frith, 2009; Hohwy, 2013, but see Deroy, 2019).³

However, the standard strategy may be a misnomer here, as the idea that there is something special about conceptual thought remains much more standard in philosophy as well as the cognitive sciences. Williams (2018) suggests that thinking has two core properties that resist the assimilation to perception: it is *general* and *richly compositional*. Generality refers to the ability to flexibly reason about phenomena

Footnote 1 continued

Table 1 where he provides a representative list of recent philosophically oriented work in PP. Clark (2015) and Hohwy (2013) present two good resources for importantly distinct detailed treatments of PP.

² For PP-based accounts of hallucinations and schizophrenia see, e.g., Adams et al. (2013), Horga et al. (2014), and Fletcher and Frith (2009). For autism spectrum disorders, see Van de Cruys et al. (2014), Pellicano and Burr (2012), and Lawson, Rees and Friston (2014). For depression and fatigue, see Stephan et al. (2016).

³ One difference still remains: cognition requires a mechanism of decoupling from the immediate environment and is afforded by *offline simulation*, while perception is more tightly coupled with the sensory input.

at any level of spatial and temporal scale and abstraction. Compositionality, on the other hand, refers to the ability to combine concepts into structured thoughts (ensuring that the expressive power of thought matches that of at least first-order logic). According to Williams, the architectural commitments of PP (which include an interconnected perceptual-conceptual generative hierarchy, conditional independence of its levels, and probability-based relationships between them) preclude the framework from ever fully accommodating these properties of human thought. If the standard strategy breaks down, Williams argues, the proponents of predictive accounts of the mind must either accept that PP only applies to **some** cognitive processes or propose how to explain compositionality and generality of thought outside the internal PP apparatus, by appealing to language as a public symbolic system.

But are such concessions warranted or inevitable? Could PP, with its core internal architectural commitments, accommodate the properties of generality and compositionality which, according to Williams, sign its limits? In this paper I propose one solution.⁴ I treat linguistic and conceptual representations as distinct (Sect. 2) and argue that compositionality may be a *surface* property resulting from an interplay of thought mechanisms, conscious access, and linguistic machinery, rather than a property of the thinking process itself (Sect. 3). I then sketch out a PP picture of conceptual thought that accommodates both generality (Sect. 4) and surface compositionality (Sect. 5). Thought and language upon my proposal are two parts of a complex and multifarious cognitive system that are **fully** supported by a PP-type architecture. The paper ends with a brief discussion of the implications of the approach and possible future directions (Sect. 6).

2 Cutting in the thinking bundle

A necessary step is to agree first on what thought is, for it to be a challenge for PP. As noted by Williams (2018) and others (see, e.g., Fodor, 2008; Harman, 2015; Kahneman, 2011), *thought* is an umbrella term. It applies to processes such as reasoning, planning, deliberating, and reflecting, which seem to have different properties, functional principles, and even goals (beyond the very general ones, such as survival of the organism). The commonality between these processes is that they all appear to be conceptual, that is, require the ability to form and manipulate concepts. With concepts, at least following a popular view, come two core properties of thought. Conceptual thought is general as “we can think and flexibly reason about phenomena at any level of spatial and temporal scale and abstraction” (Williams, 2018, p. 1). It is also richly compositional, as “concepts productively combine to yield our thoughts” in a specific way (Williams, 2018, p. 1).

⁴ Recently, in his doctoral dissertation Alex Kiefer (2019) proposed a more general defense of connectionist architectures against the critiques related to the productivity and systematicity challenge. Furthermore, his account relies on the distinction very similar to the distinction between functional and concatenative compositionality discussed in this paper (Sect. 3). Both of us argue that functional compositionality is sufficient for systematicity. Yet, Kiefer’s solution lies in the properties of the representations in the vector space semantics, while I focus on the properties of the information processing in the predictive architectures.

Under such a treatment, conceptual thought resembles language: Thought is a kind of linguistic proposition, and concepts are directly associated with their linguistic labels. The process of producing thought consists then in combining linguistic labels-concepts that reside at different levels of the representational hierarchy (in PP terms). However, recent evidence from linguistics⁵ and clinical neuroscience suggests that “many aspects of thought engage distinct brain regions from, and do not depend on, language” (Fedorenko & Varley, 2016, p. 132). For example, visual thinking commonly used in mathematical proofs (see e.g. Nelsen, 1993), while being conceptual, does not seem to involve language (Tversky, 2019). Further, some non-linguistic animals, such as cephalopods, primates, and rodents, are able to perform a range of cognitive tasks typically associated with concept-formation.⁶ Of course, the same cognitive tasks may be accomplished differently in humans and non-linguistic animals. The findings, however, do suggest that some higher-order cognitive tasks associated with conceptual thought do not, **in principle**, require capacity for language.

If being conceptual and being linguistic are not one and the same, we should avoid misattributing properties from one domain to the other. The way thought *appears* in our conscious experience, as if we were having an inner speech, may have played a role in the adoption of Language-of-Thought-like approaches to thinking (Fodor, 2008). However, such experience does not correspond to the whole of our thinking activity (Heavey & Hurlburt, 2008) and even less reflect the actual mechanisms of thinking (Machery, 2005; Wilkinson & Fernyhough, 2018). At the very best, we can expect that, when it occurs, our conscious experience of inner thought aligns with the outcome of subconscious thinking processing, but even that is not guaranteed. Blindsight and visual agnosia are good examples in which one can observe the mismatch between the outcomes of *perceptual processing* and *conscious perception*. In blindsight, a person has no awareness of the stimulus but is able to act on it: they are not conscious of a mail slot but can put a letter into it. When it comes to reasoning, a similar dissociation between *decision-making process* and our *experience of it* can be found in phenomena such as choice blindness. Choice blindness shows that, under certain circumstances, people attribute decisions they have not made to themselves: when presented with the opposite of their questionnaire responses, they defend the views they said to have disagreed with. In other words, our experiences of perceptual and cognitive processes do not correspond to the processes themselves.

⁵ As a matter of fact, there is a long history of language theorists arguing for separation between conceptual and linguistic representations. See, for example, *Some B-theorists* in Levinson (1997, p. 14).

⁶ For example, Richter and colleagues (Richter et al., 2016) have demonstrated that octopuses are able to solve simple spatial puzzles that require a combination of motor actions that cannot be understood by a simple learning rule alone, while Lauren Hvorecny and colleagues (Hvorecny et al., 2007) have shown that some octopuses and cuttlefishes not just spatially represent but also conditionally discriminate. There are also studies demonstrating future-oriented tool use and at least a minimal degree of planning in octopuses (Finn et al., 2009). Even more remarkably, Fiorito and Scotto (1992) report a case of domain-general observational learning (although this finding has not yet been replicated). Similarly, rodents and primates have shown the ability to successfully perform a variety of higher-order cognitive tasks. Call and Tomasello (2011), for example, review a large body of experimental work on chimpanzees, concluding that they understand the goals and intentions of others while simultaneously lacking capacity for language-like representation, at least beyond trivial compositionality that does not require semantic operations (Zuberbühler, 2020).

If thought and language, and thinking processes and their manifestations, are distinct, we need to consider where the real challenge for PP lies. If we try to explain thinking processes, the intuitions provided by our conscious, phenomenological experience of language-like thought are not good guides. On the other hand, if we want to explain this conscious experience of thought, we have to acknowledge that it may arise from an interplay of a few separate processes. Pace Williams (2018), I argue that compositionality is a manifest property and is better explained as resulting from an interplay of two distinct cognitive processes—conceptual thought and language.

3 Thinking processes do not require procedural compositionality

Starting with the hypothesis that we do not have direct access to cognitive processing, we can no longer use our conscious perception of thought to determine whether concepts indeed combine to yield thoughts (compositionality as a process) or whether it only appears that thoughts are combined from concepts in a compositional manner (*surface* or *manifest* compositionality). While PP may indeed face problems with compositionality as a *process of conceptual combination*, it may be able to explain surface compositionality as a result of interaction between thought and language, both of which may themselves be PP-based, while retaining the expressive power that seems to be necessary for thought and is typically associated with the conceptual combination as a process. The main challenge then consists in showing that compositionality may indeed plausibly be only a surface property.

The claim that one needs to argue for is that thoughts, experienced or articulated linguistically, exhibit a form of compositionality that does not necessarily perfectly correspond to the compositionality of the thinking process. But which compositionality is at stake? For the most part, natural language possesses *concatenative compositionality*. A composite exhibits concatenative compositionality if its constituents are its spatial or temporal parts. For instance, as is the case in written and spoken language respectively (García-Carpintero, 1996). Thought, on the other hand, may be only *functionally* compositional, that is merely require that the composite expression has proper constituents (without imposing additional spatial or temporal requirements on them). An example of such compositionality is complex tones composed of simple sine wave components, or partials. Such tones can be uniquely separated into partials both mathematically and physically (by Fourier analysis and spectrum analyzer respectively) and explained in terms of their simple wave components, despite not being temporal or spatial parts of the complex tone (García-Carpintero, 1996). Importantly, García-Carpintero (1996) argues, systematicity and productivity of thought can be supported by functional compositionality alone as the ability to form relational structures or concepts does not depend on the sequential order of information processing. As Kiefer (2019, p. 234) notes, “there is no a priori requirement that in order to represent objects and properties, a representational system must possess separable syntactic constituents that have them as their semantic values”. For example, the system may represent relations not as individual nodes directly corresponding to the linguistic labels that we use to describe these relations, but instead as a variety of the corresponding states of

affairs. Such states, however, may be further collectively generalized to yield the relevant abstract concepts and associated labels for the purpose of linguistic expression (more on the nature of concepts in PP in Sect. 4.3 and on the interaction between thought and language in Sect. 5.1). Although this type of representation may be rather “unwieldy and redundant”, it is not in principle impossible (Kiefer, 2019, p. 234).

Because our conscious experience of thought is often mediated by natural language (Frankish, 2018), and natural language syntax and semantics are, for the most part, concatenatively compositional, thought inherits the appearance of concatenative compositionality from language. However, concatenative compositionality of language is a property of conscious structured thoughts and does not necessarily reflect how thought and language are processed in the brain. It is tempting and intuitive to link concatenative compositionality to semantic *atomism*, or the idea of progressive bottom-up conceptual or linguistic combination as a way of constructing higher-order chunks of language or meaning. However, as noted by many philosophers of language, including Gottlob Frege (1879), semantic atomism quickly runs into problems. To begin with, certain linguistic concepts do not have meaning in isolation, and hence, trying to interpret them bottom up is pointless. Further, the ability to substitute parts of a sentence, which is often thought of as equivalent with bottom-up compositionality (Hodge 2001; Janssen, 2001), does not hold generally (Frege, 1892). One notable exception to the substitution property, for example, is subordinate clauses, where an expression can only be substituted with an expression with the same customary sense—the way the expression **presents** the referent (as opposed to its truth value). Semantic atomism does not sit too well with novel metaphors either. In metaphors, the meaning of an expression is not equal to the compositional combination of lower-level chunks or concepts, so it is hard to explain how such meaning arises through bottom-up conceptual combination (at least without significantly complicating the story).

An alternative approach, *top-down contextuality*, suggests that rather than starting with concepts and putting them together to form coherent thoughts and judgements, one obtains meanings of the parts of an expression by *decomposing* the thought (Hermes et al., 1981). The point of departure is a complete thought, and the idea is understood before individual words are recognized (Janssen, 2001). This form of *top-down contextuality* may be too strong as parts of the sentence obviously contribute to the expression of the sense of the sentence (Gabriel et al., 1976). However, the word “contribution” suggests that the meaning of a compound sentence is, perhaps, gestalt-like (more than the sum of the meaning of its parts), as opposed to purely compositional (all elements of the meaning are contained in the lexical items and their syntactic relations within the sentence).

This weak version of compositionality is gaining popularity due to the new findings in cognitive linguistics and the recent focus of semanticists on figurative speech and non-literal meanings. PP, with its simultaneous bottom-up and top-down processing style, is in a great position to accommodate these insights, especially under the idea of separation of conceptual thought and language. In fact, as pointed out by one of the reviewers of this manuscript, the rejection of semantic atomism is explicitly argued for in defense of PP-based semantics by Kiefer and Hohwy (2018) and is generally assumed in widely popular vector space semantic models (Kiefer, 2019).

Before outlining my approach to compositionality in PP, however, let us tackle the generality challenge.

4 The generality problem

4.1 Personal level beliefs can still be about small objects and rapidly changing regularities

Generality targets “the fact that we can think and reason about phenomena at *any* level of spatial and temporal scale and abstraction” (Williams, 2018, p. 2) in a way that flexibly combines representations across such levels. Williams argues that the existing predictive views about cognitive representation cannot accommodate this fact because of the two core commitments of PP. First, representational hierarchy tracks computational distance from sensory surfaces (higher levels predict and receive error signals from the lower levels). Second, it tracks representations of phenomena at increasingly larger spatiotemporal scales (Hohwy, 2013). Following Jona Vance (2015), Williams argues that these two commitments are in tension with the standard strategy that assumes a common hierarchy of perception and cognition:

[If] beliefs are supposed to exist higher up the hierarchy and moving higher up the hierarchy is supposed to result in representations of phenomena at larger spatiotemporal scales, it should be impossible to have beliefs about extremely small phenomena implicated in fast-moving regularities [...] This point doesn’t apply to uniquely ‘high-level’ reasoning of the sort found in deliberate intellectual or scientific enquiry: patients suffering from ‘delusional parasitosis’ wrongly *believe* themselves to be infested with *tiny* parasites, insects, or bugs—a fact difficult to square with Fletcher and Frith’s (2009) suggestion ... that delusions arise in ‘higher’ levels of the hierarchy, if this hierarchy is understood in terms of increasing spatiotemporal scale (Williams, 2018, p. 16).

The term *beliefs* discussed above refers primarily to a folk-psychological notion of beliefs typically associated with the ‘personal level’. Such beliefs, according to the standard strategy, are indeed thought to be located at the higher levels of the generative hierarchy (or hierarchies). However, as Williams himself notes, “there is a sense in which all information processing within predictive brains implicates *Beliefs*... As such, *the same fundamental kind of representation* underlies representation in both sensory cortices and cortical regions responsible for intuitively ‘higher-level’ cognition”. (Williams, 2018, p.12). The difference between such higher-level beliefs and lower-level percepts may lie, for example, in conscious or metacognitive access, complexity, and resulting phenomenology, but not necessarily in the way of representing (although such possibility is not entirely out of the question) (Barsalou & Prinz, 1997; Goldstone & Barsalou, 1998; Tacca, 2011).

The two commitments regarding representations above (tracking computational distance from the sensory surfaces and increasing on the spatio-temporal scale) are indeed often considered to be part-and-parcel of the standard strategy in PP. However, there is no real tension between them and the standard strategy. Rather, the tension pointed

out by Vance (2015) stems not from the representational commitments themselves, but from (mis)treating the properties of representations (computational distance and spatiotemporal scale) as directly reflected in the updating process. First, let us consider computational distance. The bottom representational levels of the hierarchy are linked to the sensory inputs. As we move further along the hierarchy and further away from the sensors, we get progressively more complex representations that rely on multiple types of lower-level features. For example, lower levels of the visual hierarchy may represent simple features such as luminosity, edges, or shapes, while higher levels—multi-feature objects or an entire perceptual scene. At a certain point, representations may become multisensory.⁷ As higher-level representations are informed by prediction errors coming from the levels below, complexity, in a certain sense, implies more relevant variables. In a bottom-up processing approach this would also imply a more involved computation. However, when it comes to processing, in PP representations are not constructed and unpacked bottom-up by necessarily settling the lower levels of the hierarchy first. Instead, there is often significant top-down influence. Representations are taken to be conditionally independent of the levels not directly above or below them and updating across all levels is simultaneous, not sequential. Further, prediction error propagating upwards from any level may have a system-wide effect. Of course, certainty or precision of the higher-level hypotheses to a significant extent depends on accumulating prediction error and feedback from the levels below, and hence, in some cases, higher levels may take longer to *settle*, but this is situation-specific and not true in general. In other words, at least in principle, the content of the higher levels can be updated as flexibly and dynamically as that of the lower levels despite the difference in computational distance to the sensory surfaces.

Let us now turn to the spatiotemporal properties: Can PP accommodate higher-level representations about lower-level representations given that representations increase in the spatiotemporal scale as we move higher up the generative hierarchy? The specific example (having delusions about being infested with tiny parasites) chosen by Williams to present this challenge is unfortunate, as it ties the notion of the spatial increase in the hierarchy to the physical dimensions of the represented objects. Principally, there is no difference for PP whether one has delusions about tiny parasites or giant dinosaurs. What seems like a more relevant challenge is the span (time property) and scope (space property) of the hypotheses' content. To illustrate the point through the temporal case: If higher-level representations are in some way temporally spread-out, how can they represent temporally fine-grained (fleeting) things or properties as they exist in real-time? The key is that lower-level representations (say, object-level) are represented at the higher levels as part of a more complex representation that places them within a scene/world.⁸ Here, as in the case of the computational distance commitment, the way a PP-system processes information becomes relevant. Because the system is updated simultaneously across the hierarchy and does so in a largely conditionally independent manner, the object-in-a-scene representations are updated

⁷ Although, in a certain sense, **any level** may be informed by top-down predictions (which in turn may be informed by the sensory information from other sensory domains).

⁸ As kindly pointed out by one of the anonymous reviewers (verbatim).

as quickly as the object representations, as long as the relevant information is available. Hence, ultimately, although lower levels are, of course, relevant to the formation of the higher-level representations, neither the specific computational distance from sensory surfaces nor the position on the spatiotemporal scale play a significant role in PP's ability to support high-level hypotheses about low-level dynamically changing events.

4.2 Concepts are meaningfully located at a specific region of a hierarchy

Could we assume that *multimodal* or *amodal* representations can effectively predict the phenomena represented across different sensory modalities? This could happen, for example, by implementing *spherical* or *web-like architectures* (with proximal inputs across different modalities perturbing the outer edges, Penny, 2012; but see also Gilead et al., 2020), Still, according to Williams, this solution faces additional problems:

Either [PP] simply abandons the view that conceptual representation is meaningfully located at some region of a hierarchy in favour of a conventional understanding of concepts as an autonomous domain of amodal representations capable in principle of ranging over *any* phenomena [...]; *or* it offers some principled means of characterising the region of the hierarchy involved in non-perceptual domains. (Williams, 2018 p.16)

These alternatives seemingly leave the standard strategy at an impasse. According to Williams, to abandon the idea of meaningfully localizing conceptual representations in the hierarchy is to contradict the standard strategy, as we can no longer talk about cognitive representations being located above the perceptual ones. On the other hand, one may not be able to find adequate criteria for localization of conceptual representations. “Do my thoughts about electrons activate representations at a different position in “the hierarchy” to my thoughts about the English football team’s defensive strategy, or the hypothesised block universe? If so, by what principle?” (Williams, 2018, p. 17). I argue that it is senseful to talk about conceptual representations or (winning) hypotheses associated with personal level beliefs, thought, etc., as being located further away from the sensory cortices than non-conceptual perceptual representations without necessarily having to characterize the specific “region of the hierarchy *involved* in non-perceptual domains” (Williams, 2018, p. 16). In fact, attempting to specify such a region in a certain sense would mean to fundamentally misunderstand the core tenets of PP.

To begin with, although there may be many situations where the lower perceptual levels do not play a significant role in cognitive inference (e.g., offline simulation), in principle, any level of the generative hierarchy (assuming it has enough weight) can affect processing at any other level. On the other hand, such an approach does not either preclude or necessitate activation of sensory cortices when it comes to thinking about concepts related to perception. The standard strategy merely postulates that higher-order cognitive beliefs are located above perceptual ones, not that there is a certain restricted region that is exclusively involved in non-perceptual domains or that perceptual information must necessarily be used to arrive at a higher-level (cognitive)

hypothesis. Williams' example (2018, p. 16) of a blind person thinking about light patches is meant to present a challenge to the standard strategy but it fails as PP does not require activation of the relevant part of the visual cortex in order to conceptualize about light patches. Depending on each individual case (innate or acquired blindness, etc.) the generative model and the processing path may differ. This, to a certain extent, is true even for people without visual (or other sensory) impairments. Generative models are specific to each individual and their learning paths, and even within one individual hypotheses about similar events may be formed in different ways case by case, with the error signal coming from any level of the hierarchy.

Another complication in characterizing a specific region involved in a non-perceptual domain or allocating concepts at certain levels in absolute terms is that lower-level hypotheses contributing to conceptual representations have complex branching structures with branches of different depth and potentially incorporate information from multiple sensory domains. For that reason, it would not be possible to pinpoint a specific level "n" across the hierarchy that would purely consist of representations of concepts (or any kind of representations of certain complexity or united by certain features). None of this, however, precludes us from making sense of the cognition-on-top approach in PP. Rather than fixing the location, we just need to specify how concepts arise and relate to the contributing lower-level perceptual hypotheses.

4.3 Concepts are dynamic representations

So, what would be a helpful way to think about concepts in PP with respect to the considerations above? Following Michel (2016, 2020), I suggest that concepts in PP may be best thought of as "highly flexible, dynamic, and context-dependent representations" (Michel, 2020, p. 625), rather than relatively stable theories, and include ad-hoc, non-consciously accessible, multi-modal representations with cross-domain connections (Michel, 2016, 2020). A concept of a cat, for example, can be linked to various features such as fluffy texture, purring sound, and triangular ears, as well as a range of "cat-related" situations. Conceptual representations, however, are "thinner" than the contributing representations and do not contain all the richness of detail. According to Michel (2020), the cognitive role of concepts is precisely in abstracting away unnecessary, irrelevant information. This allows the brain to more efficiently generate error-minimized predictions given the high cost of metabolic activity. For example, in order to avoid stepping on a cat, it would be sufficient to have information about its rough shape, size, and some general patterns of behavior (Michel, 2020). In some cases, entirely new prediction units may be created on the fly for abstracting a frugal representation (Kiefer, 2019; Michel, 2020). To qualify as concepts in the traditional sense of the word, such representations would need to additionally gain conscious accessibility (perhaps, through language), relative stability, and applicability across a certain range of domains.

A thin and fleeting concept could grow into a rich and stable one if it turns out to be useful in the prediction economy. Also, representations that initially have a narrow range of application might get a more generalized use through mechanisms like "neural recycling." (Michel, 2020, p. 635)

The existing concepts can continually fine-tune their cognitive content and modulate the content used for inferences depending on the context. Picking the relevant parts of the cognitive content in each context and selecting the most efficient route for categorization may be afforded by precision-weighting. To summarize, concepts are predictive units crucial for “data compression and context-sensitive modulation of the prediction detail” (Michel, 2020, p. 634). To that, I would add that such compression is not only beneficial for efficient processing, but also for communication. As a speculative point, this may be the reason why language operates with lexical units that tend to correspond to stabilized concepts. This approach to concepts as dynamic multi-modal representations (although not in the PP context) is further supported by the recent findings on distributed concepts (Handjaras et al., 2016), re-wiring experiments (Newton & Sur, 2005), and neural re-usage phenomena (Anderson, 2010) that suggest a high degree of flexibility and interconnectivity (Michel, 2020).

Importantly, such an approach does not contradict either the idea of spatiotemporal scale increase when it is understood as hypothesis generality, or Williams’s suggestion that abstract concepts are not strictly causal (they are aggregations constructed for efficiency, but not necessarily something meaningfully located in the world). Further, the notion of concepts as dynamically generated hypotheses abstracting features of the lower-level representations still allows concepts to be meaningfully located in a region of the hierarchy above hypotheses directly related to sensory input. Concepts are essentially multimodal, cross-domain representations constructed on top of the non-conceptual perceptual ones, and the more abstract ones are further removed from their perceptual beginnings (see, e.g., Meteyard et al., 2012; Kiefer & Hohwy, 2018). The dynamic “made to order” representations, on one hand, extrapolate from perceptual features and may serve as a proper top level of the perpetual hierarchy, essentially forming our perceptual ontology. On the other hand, they serve as constituents of thought and stabilize into less dynamic “full” concepts. This is in line with the weak embodiment approach to concepts (see, for example, Meteyard et al., 2012 and Dove, 2018), works nicely with the traditional “cognition-on-top” PP, which seems to presuppose some kind of transitional area (levels) that would bridge perception and cognition, and accommodates the observation that perceptual categories are much more context-dependent and flexible than concepts **proper** (Deroy, 2019). That said, where specifically the border between conceptual and non-conceptual representations lies in perception is a topic of ongoing debate. One extreme view is to argue that all perceptual representations in the hierarchy with the only exception of the lowest level (directly representing the incoming sensory signal) could be called conceptual. In such a case, conceptual representations would span almost the entirety of the generative hierarchy. Yet, it seems that for most people, including Michel (2020), when it comes to conceptual representations at least a certain degree of abstraction is implied, which means that concepts are located, for example, above the specific structured collections of visual features that we recognize (with the help of concepts or not) as instances of specific objects. This interpretation is consistent with the standard strategy in PP and puts concepts on top of the non-conceptual perceptual levels.

But how does this picture result in the kind of seemingly compositional expressions combining “representations from across levels of any conceivable hierarchy” that we register as thoughts (Williams, 2018, p. 17) on the conscious level? This question

brings us back to the discussions of compositionality and the interaction between language and thought.

5 Separation between thought and language in PP may address the compositionality problem

Let us return to William's notion of *rich compositionality*. According to Williams, compositionality is a "principle of representational systems in which the representational properties of a set of atomic representations *compose to yield* the representational properties of molecular representations" (my emphasis, Williams, 2018, p. 18). Compositionality of the representational system is further commonly taken to explain productivity and systematicity of human cognition, that is our capacity to come up with an infinite number of meaningful utterances and the fact that the ability to produce some thoughts is inherently tied to the ability to produce others. Nevertheless, as Williams notes, it is not enough to express compositionality in terms of productivity and systematicity as all sorts of cognitive architectures that are otherwise limited can potentially produce them in different varieties (Williams, 2018). For example, both productivity (infinitary character) and systematicity underlying thought only require functional systematicity (see Sect. 3). As discussed in Sect. 2, due to the lack of access to, or representative capture of the underlying mechanisms, even the assessment of directionality of the process cannot be safely made. Whether in thought atomic representations compose to *yield* the representational properties of molecular representations is contentious.

To more strictly specify the compositional requirements for thought, Williams proposes a definition of rich compositionality, that is the property of being *at least as expressive as first-order logic*. (Williams, 2018). As he notes, this seems a fairly minimal requirement: "Notice how low the bar is in the current context: all one requires is to show that *some* features of higher cognition are at least as expressive as first-order logic. This claim could not plausibly be denied" (Williams, 2018, p. 21). He further argues that PP is unable to satisfy this requirement due to its commitment to the kind of connectionist architecture that may be represented through hierarchical probabilistic graphical models. Such graphical models have expressive power equivalent to propositional logic, that is limited to facts with "operations [...] defined over atomic (i.e. unstructured) representations of such facts—namely, propositions" (Williams, 2018, p. 20). The ontology of first-order logic, on the other hand, comprises not just facts but *objects* and *relations*, thereby representing "the world as having *things* in it that are *related* to each other, not just variables with values" (Russell & Norvig, 2010, p. 58, as cited by Williams, 2018). Williams's proposed strategies are either to abandon the specific predictive coding architecture, restrict the scope of the framework, or relegate the explanation of the aforementioned phenomena to the "distinctive contribution of natural language and *public* combinatorial symbol systems more generally" (Williams, 2018, p. 22). The challenge with the latter strategy is to explain how "human thought... inherit[s] such systematicity as it displays from the grammatical structure of human language itself" (Clark, 2000, as cited by Williams, 2018, p. 23). In light of the discussion in Sect. 3, I would like to reformulate this challenge not in terms of inheritance

or transfer, but in terms of *interaction* between thought and language. The lateral connections and non-homogeneous priors so strongly emphasized by Clark allow for a variety of multidimensional subsystems (including linguistic) that integrate at various points and whose existence, nevertheless, does not contradict the commitment to the predictive coding architecture. In the following paragraphs, I outline a PP-type picture that explains how conceptual and linguistic representations work together to produce the kind of introspective experience of thought that we have.

5.1 PP-based proposal of how conceptual thought and language interact

Consider a PP hierarchy of inference (*perceptuo-conceptual hierarchy* or PCH) in line with the standard strategy. Conceptual thought in the form of hypotheses is distributed across different levels. Each node below the top one represents a lower-level hypothesis and is taken to be causally dependent on the adjacent levels only. The processing system operates simultaneously both in top-down and bottom-up fashion progressively updating the precision-weighted predictions supplied in the downward flow. As discussed in the previous section, concepts in PP are representational systems that form dynamically as hypotheses and include ad-hoc, non-consciously accessible, multi-modal representations with cross-domain connections that help to efficiently generate error-minimized predictions (Michel, 2020). Concepts representing relations between the fact-type hypotheses (described by Williams as the only kinds of hypothesis available to PP) are similarly just hypotheses/concepts statistically extrapolated from individual instances involving such relations. The rich dimensionality of the generative models (and human neural networks) allows for such dependencies to be integrated as separate inferential nodes.

Certain nodes in the PCH (especially at the level of stabilized concepts) may be associated with linguistic tags on a probabilistic basis. Such matching can happen either simultaneously with the inferential process in the PCH or after the inference stabilizes—recall the idea of decomposition of thought to obtain the meaning of separate parts (Sect. 3). The former seems to me like the more likely option, although this question should be settled empirically.

The rules and regularities of language use may form their own part of the hierarchical generative model (LH). The representations in the LH are essentially of the same type as in the PCH. However, the interaction between LH and PCH parts of the model is characterized primarily by lateral connectivity, while, internally, LH and PCH are characterized by a more traditional, primarily hierarchical architecture. The lateral connections between conceptual and linguistic hierarchies occur at various levels (starting with the level where the relationship between concepts and words may be identified). However, not all concepts (whether of relationship or object type) have to directly map onto linguistic tags used to express thought. It is sufficient that the conceptual system *overall* can be mapped to a corresponding linguistic system. Successful association of these objects/concepts with the linguistic tags allows the system to start the sentence generation process, which is done in a PP simultaneous bottom-up and top-down manner.

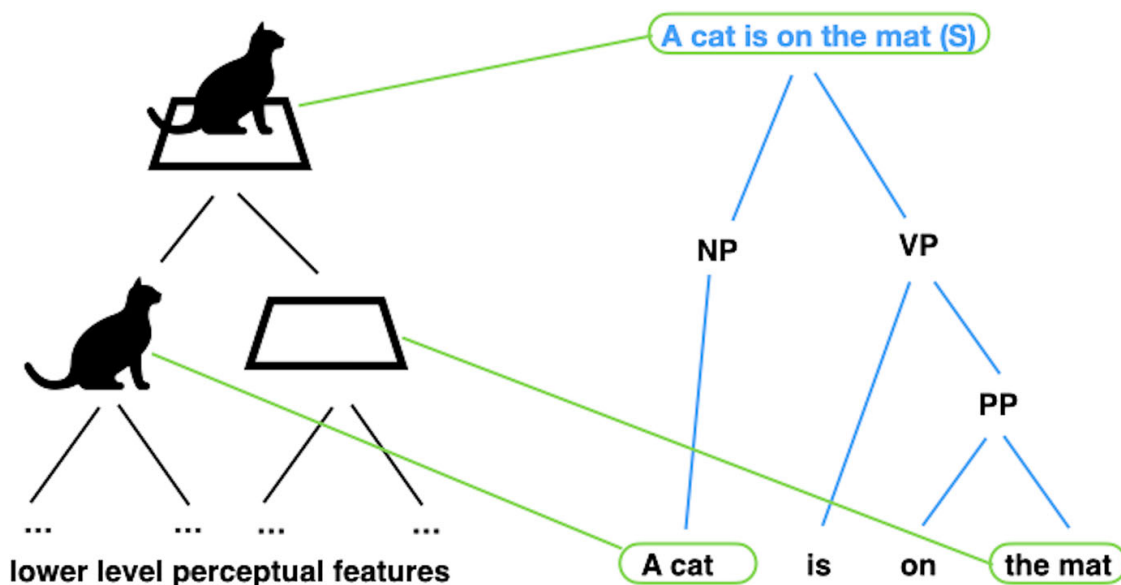


Fig. 1 A cat is on the mat

In fact, element-wise one-to-one mapping between LH and PCH may be impossible because thought and language may have different consistency and coherence requirements. Concepts, for example, may be largely redundant and reemerge in different locations in the PCH.⁹ In fact, the picture of the PCH introduced above makes it unlikely that conceptual representations would form “a coherent and consistent body of knowledge that we can fully formalize in a propositional or language-like format” (Michel, 2020, p.636). As Michel points out, some argue that such inconsistency is not only a necessary property of a human mind (Sorensen, 2004) but a desirable evolutionary feature in a highly uncertain environment (Bortolotti & Sullivan-Bissett, 2017). Natural language, on the other hand, requires (with some exceptions) consistency and coherence. Perhaps, “we should view formal systems like [languages] as cultural artifacts that contribute to shaping the mind rather than constitute it” (Michel, 2020, p.636; see also Dutilh Novaes, 2012, p. 61). Now, to illustrate the model above on a concrete example, consider the following sentence: A cat is on the mat (Fig. 1).

This may be either a description of the ongoing perceptual experience, or a thought that is isolated from the current sensory input. The former is the standard case of perceptual processing. In order to generate a sentence that describes the scene (perhaps, describes best given a specific context) we require nodes in the perceptual hierarchy that can ‘anchor’ linguistic tags. An obvious contender here is object-level hypotheses (cat, mat) presenting instances of dynamic representational concepts of the type described above. Prepositions, such as ‘on’, may also have statistically associated lateral links to the relational concepts in the PCH. The mapping, however, does not

⁹ As suggested by one of the anonymous reviewers, this redundancy may also point to a solution to the generality challenge. For example, the rapidly changing edges in the visual field may be represented directly in early vision close to the sensory periphery, but also higher up in the hierarchy as a proper concept. This would help to differentiate between the level of abstraction of the regularity vs. representation. The reviewer’s concern is that it may be rather ad hoc to supplement PP with such additional distinctions, but see the discussion directly following this footnote in the main text.

require for the relational concept to be as straightforward as the proposition itself—the latter is associated with the former on a good-enough basis. Further, the associated relational concept may be positioned above object concepts in the PCH. The mutual relationship between concepts in the hierarchy does not matter as the linguistic system is connected to the PCH laterally. Other lexical items like, ‘is’, ‘a’, or ‘the’ may be potentially treated similarly (as they may be taken to be meaning-bearing), but in general, as discussed, not all the words have to correspond to some nodes in the PCH—Grammar requirements may be settled within and expressed locally in the predictions running through the LH (see Rappe, 2019). For example, the ‘cat’ node in LH is associated with the cat representation in the PCH, but as part of LH is also associated with the specific rules of use. These rules of use are statistical properties—strong predictions about the grammatical context in which these nodes may occur. Explicit rules may be abstracted from these regularities, but the idea here is that we originally learn “language in use”, and the explicit rules follow later (if they are at all explicitly represented).¹⁰ Observation of the semantic-syntactic rules of natural language can lead to the introduction of new sentence parts that do not have direct correspondents in the PCH.

Linguistic expression stabilizes both based on the LH feedback loop and the prediction error coming from the PCH. Linguistic hierarchy has a significantly more rigid structure and captures thought as a kind of net stretched over the generative model. For the linguistic expression to match the content of the winning inferential hypothesis, linguistic subhierarchy must arrive at the state with undetectable relevant prediction error internally (which ensures coherence of the linguistic utterance) but the lateral communication between LH and PCH must also stabilize.

This relationship between PCH and LH is mutual with both hierarchies being able to influence each other. The situation is very similar for the case of thought that does not represent the current sensory input, except for the hyperprior that the lower-level sensory information does not play as important a role for error generation. The mutual bootstrapping relationship between PCH and LH in this case also helps address Williams’s point that a concept cannot be generated without the activation of the sensory path. That said, the generation of imagery and phenomenological effects such as tasting, smelling, and experiencing tactile sensations may often accompany thought. Perhaps, this could be explained by the downstream effect of prediction generation.

5.2 Surface compositionality and implications of the approach

Inferential hierarchies are functionally compositional, with top-down prediction propagation. The appearance of bottom-up and concatenative compositionality (at the manifestation level) in thought is largely due to the fact that thoughts, when they manifest at the conscious level, typically take the shape of natural language. When we receive language input, it unfolds in a temporally (oral) or spatially (written) spread-out

¹⁰ Although Fig. 1 represents linguistic hierarchy in the old-fashioned syntactic-tree style, the approach does not presuppose any specific way of encoding linguistic rules. The representational model above is chosen simply for convenience.

manner. Due to the jigsaw-puzzle-like nature of grammar (allowing for the composition of higher-order units from the lower-level ones by directly combining them), thought appears to have concatenative compositionality. This is also true for language production as it requires uttering linguistic units one by one. Both inferential models and linguistic processing models are updated in the bidirectional fashion as prescribed by PP architecture.¹¹ However, this is done in a sort of parallel-processing fashion with concept-to/from-linguistic-unit bridges often serving as the lowest-level connections.

Importantly, both language and thought influence each other. Linguistic tags help to improve the certainty of the ongoing inference, but they can also evoke or activate inference through their relations to the concept-hypotheses. The inference-language loop may also function as an additional source of prediction errors to match the hypothesis against during offline simulation and serve as a kind of bootstrapping mechanism.

The picture described above captures productivity and systematicity of thought as well as functional compositionality in inference and concatenative compositionality in language. It also allows us to explain why concatenative compositionality does not hold in some cases, such as that of metaphor: language is about the best approximation of thought **overall**, but it is not about building thought from the ground up. Although both thought and natural language hierarchies ultimately describe the same states of affairs, the mapping is not necessarily one-to-one.

To summarize, the interconnectedness of thought and language is what creates an illusion of concatenative bottom-up compositionality in conceptual thinking. The requirement of separate conceptual and linguistic hierarchies may resemble Williams's second strategy where the properties of compositionality and generality are delegated to language and public symbolic systems. The crucial difference is that all parts of my sketch are purely PP-based. For this reason, despite introducing multiple subsystems, we may still be able to talk about a single unified model implemented by the brain. This, however, requires one to accept unification at the computational (as opposed to the implementational) level. After all, both conceptual and linguistic hierarchies adhere to similar processing principles, are deeply integrated, and are in constant communication afforded by lateral connections.

5.3 Negation, predication, and quantification

One important challenge for the proposed account is to provide at least a provisional story about how properties such as negation, predication, and quantification may be realized in the brain with such an architecture. Although a full-fledged answer would merit a separate paper (likely even several), I will attempt to sketch out some promising directions below.

When it comes to predication, there are two important challenges that need to be addressed. The first one concerns the extraction of predicates from the PCH as separate representations. The second one relates to the syntactic consistency and expression of predication in language (this distinction is also highlighted in Kiefer, 2019). The question of extraction of predicates is not dissimilar to the question of representing relations, and this can be done through the process of conceptualization described in

¹¹ For more on language processing within the PP paradigm, see Rappe (2019).

Sect. 4.3. Importantly, conceptual representations do not deal with syntactic consistency, but rather define content, for example, an action or a property (although PCH may inform LH when it comes to coordinating the units of linguistic expression). The details of linguistic expression, such as grammatical coherence, on the other hand, can be resolved purely within the LH. Further, changing specific elements in a proposition may be achieved through the feedback loop between LH and PCH realized by the lateral connections. For a more detailed discussion of predication in the connectionist networks see Kiefer (2019).

The question of quantification also has two aspects. The first aspect is quantification over a finite number of entities, which is represented in language by the vague quantifiers such as “few”, “several”, or “many”. The second aspect is that of the universal quantifiers, such as “all” or “every”, which introduce an additional aspect of infinitude. When it comes to the vague quantifiers, my hunch is that they do not refer to any specific numerical quantities but instead specify contextual and communicative factors such as, for example, the relative size of the objects involved in the scene, or their expected frequency (Moxey & Sanford, 1993; Newstead & Coventry, 2000; Rakapakse et al. 2005a). Vague quantifiers are then used in language to represent ad hoc, “thin” concepts relaying an estimation of precise information in a context-dependent manner. One possibility is that the use of vague quantifiers is initially learned in the perceptual domain and then can be applied outside this context. A simple connectionist model of quantification of visual information is presented, for example, in Rakapakse et al. (2005a). The authors report that the networks in the model are able to perform the production of “psychological numbers” produced by human subjects and that, in producing such judgments, they use similar mechanisms (Rajakapakse et al., 2005a, 2005b). A provisional story regarding universal quantification is presented in Kiefer, 2019 (specifically, Sect. 6.8).

Negation, perhaps, presents the most interesting challenge. A large body of literature suggests that negation makes comprehension more difficult or, at least, takes longer to be processed (see e.g., Fodor & Garrett, 1966; Carpenter et al., 1999; Tettamanti et al., 2008; Bahlmann et al., 2011). Sometimes, this is taken as a sign that the processing of utterances with negation requires an additional step compared to those not involving negation. More recently, Yosef Grodzinsky and colleagues (2020) have found that negation is governed by a brain mechanism located outside the language areas, which further suggests that negation may not be a linguistic process. Together with the finding that negative phrases or sentences yield reduced levels of activity in regions involved in the representation of the corresponding affirmative meanings, this supports the decompositional approach to negation—the idea that a positive reversal of the negative utterance is represented first, which is then followed by inhibition of this representation. Further investigation in the nature and levels of brain activity led Papeo and de Vega (2020) to conclude that “processing negated meanings involves two functionally independent networks: the response inhibition network and the lexical-semantic network/network representing the words in the scope of negation” (p. 741). This naturally aligns with the approach proposed in this paper. Perhaps, negation, at least, at the sentential level, is realized as inhibition of the reverse representation in the PCH. Given lateral connectivity between PCH and LH, as well as simultaneous

updating at all levels, it seems plausible that the negation aspect is settled after the main factual content-bearing elements of the utterances are represented.

6 Conclusions

My aim was to show that compositionality and generality may not necessarily present problems for PP even under the assumption of a predictive coding architecture. The proposal outlined above offers a new perspective on higher-order cognition in linguistic and non-linguistic creatures as well as different types of non-linguistic cognition in humans. Completing such an account would require extensive cooperation of philosophers and cognitive scientists, but my point is theoretical: two abilities are sufficient to capture conceptual thinking in predictive terms—the ability of concept formation as construction of dynamic representational models and a full-fledged linguistic apparatus, also embedded in PP. What is important here is to consider lateral connectivity. Sticking to the very literal sense of top and bottom in the generative hierarchies means looking at generative models as conceptual networks in an overly Fodorian way and ignoring the simultaneous model updating.

The separation and interaction between language and thought raise no shortage of questions, many being already explored. My interest here is how PP provides a twist on these discussions, offering both a new way to ask questions and formulate the answers. Importantly, separating language and thought does not betray the unifying spirit of PP as the systems communicate extensively at different levels, mutually informing and bootstrapping one another. Further, as Williams notes, it is not certain that we need to commit to the single predictive coding architecture within PP at all. Biology and current discussions warn us that a plurality of principles may be at stake. That said, prediction error minimization seems like a useful explanatory concept and if compositionality and generality do not present unresolvable problems for PP in principle, there is no need to get off the PP horse right now, at least when it comes to conceptual thought.

Acknowledgements I would like to thank my primary Ph.D. supervisor, Ophelia Deroy, for the critical feedback and productive discussions throughout the process of working on the manuscript. I would like to also thank the two anonymous reviewers, Sam Wilkinson, Stephan Sellmaier, Nina Poth, members of the LMU Neurophilosophy Colloquium, and the participants of the Third Bochum Early Career Researchers Workshop in Philosophy of Mind and of Cognitive Science for their helpful suggestions and comments on the earlier drafts.

Funding Open Access funding enabled and organized by Projekt DEAL. N/A.

Availability of data and material N/A.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this

article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4, 47.
- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4), 245–266.
- Bahlmann, J., Mueller, J. L., Makuuchi, M., & Friederici, A. D. (2011). Perisylvian functional connectivity during processing of sentential negation. *Frontiers in Psychology*, 2, 104.
- Barsalou, L. W., & Prinz, J. J. (1997). Mundane creativity in perceptual symbol systems. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 267–307). American Psychological Association.
- Bortolotti, L., & Sullivan-Bissett, E. (2017). How can false or irrational beliefs be useful? *Philosophical Explorations*, 20(sup1), 1–3.
- Call, J., & Tomasello, M. (2011). Does the chimpanzee have a theory of mind? 30 years later. In *Human Nature and Self Design* (pp. 83–96). mentis.
- Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1999). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *NeuroImage*, 10(2), 216–224.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.
- Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
- Clark, A. (2019). Consciousness as generative entanglement. *The Journal of Philosophy*, 116(12), 645–662.
- Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, 26(9–10), 19–33.
- Clark, A. (2000). *Mindware: An introduction to the philosophy of cognitive science*. Oxford University Press.
- Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science*, 68(2), 451–484.
- Deroy, O. (2019). Predictions do not entail cognitive penetration: “Racial” biases in predictive models of perception. In C. Limbeck-Lilienau & F. Stadler (Eds.), *The Philosophy of Perception* (pp. 235–248). De Gruyter.
- Dołęga, K., & Dewhurst, J. E. (2020). Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*, 1–26.
- Dove, G. (2018). Language as a disruptive technology: Abstract concepts, embodiment and the flexible mind. *Philosophical Transactions of the Royal Society b: Biological Sciences*, 373(1752), 20170135.
- Dutilh Novaes, C. (2012). *Formal languages in logic: A philosophical and cognitive analysis*. Cambridge University Press.
- Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369(1), 132.
- Finn, J. K., Tregenza, T., & Norman, M. D. (2009). Defensive tool use in a coconut-carrying octopus. *Current Biology*, 19(23), R1069–R1070.
- Fiorito, G., & Scotto, P. (1992). Observational learning in *Octopus vulgaris*. *Science*, 256(5056), 545–547.
- Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48.
- Fodor, J., & Garrett, M. (1966). Some reflections on competence and performance. *Psycholinguistic papers*, pp 135–179.
- Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford University Press on Demand.
- Frankish, K. (2018). Inner Speech and Outer Thought. *Inner Speech: New Voices*, 221.

- Frege, G. (1879). Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. *From Frege to Gödel: A Source Book in Mathematical Logic, 1931*, 1–82.
- Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50. Reprinted in Frege G. (1956). The philosophical writings of Gottlieb Frege (trans: Black M).
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 360(1456), 815–836.
- Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex*, 68, 129–143.
- García-Carpintero, M. (1996). Two spurious varieties of compositionality. *Minds and Machines*, 6(2), 159–172.
- Gabriel, G., Hermes, H., Kambartel, F., Thiel, Ch., Veraart, A. (Eds.). (1976). Gottlob Frege. Wissenschaftliche Briefwechsel. Felix Meiner: Hamburg.
- Gilead, M., Trope, Y., & Liberman, N. (2020). Above and beyond the concrete: The diverse representational substrates of the predictive brain. *Behavioral and Brain Sciences*, 43.
- Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, 65(2–3), 231–262.
- Grodzinsky, Y., Deschamps, I., Pieperhoff, P., Iannilli, F., Agmon, G., Loewenstein, Y., & Amunts, K. (2020). Logical negation mapped onto the brain. *Brain Structure and Function*, 225(1), 19–31.
- Handjaras, G., Ricciardi, E., Leo, A., Lenci, A., Cecchetti, L., Cosottini, M., Moratta, G., & Pietrini, P. (2016). How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. *NeuroImage*, 135, 232–242.
- Harman, G. (2015). *Thought*. Princeton University Press.
- Heavey, C. L., & Hurlburt, R. T. (2008). The phenomena of inner experience. *Consciousness and Cognition*, 17(3), 798–810.
- Hermes, H., Kambartel, F., Kaulbach, F., Long, P., & White, R. (1981). Gottlob Frege. Posthumous writings.
- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3, 96.
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, 35(2), 209–223.
- Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3), 687–701.
- Horga, G., Schatz, K. C., Abi-Dargham, A., & Peterson, B. S. (2014). Deficits in Predictive Coding Underlie Hallucinations in Schizophrenia. *The Journal of Neuroscience*, 34(24), 8072–8082.
- Hvorecny, L. M., Grudowski, J. L., Blakeslee, C. J., Simmons, T. L., Roy, P. R., Brooks, J. A., ... & Holm, J. B. (2007). Octopuses (*Octopus bimaculoides*) and cuttlefishes (*Sepia pharaonis*, *S. officinalis*) can conditionally discriminate. *Animal cognition*, 10(4), 449–459.
- Janssen, T. M. (2001). Frege, contextuality and compositionality. *Journal of Logic, Language and Information*, 10(1), 115–136.
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, 195(6), 2387–2415.
- Kiefer, A. B. (2019). *A defense of pure connectionism*. (Doctoral dissertation, City University of New York, NY). Retrieved from https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=4098&context=gc_etds.
- Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, 8, 302.
- Levinson, S. C. (1997). From outer to inner space: linguistic categories and non-linguistic thinking. *Language and conceptualization*, 13–45.
- Machery, E. (2005). You don't know how you think: Introspection and language of thought. *The British Journal for the Philosophy of Science*, 56(3), 469–485.
- Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7), 788–804.
- Michel, C. (2020). Concept contextualism through the lens of Predictive Processing. *Philosophical Psychology*, 33(4), 624–647.
- Michel, C. (2016). *What could concepts be in the Predictive Processing framework?* (Master dissertation, the University of Edinburgh, Scotland, UK). Retrieved from <https://era.ed.ac.uk/handle/1842/21875>.

- Miller, M., & Clark, A. (2018). Happily entangled: Prediction, emotion, and the embodied mind. *Synthese*, 195(6), 2559–2575.
- Moxey, L. M., & Sanford, A. J. (1993). *Communicating quantities: A psychological perspective*. Lawrence Erlbaum Associates, Inc.
- Nelsen, R. (1993). *Proofs without words: Exercises in visual thinking*. Washington. Mathematical Assoc. of America.
- Newstead, S. E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, 12(2), 243–259.
- Newton, J. R., & Sur, M. (2005). Rewiring cortex: Functional plasticity of the auditory cortex during development. In *Plasticity and signal representation in the auditory system* (pp. 127–137). Springer, Boston, MA.
- Palmer, C. J., Paton, B., Kirkovski, M., Enticott, P. G., & Hohwy, J. (2015). Context sensitivity in action decreases along the autism spectrum: A predictive processing perspective. *Proceedings of the Royal Society of London b: Biological Sciences*, 282(1802), 2014–1557.
- Papeo, L., & de Vega, M. (2020). The neurobiology of lexical and sentential negation. In V. Déprez & M.T. Espinal (Eds.), *The Oxford Handbook of Negation*. Oxford University Press, USA.
- Pellicano, E., & Burr, D. (2012). When the world becomes ‘too real’: A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10), 504–510.
- Penny, W. (2012). Bayesian models of brain and behaviour. *ISRN Biomathematics*, 2012.
- Rajapakse, R. K., Cangelosi, A., Coventry, K. R., Newstead, S., & Bacon, A. (2005a). Connectionist modeling of linguistic quantifiers. In *International Conference on Artificial Neural Networks* (pp. 679–684). Springer, Berlin, Heidelberg.
- Rajapakse, R., Cangelosi, A., Coventry, K., Newstead, S., & Bacon, A. (2005b). Grounding linguistic quantifiers in perception: Experiments on numerosity judgments. In *2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.
- Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79.
- Rappe, S. (2019). Now, Never, or Coming Soon? Prediction and Efficient Language Processing. *Pragmatics & Cognition*, 26(2–3), 357–385.
- Richter, J. N., Hochner, B., & Kuba, M. J. (2016). Pull or push? Octopuses solve a puzzle problem. *PLoS one*, 11(3).
- Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Pearson.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565–573.
- Seth, A. K. (2015). The Cybernetic Bayesian Brain—From Interoceptive Inference to Sensorimotor Contingencies, (w:) Open MIND, red. T. Metzinger, JM Windt.
- Sorensen, R. A. (2004). *Vagueness and contradiction*. Clarendon Press.
- Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10, 550.
- Tacca, M. C. (2011). Commonalities between perception and cognition. *Frontiers in Psychology*, 2, 358.
- Tettamanti, M., Manenti, R., Della Rosa, P. A., Falini, A., Perani, D., Cappa, S. F., & Moro, A. (2008). Negation in the brain: Modulating action representations. *NeuroImage*, 43(2), 358–367.
- Thagard, P. (2019). *Brain-Mind: From Neurons to Consciousness and Creativity (Treatise on Mind and Society)*. Oxford University Press.
- Tversky, B. (2019). *Mind in motion: How action shapes thought*. Hachette UK.
- Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, 121(4), 649.
- Vance, J. (2015). Review of The Predictive Mind. *Notre Dame Philosophical Reviews*.
- Velasco, P. F., & Loev, S. (2020). Affective experience in the predictive mind: a review and new integrative account. *Synthese*, 1–36.
- Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing: 1. MIND Group*, Frankfurt am Main.
- Wilkinson, S., & Fernyhough, C. (2018). *When Inner Speech Misleads*. Oxford University Press.
- Williams, D. (2018). Predictive coding and thought. *Synthese*, pp 1–27.

Zuberbühler, K. (2020). Syntax and compositionality in animal communication. *Philosophical Transactions of the Royal Society B*, 375(1789), 20190062.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 3

Paper 2. The clear and not so clear signatures of perceptual reality in the Bayesian brain

Abstract: In a Bayesian brain, every perceptual decision takes into account internal priors as well as new incoming evidence. A reality monitoring system—eventually providing the agent with a subjective sense of reality—prevents us from being confused about whether our experience is perceptual or imagined. Yet not all confusions lead us to feel that we may be imagining: some experiences feel unmistakably perceptual, yet not quite right. What happens in such confused perceptions, and can the Bayesian brain hypothesis explain this kind of confusion? In our paper, we offer a characterization of perceptual confusion and argue that it requires our subjective sense of reality to be a composite of several subjective markers, including a categorical one, which can clearly identify an experience as perceptual and connected to reality. Our composite account makes new predictions regarding the robustness, the non-linear development, and the possible breakdowns of the sense of reality in perception.

Submitted to a journal as: Deroy, O., **Rappe, S.** (in review). The clear and not so clear signatures of perceptual reality in the Bayesian brain.

Author contributions: S.R. conducted the initial literature review. S.R. and O.D. conceived of the research idea, wrote the original draft, and revised the manuscript for publication. Both authors recognize equal contributions to the paper.

The clear and not so clear signatures of perceptual reality in the Bayesian brain

Ophelia Deroy ^{1,2,3} & Sofiia Rappe ^{1,4}

¹ Faculty of Philosophy, Ludwig-Maximilians-Universität München, Germany

² Munich Center for Neuroscience, Ludwig-Maximilians-Universität München, Germany

³ Institute of Philosophy, School of Advanced Study, University of London, UK

⁴ Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Germany

1. Introduction

One day in January 2010, Paul “Bear” Vasques was sitting in his front yard, somewhere near Yosemite National Park, California. He caught sight of a rainbow that appeared to be a double one. Incredulous and ecstatic, he filmed the spectacle and posted it on YouTube. Since then, his video has stood out for its extraordinary virality (it now has more than 48 million views). We are more interested in the reactions that Paul records on it. From expressing the surprise to see a double rainbow, his tone goes into incredulity: we hear him doubting, laughing, and finally bursting into tears as he starts wondering whether the double rainbow is a sign sent to him by God. Several interpretations of what happened to Paul can be provided, but one seems to us particularly interesting: all along, as he is documenting the scene on his phone, he must be sure that the rainbow is real and that viewers will also see it. As he films, other feelings emerge and are recorded, but none seems to challenge his certainty that there is indeed a double rainbow out there that he is recording. What these feelings resemble most, instead, is a state of confusion, though perhaps of a milder form compared with states of acute confusion which are well-documented in the clinical literature: subjective state of altered consciousness characterized by disordered attention and/or awareness and manifested by a diminished speed, clarity, and coherence of thought (e.g., Francis & Young, 2012). What this capture of Paul’s state suggests is that he is in a state where he both realizes without a doubt that he is perceiving the rainbow and yet keeps being somewhat confused by or about its reality. The description can at first strike as an oxymoron or an internal contradiction: how can certainty that one is perceiving, and is therefore presented with a real scene or object,

coexist with a state of confusion, where one seems to grapple with the reality of that scene? Dissolving this apparent tension is what our main goal is. More particularly, we want to show why a proper understanding of the subjective sense of reality (and its disturbances) should not limit itself to telling the subject whether they are perceiving or imagining. While some cases of acute confusion can be accompanied by hallucinations or states where one is uncertain whether they are imagining or perceiving a given scene, some other states of confusion occur while the subject has no doubt, and is indeed correct, that their experience is presenting us with a real scene. We suggest then that the sense of reality has both a broader function, and more dimensions than the current debates consider. Our question is whether this richer understanding of the sense of reality raises problems for the Bayesian solutions provided to explain the narrower function and nature of the sense of reality as monitoring only the perception-or-imagination boundary.

The case of Paul Vasques is here only an illustration of the kind of states we are interested in analyzing and propose to call “extraordinary perceptions”—that is, states where a categorical certainty that one is perceiving, and therefore presented with external reality, can coexist with a state of confusion surrounding this reality. The terminology matters here, and we want to highlight the difference between what we call here “categorical certainty” and “confusion.” Though the distinction will become clearer as we proceed, we can say that categorical certainty is about the perceptual status of the state, while confusion is about the difficulty of coping with the perceptual reality as real. Keeping this difference in mind should dissipate the difficulty that one could have thinking about a state where one is both certain and uncertain about the same thing, which would be the perceptual status of their experience and the reality of what they experience. Before proceeding further, we want to explain how putting “extraordinary perceptions” on the map changes the debates about the nature and role of the sense of reality.

2. Putting extraordinary perception on the map

The case of Paul Vasques is not tapping into the problem of hallucinations (illustrated in Figure 1a): Paul’s retina is receiving sensory signals from the environment, and his

visual experience of the rainbow is what philosophers would call “factive” and accurately represents an existing rainbow. In other words, the experienced object (the seen rainbow) corresponds to a real object in the external world (an existing rainbow).

Yet something differs from a case of canonical perception (Figure 1b). There, we would not have expected Paul to express so much confusion as he is experiencing the rainbow. Where does his confusion come from, and what kind of confusion could it be? One interpretation is that Paul is confused as to whether he is seeing or imagining what he is currently experiencing (Figure 1c). This confusion itself can be interpreted as being a confusion about the contents of his experience: is the rainbow I am experiencing something that is really out there or something that is the product of my imagination? After all, if the contents of his experience are a bit blurry or faded because of the fog, the rainbow could lack the precise or stable contours that characterize classic objects of perception and start to look like an imagined object. Alternatively, however, the confusion may not come from the contents: the rainbow may be perfectly defined in Paul’s experience, and yet some attitude or the feeling toward this experience, or accompanying it, generates an odd impression that makes it feel like an imagining or a dream (see Dokic & Martin, 2017; Matthen, 2010).

The empirical literature on predictive and Bayesian brains provides good explanations for these two types of confusion. If an individual’s experience, even in the presence of incoming stimuli, depends on inner predictions or priors, the brain must monitor whether their experience is genuinely perceptual (meaning that it is about the real world, or in a less representationalist vocabulary, that it is a guide to reality) or imaginary (and for instance conjured up by their own fantasy or dreams) (Figure 1c).

As already mentioned, we suggest a third possibility behind Paul’s confusion, which is not about doubting the perceptual status of his experience. According to this interpretation, the confusion is not at all about whether he is perceiving or imagining. Paul’s experience would be, instead, a case of what we call “extraordinary perception”: he is subjectively aware that he is seeing and not imagining a double rainbow. Nevertheless, this perceptual experience remains unsettling and resists the normal path from perception to belief.

We borrow the term *extraordinary perceptions* from the psychologist Holt-Hansen (1976), who documented them in the context where some multisensory perceptions, though not challenged as being perceptual, continued to feel “out of this world” or strange to the people who were experiencing them in the lab. Confusions arising during multisensory perceptions are a paradigmatic case where the possible dissociation between certainty about a perceptual status and confusion can coexist. Given that every perception, including what we wrongly call “visual perception,” is multisensory, the fact that multisensory perceptions can be extraordinary shows the importance of the category.

Extraordinary perceptions occur when one is absolutely aware of perceiving: Paul is unsettled **as** he sees the rainbow, not uncertain **whether** he sees or imagines it (Figure 1d). Consider what happens when we meet someone we expect to be at the other end of the planet. We are certain that this is, say, Betty, yet our minds also feel dizzy trying to perceptually make sense of Betty being here in front of us. In other words, we are not doubting the ontological status of the object of experience (it is real, and we are perceiving and not making it up), but this reality is still, in a way, confusing. Instead of seeing the confusion as arising only at the level of beliefs—say, as a struggle to believe that Betty is not at the other end of the planet and that it is indeed Betty that we are seeing—we argue that the confusion is already and also present in the perceptual experience itself. Our argument has both conceptual and empirical support. Conceptually, accepting that the confusion is already in the perceptual experience provides a good explanation as to why there should be confusion at the level of beliefs, which can be immediate and create an experiential feeling of unease: some dissonance between what our perceptions asks us to believe and our prior beliefs can lead to cognitive unease, but some unease can also occur in the perceptual experience itself—as when we feel dizzy for a few seconds when we see Betty. Empirically, this experience of extraordinary perception may be treated as a milder, shorter, and more frequent version of the kind of state of confusion, *acute confusion*, widely documented in the clinical literature (see Crammer, 2002, in particular, for the subjective description). This type of confu-

sion is not primarily or solely a disorder of thought but starts with disordered attention and/or awareness, which then gets manifested by diminished speed, clarity, and coherence of thought (Francis & Young, 2012).

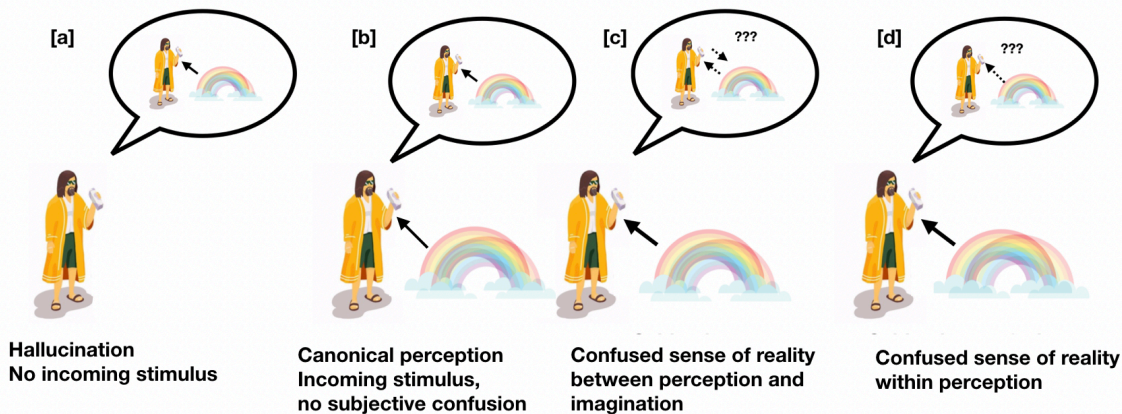


Figure 1: Distinguishing hallucination (1a) from perception (1b) from a third-person point of view comes from the absence or presence of incoming stimuli. From a subjective point of view, the presence of incoming stimuli can be accompanied by no confusion or two kinds of confusion. The first is about whether one is really perceiving or imagining (1c), while the other occurs without questioning the perceptual status of the experience (1d).

In this paper, we focus on the milder, non-clinical confusion and only as it relates to the alteration of the sense of **perceptual** reality (the sense of perceptual reality being here the cognitive function that is disturbed). There may be instances in which one is confused, for example, about their sense of self, but these are not in our focus (more on this below). Our goal in this paper is to define the phenomenon at stake in the kinds of experiences where people are certain of perceiving and yet experience confusion concerning the reality of their perceptual experience and/or the object. While most people see the sense of reality as having a single component, we argue that it must be a composite of several subjective components and a plurality of mechanisms.

What does this entail for Bayesian accounts of the Brain? We do not deny that our brains are challenged to monitor whether and how much an experience is likely to be perceptual and tell us about the external or objective reality rather than being made up by our internal expectations. If a Bayesian inferential process is taking place during perception and draws on both internal as well as external information, the “reality

monitoring” question in cognitive neuroscience is genuine and important (see, e.g., Corlett et al., 2019; Dijkstra, Kok & Fleming, 2021, as well as this special issue).

What we deny is that all cases of confusion involve questioning that one is perceiving the real world. The challenge raised here requires careful unpacking, which will come later in Sections 3 and 4. Already, one can guess that a Bayesian brain, which operates with degrees—of probabilities, certainty, or confidence—does well at capturing graded states but may do less well at explaining the categorical sense that percepts, even confusing ones, present us with a mind-independent reality.

We examine how a Bayesian brain could assign a “real” categorical tag to certain representations (Section 5) and then look at how categoricity can combine with some probabilistic aspects coming from other processing. So, while contents are provided as real, agents must be able to entertain feelings of uncertainty, control, or sensory incongruence to explain states of confusion like Paul’s (or real empirical cases such as experiences in virtual reality or derealization). In other words, we claim that the signature of perceptual reality is a composite. We explain how our composite model works (Section 6) and discuss where it raises problems and generates important predictions (Sections 7 and 8).

3. Two problems for the Bayesian brain

Let us go back to the problem that Paul’s brain must solve when faced with unexpected new retinal evidence: what is the likely cause of a noisy retinal signal? The relation between stimuli and signals is many to many, and each step from stimulus to signal to percept is also noisy. To address the issue, it has become common to think of perceptual processing as solving a Bayesian inference problem. Start with an assignment of prior probabilities to candidate hypotheses about the cause of the signal. For instance, there is more prior likelihood of encountering a single rainbow than a double rainbow. Draw on a likelihood function assigning conditional probabilities to signals given the hypothesized causes. Then compute—following Bayes’ theorem—the posterior probability of the hypotheses.

By now, Bayesian models of perception are well-accepted. However, there is a debate about whether they are indeed just models or whether the processing itself follows Bayesian rules. Arguably, according, for example, to Morrison (2016), the processing that takes place is actually Bayesian, and the resulting experience explicitly represents the posterior probabilities of two rainbows vs. one rainbow happening. The content of perception is here probabilistic. Instrumentalists object and consider Bayesian models only as ideal models, not what really happens in perception (Block, 2018).

A different challenge, however, comes with inferring not just the probable cause of the sensory signal but that there is an external source of the signal in the first instance. Evidence of an important neural overlap between stimulus-dependent and self-generated imagery shows that establishing whether there is an external source is itself a problem that a Bayesian brain needs to solve. In other words, it needs to not only determine whether the agent is experiencing a single or double rainbow but also whether the rainbow (or rainbows) is real or imagined. Predictive coding accounts run into a similar challenge: perceptual inference results from a combination of information provided both by the current sensory input and prior expectations, and only the prediction error is neurally represented (see, e.g., Aitchison & Lengyel, 2017). On the one hand, the kind of generative capacity required for imagining may also have a part to play in regular perception (although an overlap between capacities does not yet determine that identical cognitive processes are at stake). On the other, the generative part involved in perception eventually means that the status of the output as perceptual could become blurred—at least for the subject—with the output of a purely or primarily self-generated imaginative experience. Captured in the Bayesian terms, the problem is to infer whether the source of the neural activity is more likely to be stimulus-dependent or self-generated.

The final result of such inference faces a problem similar to that of the probabilistic nature of perceptual content. One can accept that the reality-monitoring mechanisms of the Bayesian brain underlie a graded, probabilistic sense of reality. Suppose the sense of reality pertains only to the content of experience. Then it can be closely linked to the cases where the contents themselves are probabilistic, if this means that

these contents are somewhat phenomenologically blurry. If the sense of reality is akin to an attitude or feeling that can be dissociated from the clarity of the content, then this attitude or feeling also needs to be graded.

The problem comes with understanding how a graded sense of reality could accommodate extraordinary perceptions—cases that are characterized, as we suggest, by an absolute certainty that we are perceiving and yet continue to be confusing.

4. Extraordinary perceptions: accommodating confusion within subjective certainty of perception

Derealization is a good example here. On its chronic side, derealization can accompany various neurological and psychiatric disorders (Lambert et al., 2002; Simeon, 2004). On its episodic or transient side, derealization can follow sensory deprivation (Reed & Sedman, 1964), extreme stress (Bernat et al., 1998), burn-out, or drug/alcohol abuse (Melges et al., 1974). Individual episodes of derealization in the healthy population are relatively common (see, e.g., Aderibigbe, 2001). They can be subjectively and neurologically distinguished from depersonalization (Dewe et al., 2018) as they manifest themselves as a feeling detached from surroundings rather than one's bodily self. Derealization is also one of the key cases that Dokic and Martin (2017) use to argue for a possible dissociation between the subjective sense of reality and the clarity of the contents of experience. In derealization, the objects and scenes are not experienced as blurry or unclear, yet they continue to lack a strong subjective sense of reality.

The actual cases of derealization draw a complex landscape, where much remains to be empirically documented. The present state of knowledge still recommends, we believe, to be careful about recognizing two categories of states. First, some cases of derealization imply a feeling that one may be dreaming or imagining the content that is otherwise provided very clearly in experience (let us call this Type 1 derealization). Yet other cases (Type 2 derealization) can involve unambiguously identifying one's experience as perceptual and yet still feeling that something is wrong. (Spiegel, 2021, see also Shorvon, 1946; Parnas & Sass, 2001). If we accept

the existence of such Type 2 cases and that the subjective disturbance does not come from the blurriness of the content, how could the sense of reality both clearly indicate that one is having a perceptual experience (and therefore close to 100% certainty that it is not imagination) yet continue to show grades of confusion?

The second question concerns the dissociation between the clarity of the content and the sense of reality. According to Dokic and Martin, and in line with some recent Bayesian accounts, the disturbance in the subjective sense of reality observed in (Type 1) derealization results from lower confidence in the reality of the content: a probabilistic metacognitive inference surfaces in experience as a lower intensity of the accompanying subjective feeling of reality. The reasoning is that, in non-derealized cases, the “normal” sense of reality is sufficiently explained by high confidence in the reality of the content. Here we would object that at the conscious level, the subjective sense of reality can interact with the subjective appearance of the scene or objects: how clearly or determinately something is experienced can impact the sense of reality. We agree that, to go back to Paul’s example, the experienced rainbow can be clear and vivid, and yet the scene is subjectively confusing. But if, at some point, the subjective appearance of the rainbow gets not so vivid and kind of fuzzy (Koenig-Robert & Pearson, 2021; Pearson & Kosslyn, 2015), Paul may be even more confused. The appearance properties cannot fully explain or constitute the categorical feeling of perception as presenting us with something real but may sometimes help establish the source differences, for example, between self-generated, imagined contents and stimulus-driven scenes. This decorrelation remains to be better understood but serves to argue for the existence of distinct mechanisms but also different subjective aspects within the sense of reality.

Importantly, in the case where Paul experiences the rainbow(s) as indeterminate, we would suggest that his confusion is not shattering the clear sense that he is experiencing the real world. Some people cannot tell without glasses whether a blurry object located on the table is an apple or a curled-up parrot taking a nap. However, they can be categorically clear they are perceptually experiencing the scene and even the object. The same is true in other scenarios where the senses are partially impaired, such as poorly illuminated rooms or foggy weather.

So far, we argue that cases of extraordinary perceptions, of which Type 2 derealization would be an instance, require to accommodate two observations:

- (i) Though a lack of clarity in the content is not necessary for a confused sense of reality to surface, it can affect the degree of this confusion.
- (ii) Whether the content is clear or not, it is possible to be in a state of subjective confusion that is not at all suggesting that the experience is anything less than perceptual. More particularly, there is no hint at all that imagination could be involved.

To accommodate these two claims, we suggest that the sense of reality needs to be analyzed as a composite of different subjective markers. As we explain below, one of the subjective markers needs to be a categorical marker of perceptual certainty: something that clearly makes the subject aware that their experience is putting them in touch with reality or, in other words, gives them no room to doubt that they are seeing—and not imagining—the scene in front of them. In addition, other markers of subjective confusion can give rise to more complex varieties of the overall subjective sense of reality within the clearly perceptual experience without affecting the categorical marker.

To clarify the scope of the problem, we focus on “experience” as a snapshot or brief period rather than an extended period (Snowdon, 2010). In other words, we posit that the sense of reality can be explained in terms of what is going on in the brain over a brief amount of time. This is not a clear-cut boundary, as we cannot really consider perception by nanoseconds. In the lab, the examples of such brief experiences are the ones where a single perceptual decision is made after sufficient sensory evidence has been accumulated. This accumulation is, of course, not instant and involves (at least) multiple eye saccades. But we also do not focus on the **extensive** exercise of attention, active exploration with multiple senses, or cross-examination. Perceptual and active checking occurs, but some instances come to confirm or modulate a subjective signature of reality, which can arise relatively fast. As we open our eyes in the morning and have the visual experience of a new hotel room around us, we perceive

it as real and do not need to get up, look at several corners of the room, or touch the bedside table to have a sense of reality.

We also acknowledge that relying on phenomenology remains problematic: personal reports and anecdotes differ, notably when it comes to how things appear and feel as people fall asleep, wake up, experience virtual reality (Cheng et al., 2014) or derealization (Lambert et al., 2001a, b). The very method of capturing how real or odd the environment seems uses scales, and averaging across multiple individuals can explain why reports come in a graded score. Similar to what happens with consciousness, talks of degrees of felt reality or presence comes in handy for measurement purposes but still do not tell a philosopher precisely what is going on (Bayne, Hohwy, and Owen 2016).

These cautionary notes aside, we proceed with the above assumptions, which, we believe, will not offend the common sense of the readers and will seem acceptable to both philosophers and psychologists.

5. Why the subjective sense of reality needs a clear-cut component, and can one be Bayesian about it

Could the mix of certainty and confusion we consider to be characteristic of the cases of extraordinary perceptions, perhaps like that of Paul Vasques, result from a conflict between experience and abstract judgment? After all, one has some knowledge or beliefs about the environment, and when their experience is taken as perceptual and putting one in touch with reality, it could be that what this experience asks them to believe generates some friction with the existing beliefs. It is indeed possible to interpret Paul Vasques' expressions of confusion as a sign that it is difficult for him to believe that there is a double rainbow in front of him. Yet, if Paul Vasques is closer to what we call a Type 2 derealization, he could have no problem believing that the double rainbow is real and still feel not quite right. In derealization, the agent can judge or know that the scene is real (Shorvon, 1946; Parnas and Sass, 2001).

We do not think that the confusion in extraordinary perception is only due to the incongruence between what we see and what we believe/are ready to believe. Of

course, if we see someone who looks like a vampire in the street, smiling with extra-long teeth, because we know that vampires do not exist, we conclude that what appears to be a vampire must be a normal human going to a fancy-dress party. We introduce a difference between having perceived someone with the appearance of a vampire and having perceived a genuine vampire at the level of judgment, but our experience of the vampire-looking person is the same in both cases.

As previously said, the same agent can experience a clear visual content—with no blurriness or indetermination—and be certain that this experience is perceptual and puts them in touch with the real world, yet still feel uncomfortable with their sense of reality. We suggest analyzing the subjective markers of reality in the following way.

Categorical tag of reality in perception. We can feel that what we are experiencing is unambiguously the real world. In other words, we are then subjectively aware of having a perceptual experience (i.e., seeing, hearing, touching, etc.). Moreover, the perceptual nature of our experience is not something we experience with probabilities but rather with certainty—the experience either presents itself as perceptual, or it does not—even though the underlying computational processes that classify our experience as perceptual are probabilistic.

The argument here is not dissimilar to those raised elsewhere against the notion of probabilistic contents of perception. While Block asked, “if perception is probabilistic, why doesn’t it seem probabilistic?” (Block, 2018), we ask: “if perception combines both internal priors and incoming external signal, why doesn’t perception seem to come both from us and the outside world?” Or, to put it more generally, “why does what we see feel absolutely real, and not also probably a bit made up?” The challenge to explain the categorical subjective marker of perception duplicates the challenge raised by Block (2018) about probabilistic contents, but this time concerning the source: if outputs are Bayesian and probabilistic, how can they seem non-probabilistic?

There are, of course, multiple ways to avoid the challenge, for example, by denying that experiences ever come with an absolute subjective mark of reality. However, as noted above, we consider it legitimate to take the categoricity seriously as far as the

reality signature of perception is concerned. Our question is: accepting that there is a challenge, can it be addressed within a Bayesian realist framework? If being a realist about Bayesian models means that our cognitive systems combine priors with externally driven signals, does not this entail that our experience, even when dominated by external signals, will only look real to a very high degree?

A good place to start is with the approach sketched in Clark, Friston, and Wilkinson (2019) in defense of predictive coding or by Gross (2020) in defense of Bayesian realism against the challenge that perceptual contents necessarily end up being probabilistic because the underlying computations are. The authors argue that proponents of the Bayesian brain hypotheses are not limited to the talks about degrees and probabilities when perception is concerned. For example, Clark, Friston, and Wilkinson (2019) argue that specific mid-level perceptual hypotheses can be tagged as “100% certain”. Their solution can address both the fact that contents do not appear probabilistic and that even blurry or indeterminate contents can appear to us as real, that is, come with a subjective tag of reality. It does not matter whether the objects appear to Paul like strange colorful bows or blurry shapes on the horizon or in the periphery of his visual field. However, there must be “100% certainty” in the mid-level hypothesis that one is experiencing the scene and its constituents as real. Gross (2020) also offers resources to explain the reality tag of perceptual experiences outside predictive coding assumptions. As he underscores, the probabilistic nature of Bayesian neural processing is compatible with a selection process that makes only the most probable output available to other systems or levels, such as the agent level of consciousness. What matters in Gross’s proposal is that the selection mechanism means that the probabilistic information is suppressed and not passed on to the next level: this makes the result categorical, and not just highly or most probable.

Both suggestions are compatible with a realist interpretation of Bayesian models, where neural processes follow Bayesian rules. Other mechanisms, having to do, for instance, with action-selection or action guidance, could also contribute to the selection process, but our point is less about the mechanisms and more about the clear-cut perceptual character of the resulting representation. Indeed, as far as predictive

implementations are concerned, the solution suggests discontinuity in the hierarchical processing because of this intermediate selection stage (see also Deroy, 2019). There is something special about perceptual experience if contents are marked as absolutely real at the mid-levels.

Suppose these Bayesian-compatible mechanisms explain how some representations or experiences can be categorically marked as real at one level of processing. How can they also explain the confusion?

6. Why we can be confused while being absolutely certain and right that we are perceiving the real world

What we call the *possibility of perceptual confusion* can be expressed in the following way: while one feels absolutely certain that one's experience is putting them in touch with reality and is perceptual, something else is subjectively confusing the accompanying sense of reality. Here importantly, confusion is separate from and combines with the categorical marker that the experience is strictly perceptual. Again, we are not ruling out that, in some cases, the boundaries between perceptual and imaginary may also be confused, also because of the content. However, we only target cases of perceptual confusion where our experience tells us something is real yet "fishy" (like Type 2 derealization or Paul's state of wonder about the rainbow).

We already mentioned that the blurriness or lack of determinate content could introduce a sense of confusion. We want to add here that the indeterminacy of the content can affect the overall sense of reality within a clear feeling that one is perceiving the real world.

This suggests that the sense of reality in perception is a composite of different subjective markers, giving rise to a feeling that may vary in different ways or across multiple dimensions. The experience in Type 2 derealization can be "fishy" in a phenomenologically different way than Paul's experience with the rainbow, or say, experiences in virtual reality (VR). All of these experiences are characterized by clear

perceptual origin. It is the overall sense of reality within perception that is disturbed, not reality monitoring or the categorical subjective marker of reality.

Providing a complete set of markers that play a role in the composite sense of reality is an open empirical challenge, which requires a careful, detailed examination of the phenomenological aspects of one's perceptual experience, as well as the controlled alteration of factors hypothesized to be relevant. Below we discuss some such potentially relevant factors, such as multiple compatible causal scenarios, feeling of control and metacognitive congruence, and their accompanying subjective effects. Because we are looking at a sense of reality in perceptual experience, we are not considering factors that can only occur at the level of judgment. At the same time, we take perceptual experience here in a broad sense as that which is experienced along with rich perceptual contents and can also include metacognitive or noetic feelings.

For example, as discussed above, Clark, Friston, and Wilkinson (2019) propose that mid-level perceptual hypotheses come with the categorical signature of certainty and, as we read them, of reality. Their concern is partly to explain how this mid-level can be paired with multiple potentially applicable higher-level hypotheses, bearing this time the mark of probabilities and interpretative possibilities. According to them, "when the brain estimates that a suite of mid-level re-codings, couched in terms of features such as redness, roundness, loudness, pulsatingness etc. etc., as highly certain, it can simultaneously compute that this vivid set of (perhaps 100% agent-certain) contents is consistent with multiple ways the real world might be" (Clark, Friston, and Wilkinson, 2019 p. 30). For instance, while Paul could settle on his perceptual experience being that of a double rainbow, such experience would still be compatible with the following alternative hypotheses: "the double rainbow is a natural physical phenomenon" and "the double rainbow is a supernatural sign of God." These two hypotheses do not compete with the subjective certainty that the visual experience is a perception of a real rainbow.

Paul can wonder about another causal aspect of what he is experiencing without stopping to feel that his experience is perceptual. What he wonders about and what

he experiences as real are not the same hypotheses. What Paul experiences as real is “there is a double rainbow in front of me,” while what he doubts is “whether the double rainbow is natural or supernatural.” Such counterfactual explorations or wonderings can be judgements, but a sense of wonder they elicit can also be part of the overall experience (like surprise would be). These feelings are about the world and can occur in ways that do not change how the world perceptually appears. Leaving the mechanism aside, we assume that wondering about, for instance, the natural or supernatural cause of the occurrence of a rainbow in the real world creates a subjective feeling (e.g., sense of wondering) that permeates the perceptual experience of the rainbow and explains part of what constitutes the subjective feeling of confusion within perception. Note that this seems to suggest a degree of cognitive penetrability of perception, which could understandably be unsatisfactory to some readers. However, this interaction between uncertainty regarding higher-level beliefs and the actual perceptual experience is significantly less dramatic than some Bayesian frameworks theoretically allow: it is not the **content** of perception that is altered by this uncertainty but rather a composite sense of reality that co-occurs with perception.

Another good example here is olfactory experiences (see Jraissati & Deroy, 2021): we may find ourselves confused about how to name what we are smelling (is it smoke or coffee?), or our brains could even be unable to categorize the smell, yet we clearly feel that we are experiencing a real smell. The experience of smelling a real smell without making sense of it is still subjectively unsettling. Understanding how confusion about vivid supra-threshold perception can occur while the sense that one is in touch with reality is maintained is the challenge we are pointing at. Whether confusions about near-threshold perception is of the same is not a given, and it is also an important question to raise.

6.1. Feelings of control

Control seems to be another good candidate when it comes to accounting for the composite subjective signature of reality and its various confusing variations. Here, we want to distinguish the actual controllability of the processing within the brain from

the feelings of control, which are how the agent represents their control (often erroneously, as shown in the multiple illusions of control documented in the literature).

We acknowledge that actual lack of control may play a role in tagging something as real and solve the missing selection mechanism posited in Clark, Friston, and Wilkinson (2019). In some cases, at least for predictive coding versions of the Bayesian brain, the neurocognitive mechanisms of control are not so independent from the perceptual mechanisms of causal inference, drawing on predictability and error reduction of the incoming signals (Teufel & Fletcher, 2020), but others argue that they should be distinguished (Ligneul, 2021). Cognitive and motor control are also intertwined with the perceptual inference process, as one can causally intervene on the world to check their predictions, including by voluntary saccades.

We are, however, concerned with the subjective feelings of control as a possible ingredient adding confusion to the categorical subjective marker of reality within perception. An interesting case here comes, for example, from VR: the general awareness that we can switch off from what we are perceptually presented with, for instance, by pressing the “on/off button,” may provide a general feeling of control over what is perceptually happening in front of our eyes and contribute to some of the weirdness of the experiences. We can close our eyes in genuine everyday perception, but we also represent that this will not make the real rainbow cease to occur. Because the objects in the VR system are fictional, their existence depends on our perception, in a way that may also manifest itself as a feeling of control over them.

6.2. Multisensory congruence: direct and metacognitive aspects

This interpretation about a subjective marker of controllability affecting the overall sense of reality in VR remains speculative. Less speculative is the claim from the literature showing that the sense of reality in VR also correlates with interactivity: an agent should experience the changes caused by her actions and movements in the virtual environment. If the coupling between movement and response is good, the agent feels effortlessly immersed in the virtual environment. In contrast, a bad coupling could generate a sense of cognitive effort or uneasiness. When the coupling is bad, the subjective manifestation may not be that of control, but spatial-temporal

incongruence: sensory signals do not come as predicted in a genuine perceptual experience. The head movements, visual, vestibular, and proprioceptive information need to combine in a specific way (see Garzorz & Deroy, 2020). A few milliseconds of delay can cause subjective feelings of dizziness or motion sickness. Even in such low congruity cases, the virtual environment can still be experienced and tagged as perceptually real and not made up by one's mind. Someone can subjectively experience a scene as real and perceived yet have a direct feeling of incongruence (dissonance) or a low metacognitive feeling of multisensory confidence because of the spatio-temporal incongruence between the various signals they receive (Deroy, Spence & Noppeney, 2016).

As perception is through and through multisensory, we suggest that incongruence across the senses and the metacognitive monitoring of this incongruence offer a promising avenue for looking into the confusing feelings surrounding the 'is this here and now?' as they also occur in derealization and experiences in VR (see already Ciaunica et al., under review, for the different case of depersonalization).

Less speculative here is other evidence showing that cross-modal imagery brings a form of confusion to the overall feeling of reality while clearly keeping the absolute subjective certainty that we are having a perceptual experience. For instance, we often have the feeling that we can almost hear the voice coming from a muted person speaking on a TV screen (Spence and Deroy, 2013; Bourguignon et al., 2020). In this case, the auditory imagery is indeed imagined. Here, we do not have a lesser feeling of reality for the visual experience of the person on the TV because of the presence

of an auditory image or a feeling of control. However, it can still create confused feelings about our experience.

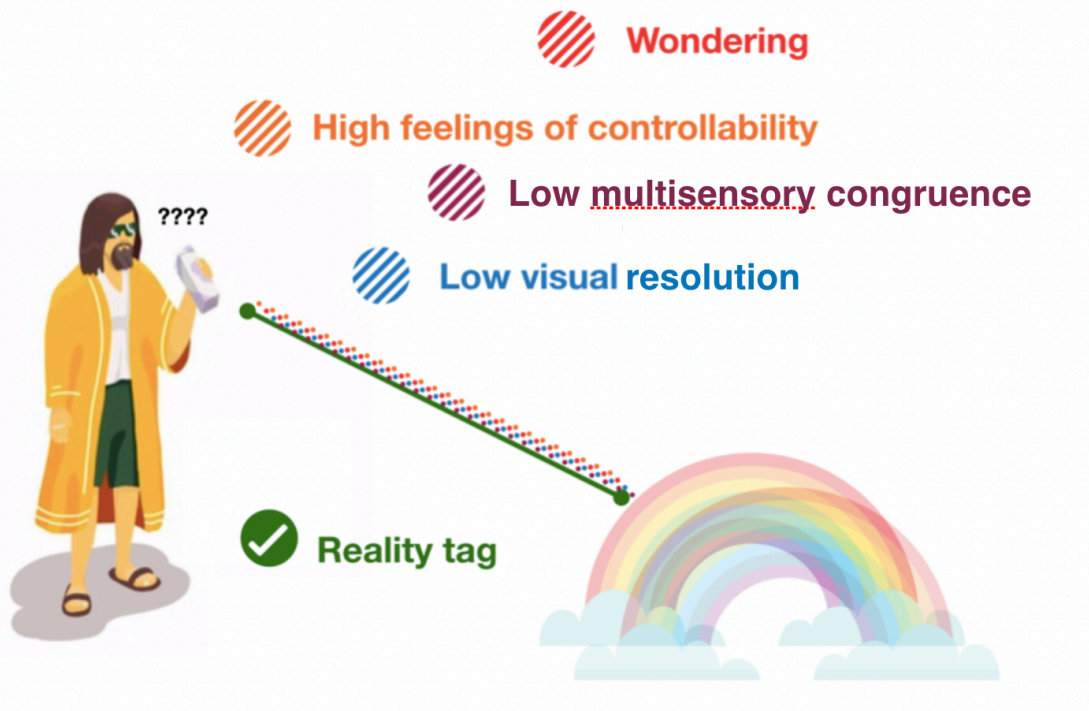


Figure 2: The subjective sense of reality in perception is a composite of subjective markers (in colour): a categorical “reality” tag means that one is certain of having a perceptual experience of a real object, while other subjective feelings can come to bring more or less confusion without shattering the reality tag.

7. The composite sense of reality: Standing questions and implications

7.1. If the sense of reality is a composite, why doesn't it seem so?

According to our analysis, the subjective signature of reality, which appears clearly in everyday perception and can get us confused in some other cases such as derealization, VR, or some extraordinary cases of wonder, happens to be a composite. Our experience can come with a categorical marker that it is real. Other subjective markers—for instance of low control, high multisensory congruence, and sense of wonder—contribute to us having a strong feeling of reality within this clearly perceptual experience.

In cases of extraordinary perception, our experience continues to manifest itself as categorically perceptual and having nothing to do with imagination. However, other subjective markers, such as a residual feeling of control, sensory incongruence, low multisensory congruence, or a feeling of wonder, introduce layers or dimensions of subjective confusion. Crucially, this confusion occurs for us within perception. Sometimes, what is at stake is us losing a smooth sense of spatio-temporal actuality (“here and now”), sometimes having other strange feelings of being able to look at the world in the same way and see something as mysterious. Nevertheless, we are still in a subjective state of perception and not challenging that our experience is putting us in touch with reality.

We are not intending to provide an exhaustive list of the various ingredients and subjective markers underlying our sense of perceptual confusion. However, we insist that various subjective markers eventually play a role on top of mechanisms selecting contents as perceptual/real.

We want to contrast our approach with those saying that the subjective signature of reality is entirely metacognitive. Dokic and Martin (2017), for example, argue that the sense of reality is a metacognitive feeling based on a “set of processes that allow us to distinguish between memories of internally generated events [...] and memories of externally generated events [...]” (p. 304). Our solution, however, favors the idea that multiple subjective markers of reality are provided, through distinct mechanisms, including possibly some non-metacognitive ones. For instance, non-metacognitive processes may produce a sense of wonder, and so can the blurriness or indeterminacy of the perceptual content.

Another difference here consists in accepting that the subjective sense of reality can be disturbed without the contents of experience being in any way indeterminate or blurry, while recognizing that blurry contents can have a possible effect on the general subjective feeling of reality. Here, in particular, we should not generalize from the dissociation between the sense of reality and determinacy of contents in the clinical population affected by derealization to the rest of the population.

Still, we agree that our composite approach faces a “phenomenal unity” problem that a fully metacognitive account like Dokic and Martin (2017) may not. In our account, even in a typical perceptual case, our subjective sense of reality is a blend of several subjective markers. At the same time—phenomenologically—most authors seem to have accepted that feeling real is all one well-unified thing. Some people claim that our conscious perceptual experience is unified across modalities, but this phenomenological evidence can be challenged based on experimental results (Spence and Bayne, 2015; Deroy, Chen & Spence, 2016). Perhaps the sense of reality we enjoy in everyday perception is just a familiar mixture, but still a mixture of different markers.

Pace recent accounts, we could ask if our sense of reality in perception is, after all, that unified. Reports collected from patients suffering from hallucinations show that they distinguish about seven different aspects of the sense of reality that infuses the hallucinated object (Aggernaes, 1972; see also the discussion in Farkas, 2014), while some phenomenological accounts do not shy away from breaking the sense of reality in multiple subjective aspects or components (six, in Jaspers, 1959/1997). Our account also claims that different aspects or dimensions compose our overall sense of reality even once one value (reality as source) is fixed. What differences exist between analyzing the sense of reality as a composite of different aspects or in the dimensional approach remains a topic for further investigation. One difference would be that a dimensional account could explain that specific dimensions are wholly independent (e.g., wonder and control do not seem to co-vary), while others can be more dependent (e.g., multisensory congruence and unisensory confidence).

7.2 Sense of reality is robust and breaks down in different ways

An interesting feature of the above proposal is that the sense of reality remains related to the determinacy of perceptual contents but only indirectly. To an extent, they can mutually influence each other and help the agent navigate the complex challenge of keeping track of what is real. The functional redundancy is rather useful and can make the sense of reality more robust: not throwing the agent into doubt about the reality of their own experience for minor disturbances, while using determinacy as a useful cue to ascertain whether they are really in touch with the external world.

Similarly, in our account, because the subjective signature of reality is also a composite of various markers or “dimensions,” each ingredient informs the agent about subtly different things. The complementarity (rather than redundancy this time) is also beneficial. While some categorical markers inform the agent that their experience is perceptual and not imagined, other subjective markers of reality can create confusion through feelings of control, multisensory incongruence, and wonder, each bringing in different variations in the overall sense of perceptual confusion and encouraging different forms of attention, active checks, or cognitive explorations.

7.3. Implications

Our composite account of the subjective signature of perception in the Bayesian brain suggests that the presence of a subjective categorical marker of perceptual reality can (and evolutionarily should) be robust and early in development but also independent of metacognitive abilities. Animals with Bayesian brains and no metacognitive capacities should be able to tag experiences as real and indeed may not be as easily confused about perception or even exercise complex subjective reality monitoring.

Still, if metacognitive processes and multisensory congruence develop in the first few years of life, so the sense of perceptual reality should also become richer but also more confusing at times, as infants and children develop (see Goupil & Kouider, 2019). Similarly, the capacity to entertain counterfactual hypotheses (e.g., Nyhout & Ganea, 2019a, b) and therefore add a subjective sense of wonder or causal depth to the same percept should also come later and independently in the development, as counterfactual capacities start to shape. However, it does not imply that young children do not have any subjective sense of reality, but merely that some ingredients of their sense of perceptual reality are not yet adequately established.

Another set of implications concerns the multiple ways in which our way of monitoring experiences, such that they can be both perceptual and confusing, could break down or dysfunction. Not minimizing the importance of looking at how we subjectively feel our experiences as more or less real or imagined, we expect that some clinical or sub-clinical cases could be better examined as dysfunction that does not challenge

the capacity to feel that one perceives the real world and yet comes with other confusions.

8. Conclusions

We have focused on explaining both categorical and confusing aspects of our subjective sense of reality as we perceive the world as external. The scenes we perceive strike us as real, not just probably real. However, in some cases, such as virtual reality, derealization, and other kinds of extraordinary perceptions, there is still some confusion within a perceptual sense of reality.

We suggest that the Bayesian brain hypothesis should accommodate both the categorical aspects of this signature of reality in perception and the various confusions that perception also welcomes. This proposal makes the subjective signature of reality a composite. We acknowledge this as a new phenomenal unity problem: how can the categorical signature combine with other graded subjective markers, notably feelings of congruence, feelings of control, and wonder? The question of whether the subjective sense of reality manifests itself as a composite or as unified but multi-dimensional is hard to resolve at the phenomenological level, and the main challenge remains to better understand the underlying mechanisms which contribute to the overall feeling of reality in perception and its disturbances.

Our proposal mostly amounts to raising new prospects and challenges for the Bayesian brain models, which could be examined through developmental, experimental, and clinical studies. Our composite model points to the importance of accounting for the robustness of our subjective sense of reality and its capacity to guide our reality checks in multiple directions. It also makes new predictions regarding the non-linear development and breakdowns of the subjective sense of reality.

References

- Aderibigbe, Y. A., Bloch, R. M., & Walker, W. R. (2001). Prevalence of depersonalisation and derealisation experiences in a rural population. *Social Psychiatry and Psychiatric Epidemiology*, 36(2), 63-69.
- Aggernaes, A. (1972) "The experienced reality of hallucinations and other psychological phenomena". *Acta Psychiatrica Scandinavica*, 48, 220-238,
- Aitchison, L., & Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, 46, 219-227.
- Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are there levels of consciousness? *Trends in Cognitive Sciences*, 20(6), 405-413.
- Bernat, J. A., Ronfeldt, H. M., Calhoun, K. S., & Arias, I. (1998). Prevalence of traumatic events and peritraumatic predictors of posttraumatic stress symptoms in a non-clinical sample of college students. *Journal of Traumatic Stress*, 11(4), 645-664.
- Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1755), 20170341.
- Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, 40(5), 1053-1065.
- Cheng, L. K., Chieng, M. H., & Chieng, W. H. (2014). Measuring virtual experience in a three-dimensional virtual reality interactive simulator environment: a structural equation modelling approach. *Virtual Reality*, 18(3), 173-188.
- Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, 26(9-10), 19-33.
- Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers III, A. R. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences*, 23(2), 114-127.
- Crammer, J. L. (2002). Subjective experience of a confusional state. *The British Journal of Psychiatry*, 180(1), 71-75.
- Deroy, O. (2019). Predictions do not entail cognitive penetration: "Racial" biases in predictive models of perception. In *The Philosophy of Perception* (pp. 235-248). De Gruyter.
- Deroy, O., Chen, Y. C., & Spence, C. (2014). Multisensory constraints on awareness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1641), 20130207.
- Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in multisensory perception. *Trends in Cognitive Sciences*, 20(10), 736-747.
- Dewe, H., Watson, D. G., Kessler, K., & Braithwaite, J. J. (2018). The depersonalized brain: New evidence supporting a distinction between depersonalization and derealization from discrete patterns of autonomic suppression observed in a non-clinical sample. *Consciousness and cognition*, 63, 29-46.

- Dijkstra, N., Kok, P., & Fleming, S. M. (2021, May 19). Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. <https://doi.org/10.31234/osf.io/zngeq>
- Dokic, J., & Martin, J. R. (2017). Felt reality and the opacity of perception. *Topoi*, 36(2), 299-309.
- Farkas, K. (2014) A sense of reality. In Fiona MacPherson & Dimitris Platchias (eds.), *Hallucinations*. MIT Press. pp. 399-417 (2014)
- Francis, J., & Young, G. B. (2012). Diagnosis of delirium and confusional states. *UpToDate: Waltham, MA*.
- Garzorz, I., & Deroy, O. (2020). Why There Is a Vestibular Sense, or How Metacognition Individuates the Senses. *Multisensory Research*, 1, 1-20.
- Goupil, L., & Kouider, S. (2019). Developing a reflective mind: from core metacognition to explicit self-reflection. *Current Directions in Psychological Science*, 28(4), 403-408.
- Gross, S. (2020). Probabilistic representations in perception: Are there any, and what would they be? *Mind & Language*, 35(3), 377-389.
- Holt-Hansen, K. (1976). Extraordinary experiences during cross-modal perception. *Perceptual and Motor Skills*, 43(3_suppl), 1023-1027.
- Jaspers, Karl (1959/1997). *General Psychopathology*. Translated by J. Hoenig, Marian W. Hamilton. Reprint. Johns Hopkins University Press
- Jraissati, Y., & Deroy, O. (2021). Categorizing smells: A localist approach. *Cognitive Science*, 45(1), e12930.
- Koenig-Robert, R., & Pearson, J. (2021). Why do imagery and perception look and feel so different?. *Philosophical Transactions of the Royal Society B*, 376(1817), 20190703.
- Lambert, M. V., Senior, C., Phillips, M. L., Sierra, M., & al, e. (2001a). Visual imagery and depersonalisation. *Psychopathology*, 34(5), 259-64.
- Lambert, M. V., Senior, C., Fewtrell, W. D., Phillips, M. L., & David, A. S. (2001b). Primary and secondary depersonalisation disorder: a psychometric study. *Journal of Affective Disorders*, 63(1-3), 249-256.
- Lambert, M. V., Sierra, M., Phillips, M. L., & David, A. S. (2002). The spectrum of organic depersonalisation: a review plus four new cases. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 14(2), 141-154.
- Ligneul, R. (2021). Prediction or causation? Towards a redefinition of task controllability. *Trends in Cognitive Sciences*, 25(6), 431-433.
- Matthen, M. (2010). Two visual systems and the feeling of presence. In N. Gangopadhyay, M. Madary & F. Spicer (Eds.) *Perception, action, and consciousness: sensorimotor dynamics and two visual systems*. Oxford University Press, Oxford.
- Melges, F. T., Tinklenberg, J. R., Deardorff, C. M., Davies, N. H., Anderson, R. E., & Owen, C. A. (1974). Temporal disorganisation and delusional-like ideation: Processes induced by hashish and alcohol. *Archives of General Psychiatry*, 30(6), 855-861.

- Morrison, J. (2016). Perceptual confidence. *Analytic Philosophy*, 57(1), 15-48.
- Nyhout, A., & Ganea, P. A. (2019a). Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition*, 183, 57-66.
- Nyhout, A., & Ganea, P. A. (2019). The development of the counterfactual imagination. *Child Development Perspectives*, 13(4), 254-259.
- Parnas, J., & Sass, L. A. (2001). Self, solipsism, and schizophrenic delusions. *Philosophy, Psychiatry, & Psychology*, 8(2), 101-120.
- Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, 112(33), 10089-10092.
- Reed, G. F., & Sedman, G. (1964). Personality and depersonalisation under sensory deprivation conditions. *Perceptual and Motor Skills*, 18(2), 659-660.
- Shorvon, H. J. (1946). The depersonalisation syndrome. *Proceedings of the Royal Society of Medicine*, 39(779), 37.
- Simeon, D. (2004). Depersonalisation disorder. *CNS Drugs*, 18(6), 343-354.
- Snowdon, P. (2010). On the What-it-is-like-ness of Experience. *The Southern Journal of Philosophy*, 48(1), 8-27.
- Spence, C., & Bayne, T. (2015). Is consciousness multisensory? In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its Modalities* (p. 95–132). Oxford University Press.
- Spence, C., & Deroy, O. (2013). Crossmodal mental imagery. *Multisensory imagery*, 157-183.
- Spiegel, D. (2021). Depersonalization/derealization disorder. In *MSD Manual Professional Version* (March 2021 version). <https://www.msmanuals.com/professional/psychiatric-disorders/dissociative-disorders/depersonalization-derealization-disorder>.
- Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews Neuroscience*, 21(4), 231-242.

Chapter 4

Paper 3. Counterfactual cognition and psychosis: adding complexity to predictive processing accounts

This paper has been published as:


Rappe, S. & Wilkinson S. (2022). Counterfactual cognition and psychosis: adding complexity to predictive processing accounts. *Philosophical Psychology*, 1-24. <https://doi.org/10.1080/09515089.2022.2054789>.

Author contributions: S.R. and S.W. conceived of the research idea and developed the arguments presented in the paper. S.R. drafted most of the manuscript and revised it for publication with the help of S.W.

ORIGINAL PAPER



Counterfactual cognition and psychosis: adding complexity to predictive processing accounts

Sofia Rappe ^{a,b} and Sam Wilkinson^c

^aFaculty of Philosophy, Ludwig-Maximilians-Universität München, München, Germany; ^bGraduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, München, Germany; ^cDepartment of Sociology, Philosophy, and Anthropology, University of Exeter, Exeter, UK

ABSTRACT

Over the last decade or so, several researchers have considered the predictive processing framework (PPF) to be a useful perspective from which to shed some much-needed light on the mechanisms behind psychosis. Most approaches to psychosis within PPF come down to the idea of the “atypical” brain generating inaccurate hypotheses that the “typical” brain does not generate, either due to a systematic top-down processing bias or more general precision weighting breakdown. Strong at explaining common individual symptoms of psychosis, such approaches face some issues when we look at a more general clinical picture. In this paper, we propose an update on the current accounts of psychosis based on the realization that a neurotypical brain constantly generates non-actual, de-coupled, counterfactual hypotheses as part of healthy cognition. We suggest that what is going on in psychosis, at least in some cases, is not so much a generation of erroneous hypotheses, but rather an inability to correctly use the counterfactual ones. This updated view casts “accurate” cognition as more fragile and delicate, but also closes the gap between psychosis and typical cognition.

ARTICLE HISTORY

Received 26 October 2021
Accepted 11 March 2022

KEYWORDS

Psychosis; counterfactuals; delusions; hallucinations; predictive processing; reality monitoring

1. Introduction

Psychosis is a puzzling phenomenon. It involves having inaccurate and often strange beliefs and perceptual experiences. It is not clear, and certainly not from a pre-theoretic, un-scientific perspective, where this disconnection from consensus reality comes from, why it happens, what causes it. Over the last decade or so, several theorists have considered the predictive processing framework (PPF)¹ to be a useful perspective from which to shed some much-needed light on the mechanisms behind psychosis. We fully share this optimism. The PPF has a lot in its favor. It provides clear sources of potential problems for the functioning of a cognitive system (predictions, priors, and prediction errors with associated precision weightings). It also

CONTACT Sofia Rappe  Sofia.Rappe@campus.lmu.de  Faculty of Philosophy, Philosophy of Science, and the Study of Religion, Ludwig-Maximilians-Universität München, Ludwigstraße 31, Munich 80539, Germany

© 2022 Informa UK Limited, trading as Taylor & Francis Group

shows promise in tying neurobiology and neurochemistry (especially the role of neurotransmitters like dopamine, see, e.g., Corlett et al., 2009) to computational aspects of cognition, as well as dovetailing nicely with certain phenomenological features of experience (e.g., Ratcliffe, 2013, 2017).

Our aim in this paper is to update existing accounts of psychosis by helping ourselves to one recent innovation concerning neurotypical cognition, namely, the realization that healthy, daily cognition is suffused with counterfactual hypotheses, and to apply it to thinking about psychosis in the PPF. To state our claim plainly: whereas standard accounts take psychosis to involve the “atypical” brain generating inaccurate hypotheses that the neurotypical brain does not generate, we explore the idea that the neurotypical brain is actually constantly generating inaccurate, de-coupled, counterfactual hypotheses, and that what is going on in psychosis (at least sometimes) is more helpfully construed as an inability to distinguish the factual from the counterfactual hypotheses (or, perhaps more accurately, a failure to appropriately use the counterfactual hypotheses as they should be used). In other words, psychosis could sometimes be less about generating inaccurate hypotheses *de novo*, as existing PPF accounts have suggested, and more about wrongly identifying or using counterfactual hypotheses as such. Paradoxically, it is these counterfactual departures from reality, that give our experience of reality its *counterfactual depth*, which contributes to the sense of reality.

This updated view has a couple of important consequences. First, it paints a picture where the brain of an individual in a state of psychosis is more similar to the brain of the individual who is not in such a state, since human cognition is rife with unreal hypothesizing. Second, it fits better with the phenomenology of psychosis, its subtlety and heterogeneity, and also in the way that hallucinations are not simply like normal perceptual experiences that happen to be inaccurate (and delusions are not simply like normal beliefs that happen to be false): they are phenomenologically more exotic and unfamiliar than that (Humpston & Broome, 2016; Ratcliffe, 2017).

We proceed as follows. We start by presenting existing predictive processing accounts of psychosis and discuss their virtues (section 2). Then we introduce the notion of *counterfactual depth* in theoretical work on cognition in general (section 3). We further explore the idea that psychosis can be understood in terms of failures to recognize or use counterfactual hypotheses as such and discuss four distinct types of breakdowns in the counterfactual depth and how they may produce the symptoms associated with psychosis (section 4). We argue that failures of reality monitoring result not only in the taking (by the brain) of inaccurate, counterfactual hypotheses to be accurate and factual, and hence to feature as the primary rather than auxiliary drivers of experience and belief (i.e., hallucinations and delusions), but also erode the structure of experience, because the auxiliary

counterfactual hypotheses are no longer playing that role but also because such failure results in various subjective reality markers being misapplied. This fits nicely with the idea that what Ratcliffe calls *real hallucinations* are not simply like normal perceptual states that are inaccurate: they are new and unfamiliar kinds of states. It also fits with the uncanny, but pre-hallucinatory aspects of the psychosis prodrome: reality looks flat or strange (Ratcliffe, 2013), experience has a lost or altered counterfactual depth. We conclude by outlining some implications of our approach and future directions (section 5).

2. Existing predictive processing accounts of psychosis

The predictive processing framework (PPF) for thinking about cognition, perception, and action has recently gained a lot of attention in computational psychiatry, with PPF-based models being proposed in relation to anxiety (Chekroud, 2015), depression (Barrett et al., 2016; Stephan et al., 2016), PTSD (Wilkinson et al., 2017), autism (Lawson et al., 2014; Pellicano & Burr, 2012; Van de Cruys et al., 2014), schizophrenia (Adams et al., 2013; Fletcher & Frith, 2009; Horga et al., 2014), and general accounts of emotion (Miller & Clark, 2018; Seth, 2013; Smith et al., 2019).

The popularity of the PPF in psychiatric research does not come as a surprise; the framework, even in its most basic form, provides at least three distinct sources of potential problems for the functioning of a cognitive system (namely, predictions, priors, and prediction errors with associated precision weightings), setting a clear direction for addressing the first of the two main questions of computational psychiatry: Given a specific model of the mind, what could possibly go wrong (Ford et al., 2014; Huys et al., 2016)? On the other hand, problems with the above elements of the framework map well onto various psychopathologies, providing simple potential answers to the second question: How can a specific disorder be explained using the model at hand (Wilkinson, 2014)? Furthermore, the PPF is increasingly conceived by its proponents as a general paradigm about brain functioning. Hence, any explanation of psychopathology in PPF terms automatically fits within a much broader research program.

Such symptoms of psychosis as delusions and hallucinations are probably the most explored psychopathological phenomena within the PPF. An early account of psychosis in predictive processing can be found in Fletcher and Frith's paper (Fletcher & Frith, 2009), which sketched a hierarchically arranged prediction error minimizing architecture. The main point of this model was to show how one basic mechanism can account for both delusion-like (belief-like) phenomena and hallucination-like (experience-like) phenomena. According to Fletcher and Frith, both hallucinations and delusions present erroneously selected winning hypotheses, due to excessive

prediction error signaling, while the difference between the two types of phenomena is a question of degree determined by where in the hierarchy the mis-selection occurs. The “higher up” the hypotheses are, the more belief-like they are (e.g., delusions); the “lower down” they are, the more experience-like they are (e.g., visual hallucinations or voice-hearing) (see, Adams et al., 2013; Fletcher & Frith, 2009; Horga et al., 2014). In other words, hallucinations and delusions are treated as a problem of inference with a single cause – associating too little precision with sensory information (/too much with predictions), which results in the selection of a “wrong” hypothesis about the world. Depending on where in the system erroneous selection occurs, it may be experienced, for example, as voice-hearing, visual hallucinations, delusions, etc.

This idea that psychosis is the result of excessive prediction error in the system or, to similar results, excessive precision-weighting on prediction errors is common to many accounts of psychosis (see, Sterzer et al., 2018 for a nice review of PPF and excessive prediction error in psychosis) and has some empirical support. For example, there is evidence that certain conscious effects that are explained in terms of prediction error minimization are experienced less or differently in people with diagnoses of schizophrenia (see, e.g., the hollow mask illusion in Dima et al., 2009). Other evidence that supports this hypothesis comes from eye tracking data, namely, impaired tracking during visual occlusion (Hong et al., 2005), impaired repetition learning (Avila et al., 2006), and “paradoxical improvement” (Adams et al., 2013), where clinical populations are better at responding to sudden changes of direction in visual tracking targets. All of this points to a general over-reliance on bottom-up prediction error, and less reliance on the top-down predictions. Furthermore, these approaches provide plausible stories about why delusions and hallucinations co-occur (similar bottom-down processing “bias”) and why delusions can arise both due to biological factors and life events, since both can impact on predictive processing mechanisms (Wilkinson, 2014).

Having said this, they have some outstanding issues (see, Sterzer et al., 2018 for a full review). First, delusions and hallucinations co-occur to varying degrees in different cases of psychosis, and not as often as the traditional approach might suggest. Second, the persistence of delusions in psychosis may be tricky to accommodate by precision estimation breakdowns, especially the kind that result in a bottom-up processing bias. It is a defining feature of delusions that they persist despite contradicting evidence. This suggests an excessive influence of delusional beliefs on the perception of new information, which would entail an increased precision of delusion-related priors (in direct contradiction with the excessive error signaling proposal). Additionally, given that the hypotheses are updated on many conditionally independent levels simultaneously, arriving at (and

sustaining) a drastically wrong hypothesis about the world requires a fairly large breakdown in precision estimation machinery. It is a common assumption in PPF that different parts of the generative model are deeply inter-connected and integrated. This, of course, does not imply unconstrained holism: the network reflects the inferred causal structure of the world, which imposes certain constraints.² Still, a specific story must be told about how these constraints would prevent an individual suffering a radical systematic precision estimation breakdown from losing the ability to function in the world at all, as selecting a hypothesis leading to persistent delusions and hallucinations could possibly affect the rest of the system.

Third, this approach to psychosis is not always consistent with the PPF-based explanations of other kinds of symptoms that may co-occur with it. For example, many features associated with autistic perception, are often attributed to an imbalance of precision ascribed to sensory evidence relative to prior beliefs toward bottom-up processing (that is, having too much precision on sensory evidence) (see, e.g, Pellicano & Burr, 2012; Van de Cruys et al., 2014). This paradigm in autism has some empirical support, such as increased visual cortical activation and decreased prefrontal activation in participants with autism (Lee et al., 2007; Manjaly et al., 2007). This is consistent with the increased bottom-up visual processing, which corresponds to precision weighting skewed toward sensory signal. Yet, although autistic behavior sometimes can co-occur with psychosis, this is rather unusual (Larson et al., 2017).

Finally, when it comes to hallucinations, an alternative view, in which hallucinations occur due to enhanced rather than weakened top-down predictive signaling, has also been proposed (Corlett et al., 2019; K. J. Friston, 2005). This approach suggests that perception would rely less on the sensory input and more on the prior beliefs, a claim even more problematic for explaining the occasional co-existence of psychotic and autistic symptoms in one individual. To complicate matters, some evidence indirectly supports this reverse approach to hallucinations. For example, it was shown that people who hear voices are more susceptible to conditioning-induced hallucinations (Powers et al., 2017) and that hallucinations in schizophrenia patients correlate with a top-down perceptual bias in auditory tasks (Cassidy et al., 2018).

What is relevant is that all the accounts discussed above take psychosis to involve the generation of hypotheses that are inaccurate portrayals of the world, hypotheses that do not feature in the brains of those who are not in states of psychosis. This core commonality across several different accounts of psychosis is precisely what we would like to question to update the PPF treatment of psychosis. Whereas standard accounts take psychosis to involve the “atypical” brain generating inaccurate hypotheses that the neurotypical brain does not generate, we explore the idea that the neurotypical

brain actually constantly generates inaccurate, de-coupled, counterfactual hypotheses, and that this is an integral part of the rich tapestry of healthy cognition. This is consistent with the evidence that psychosis-like experiences are much more common than we typically think (McGrath et al., 2015). So, perhaps, at least sometimes, what matters is how the mind *treats* and *uses* such “inaccurate” hypotheses and how the treatment differs in the pathological cases. In the next section, we argue that generation of inaccurate hypotheses is indeed a crucial feature of our cognition as it heavily relies on counterfactuals everywhere from the lowest levels of perception to intentional, conscious reasoning.

3. Predictive processing, counterfactual depth, and offline cognition

The focus of our paper is on the difference between cognition in psychosis and neurotypical cognition. In this section we discuss the latter. So, putting psychosis to one side, we would like to draw attention to the developments in predictive processing that go beyond the earliest versions of PPF and, specifically, to one recent feature, namely, the emphasis on the rich “counterfactual depth” of the generative models (Seth, 2014; Wilkinson, 2020, 2021).

In the literature related to PPF, the word “counterfactual” is mostly used not in the strict linguistic sense but in relation to the capacity to form hypotheses about the non-factual, about past and present possibilities, as well as about other possible (or even impossible) worlds. In such cases, the notion of counterfactual hypotheses often also refers simply to the hypotheses that present alternative possibilities that are mutually exclusive. This means that at least some of them do not correspond to the actual state of affairs (they are *counter to the facts*), although which ones are such may not be known from an agent’s perspective (see, e.g. Clark et al., 2019). In other cases (see, e.g., Corcoran et al., 2020), the hypotheses may pertain to the states the agent could possibly find herself in *if* she were to act in a certain way. Philosophers and cognitive scientists have emphasized the importance of such (broadly construed) counterfactuals in human cognition long before the PPF. Counterfactual reasoning is thought to be central to planning, decision-making, and intentional, goal-oriented behavior more broadly (see, Byrne, 2016). When it comes to predictive architectures specifically, counterfactual reasoning may be what underlies the human ability to learn priors for decision making in the absence of direct feedback (Zylberberg et al., 2018). Generation of counterfactual alternatives may also be implicated in imagination. Here, a common idea is that imagination has emerged as a way of predicting consequences of anticipated possible actions (Burr & Jones, 2016; Friston et al., 2012; Seth, 2014). “Since many actions will be mutually exclusive, many such representations will inevitably be about

merely fictional circumstances, representing possible sensory consequences of actions that never occur” (M. Jones & Wilkinson, 2020).³ Corcoran et al. (2020) further propose a counterfactual active inference model that allows an agent to evaluate a variety of different contexts before settling on a specific action by reflecting on previous actions (“retrospective” inference) or imagining possible future scenarios (“prospective” inference). Although their discussion is shaped through the lens of the free energy principle (see, e.g., Friston, 2009, 2010) not touched upon in this paper, similar observations hold for the standard predictive processing formulation.

Less intuitively, but consistent with the perception-cognition continuity in PPF, rich counterfactuality may be not only a property of our conscious reasoning, decision-making, and imagination, but lower-level processes, such as those that generate perception and perceptual phenomenology. For example, Seth (2014) (building on Noe’s 2006 notion of sensorimotor contingencies) argues that when the subject is engaged in “factual”, actual world-directed experience, that experience is the way it is, has “perceptual presence” (Noe, 2006), in part due to the activation of a range of counterfactual predictions within the generative model. Indeed, Seth claims, a lack of such counterfactual underpinning of the generative model leads to some atypical perceptual phenomena such as synesthesia, which notably lack this presence, and hence do not feel real. The same could be also said about after-images. If you have looked directly into a bright light bulb, when you look elsewhere, you may see an after-image, but the fact that it occurs in the same patch of your visual field wherever you look is one of several things that tells you that it is not a real thing in the world, but an anomalous product of your visual system.

Wilkinson (2020) builds on Seth’s account and draws on observations from virtual reality research (e.g., Meehan et al., 2003), arguing that different kinds of counterfactual predictions account for different aspects of perceptual experiences. In particular, *object-active* predictions are crucial for perceiving objects as having volumetric content, whereas *agent-active* predictions are involved in the subjective experience of presence associated with perceiving the world, of taking what one perceives to be real and present to us, and of us as present in the world.

In other words, it seems very plausible to suggest that “healthy” cognition involves far more, and constant, counterfactual hypothesizing across the generative model, which stands in contrast with the previous more minimal versions of predictive processing, which emphasize processing efficiency and consistency of winning hypotheses across the generative model. Instead, on the updated view, our experience of the world is underpinned by *counterfactual depth*. The fact that we experience a real world at all (e.g., a real apple, in front of me), and that we experience parts of it in the way we do (e.g., as a real apple rather than a fake one) is grounded in a suite of

counterfactual hypotheses that need never be tested or actualized (e.g., concerning what would happen if we were to bite into it, and the expectation that we could, even if we never in fact end up doing so).

When it comes to explaining psychosis, this realization tempts us to shift the focus from generation of inaccurate hypotheses by an “atypical” brain, toward the idea that the neurotypical brain is constantly generating inaccurate, de-coupled, counterfactual hypotheses, and that what is going on in psychosis (at least sometimes) is an inability to distinguish the factual from the counterfactual hypotheses. The question then is how and why such misidentification or mismanagement occurs. We argue that the counterfactual richness of the predictive mind requires additional mechanisms of monitoring, both when it comes to entertaining the alternative scenarios about how the world could be but also how the world could have been (but is not). Such mechanisms of monitoring may present an added source of vulnerability in the predictive brain that is not fully captured by precision weighting imbalances.

Clark et al. (2019) argue that the commitment of PPF to the deep architecture of the generative models and latent variables (“hidden causes”) allows for counterfactual reasoning to naturally arise in predictive agents. Specifically, they propose that one’s posterior beliefs at intermediate levels of the hierarchical generative model may be estimated as highly certain in a way that leaves room for them to be paired with multiple potentially applicable higher-level hypotheses. For example, the hypothesis “it is snowing” may be compatible with both “the snow is H₂O” and “the snow is synthetic”. Being possible alternatives, these two hypotheses do not have the same degree of certainty as “it is snowing”, and this is precisely what allows one to have doubts about their experience and entertain alternative causal scenarios.

However, such openness to causal alternatives on its own does not provide the mechanism of further counterfactual exploration, which may require simulation – redistribution of precision weighting in a way that allows the system to treat the counterfactual parts of the model as if these hypotheses were indeed selected as the winning ones in order to generate rich counterfactual spaces. The case for simulation becomes even stronger for the task of generating counterfactuals **in the strict linguistic sense**, that is the kind of hypotheses explicitly incompatible with the current state of the system but rather related to how the world could have been but is not, and also for the kind of retroactive and prospective inference discussed by Corcoran et al. (2020). Such hypothesizing presents an example of what Hoerl and McCormack (2019) call *temporal reasoning*. As they note, connectionist architectures generally have problems with the tasks that require temporal reasoning because they do not explicitly represent change, but rather simply update representations as new information comes along

(Hoerl & McCormack, 2019). In the absence of explicit representation of the past states of the system, direct simulation (Corcoran et al., 2020; Knight & Grabowecy, 1995) becomes a very likely candidate for the mechanism of counterfactual exploration. Here, by simulation we do not mean offline model updating that merely disregards sensory input (the feature often taken to be among those differentiating between cognition and perception). Rather, we suggest that, in the processes of both perception and cognition, the system is exploring, “trying on” different generative sub-models by altering the relevant weights. Such “role-play” requires keeping track of the parts of the model that pertain to the actual, as opposed to some possible, world.⁴

In the literature, the process of differentiating between what is real from what is not is often referred to as *reality monitoring*. There is no consensus as to which variables contribute to reality monitoring or what features reality monitoring specifically tracks. As we discuss in more detail in [section 5.1](#), reality monitoring is typically understood as monitoring the source of one’s experience (is it dependent on external stimuli or is it entirely self-generated?) – a task especially important for accounts such as predictive processing in which perception is a constructive process that has a significant top down component. Such monitoring is often taken to be metacognitive, that is, a kind of second-order process integrating multiple types of information and manifesting as a sense (or feeling) of reality (see, e.g., Dokic & Martin, 2017). However, as argued by Deroy and Rappe ([under review](#)), not all processes related to reality monitoring are necessarily aimed toward establishing the source of one’s experience. For example, the experiences such as those in derealization or virtual reality are recognized by the agent as perceptual (coming from outside of the agent) but are characterized by distinct subjective reality signatures (namely, are experienced as not being “quite” real). When it comes to counterfactual reasoning, the focus shifts from monitoring the source of experiences (hypotheses) to monitoring the “actuality” status of the relevant generative (sub)model; that is, rather than caring about the distinction between perception and cognition/imagination (from world vs. from me), we care about the distinction between actual vs. counterfactual world models (which include not only here-and-now perceptual experiences, but general knowledge about the world). This distinction between “actuality” and perceptual reality monitoring, at least theoretically, leads to two different monitoring goals, although they may be accomplished by largely overlapping mechanisms. For example, metacognitive reality monitoring may by default fulfil the role of tracking the actual world model, while various subjective markers of reality may ground the agent in the perceptual experience during the exploration of counterfactual scenarios. Here, we do not aim to provide the specific mechanisms of reality (or actuality) monitoring for counterfactual

exploration (this is a whole research field in its own right), yet, following Deroy and Rappe ([under review](#)) we take the subjective experience of reality to be a composite that includes both categorical mechanisms of reality monitoring (is this real or not?) and a variety of qualitative, gradual, subjective signatures of reality that accompany different “non-imaginary” experiences. We argue that disturbances in the actual-model monitoring (as well as the subjective signature of reality) in some cases may play a role in explaining the sources and symptoms of psychosis.

4. Breakdowns in counterfactually rich models

Assuming the counterfactual richness of the generative models in PPF and its relevance to psychosis, the next step is to specify in more detail the different ways in which such rich counterfactuality may break down and lead to the symptoms such as delusions, illusions and hallucinations, derealization, and the “uncanniness” of experience in psychosis.⁵ As a starting point, we identify four possible ways in which such a breakdown could occur. The list is not necessarily exhaustive, and some options may work better than others and be better fits for different cases.

- (1) Actuality monitoring breakdown: misidentifying the counterfactual parts of the model as pertaining to the actual world;
- (2) Disconnectedness from (lack of access to) certain parts of the richly counterfactual model;
- (3) Poor counterfactual underpinning: inability to generate enough alternative hypotheses for sufficient counterfactual depth;
- (4) Perceptual reality grounding problems: abnormality in the subjective markers of reality.

Below we elaborate on these four options. Importantly, however, our proposal is not meant to substitute the more traditional precision estimation-based approaches to psychosis in PPF, rather it is complementary, providing *additional* possible sources of disorder. There is an increasing understanding in psychiatry that superficially similar symptoms may have distinct pathophysiological mechanisms and that diagnosis based simply on a cluster of symptoms may mistakenly group “heterogeneous syndromes with different pathophysiological mechanisms into one disorder” (Wong et al., 2010), which could ultimately result in a low efficiency of the administered treatment in individual cases and mismanagement of cognitive and financial resources when it comes to developing new treatments more broadly (Ford et al., 2014). Exhaustively mapping out the space of the possible pathophysiological mechanisms then becomes crucial for accurate diagnosis and treatment. We see our proposal as contributing toward this goal from the perspective of a specific theoretical framework, the PPF.

4.1. Option 1: Actuality monitoring breakdown

The first option corresponds to the situation in which rich counterfactual models are generated but wrong hypotheses are selected “as real” because the parts of the model that pertain to the counterfactual alternatives to the current state of affairs are misidentified as pertaining to the actual world. Depending on how the actual-world monitoring mechanisms are conceived, this case may closely resemble some of the more traditional, precision-estimation (PE) accounts of psychosis.

According to the PE accounts, a self-generated stimulus may be mistaken for a stimulus caused by an external source due to a malfunction in precision estimation (Giersch & Mishara, 2017). The consequences of self-generated stimuli are expected to be easy for the system to predict (the predictions have high precision), and so when they aren’t well predicted, they generate levels of prediction error akin to external stimuli, and hence are experienced as such. This would lead to, for example, inner speech being experienced as having an external source, leading to auditory verbal hallucinations (voice-hearing) (see, e.g., S.R. Jones & Fernyhough, 2007). This is how self-monitoring accounts of psychosis (Frith, 1992) are accommodated as a special case within the PPF (Wilkinson, 2014). There is, in effect, too much prediction error generated by self-produced stimuli, and hence the self-produced stimuli are erroneously deemed to not be self-produced.

Similar effects may also be achieved, however, not by the general dysregulation resulting in wrongly accommodating error signals across the system, but by misidentifying a part of the generative model related to a counterfactual scenario as pertaining to the actual world. This may manifest, for example, as delusions or hallucinatory experiences (again, depending on where in the hierarchy such mis-selection occurred). Here, however, the problem arises not as a global tendency toward top-down processing or a random precision estimation breakdown, but specifically in relation to the treatment of counterfactuals. This provides advantages over the PE dysregulation accounts, at least when it comes to accommodating certain cases. For example, hallucinations and delusions are not necessarily expected to co-occur because misidentification is now limited to very specific counterfactual sub-models. For the same reason, simultaneous occurrence of hallucinations and autistic behavior/decreased susceptibility to visual illusions in one individual are no longer in theoretical conflict: the hallucinatory/delusional part of one’s experience is explained by the actuality monitoring breakdown, rather than a general top-down bias in the system. Hence, it is no longer incompatible with the bottom-up processing bias assumed to take place in individuals diagnosed with autism spectrum disorder (ASD; Pellicano & Burr, 2012).⁶ This is compatible with the

reported co-occurrence of ASD and psychosis in individuals, and with the observation that the manifestation of psychosis in individuals with ASD is somewhat atypical (Larson et al., 2017).

Finally, if we treat actuality monitoring as (at least in part) a metacognitive process, this case also aligns with the metacognitive accounts of psychosis, supported by evidence that the patients diagnosed with schizophrenia often show certain metacognitive deficiencies (Cella et al., 2015; Lysaker et al., 2011, 2014).

4.2. Option 2: Loss of access

The second option relates to a problem with accessing the right parts of the model. If the ability to navigate the entirety of the generative model is somewhat impaired, the agent may be stuck in certain possible world interpretations (subparts of the generative model) that would result in persistent hallucinatory and/or delusional experiences. The idea here is that, although a rich counterfactual model is generated, some of the alternative hypotheses are blocked from being selected and relied upon in further processing. This may include the alternatives corresponding to the real state of affairs, forcing the system to operate on a set of inevitable “false” options. This could explain the inability of an affected individual to properly process evidence against the model and re-assess. Indeed, delusion seem to be very resistant to evidence, even if such evidence is judged completely trustworthy (Wilkinson, 2015). Further, like with the previous option, because such a disconnect could theoretically occur at any specific part of the model, it would be rather natural, again, to assume that delusions and hallucinations would not always come together. There are no general top-down biases involved, only the impaired ability to break out of certain (counterfactual) “frames”. Furthermore, an impairment in the ability to navigate the entirety of the generative model is directly associated with the problems integrating information from multiple parts of the model. This could provide an alternative explanation to the observed decreased proportion of integrative “solutions” to the McGurk effect in populations with psychosis compared to controls (White et al., 2014; as opposed to the standard explanation within the PPF that there is too much prediction error).

Beyond the waking (yet altered) states, such as those in psychosis, the substantial loss of access to subparts of the generative model is characteristic of dreaming. Impaired connectivity and limited access to episodic memory are indeed established features of REM sleep and may explain the subjectively “real” feeling of the dream environments (even though they can be deeply bizarre in content). Reality

monitoring simply does not have the correct targets as viable options for applying itself. If the access is partially restored, however, different cues and monitoring processes may pick up on this, leading, for example, to (partial) lucidity, even if the dream is experienced as highly immersive. It also accounts for other interesting features of the strange phenomenology of dream content: often the counterfactual depth that tells us who an individual is, may clash with the surface imagery. In other words, you are convinced in your dream (indeed you never question) that someone is a certain individual, even though they look nothing like them. Consider this dream report: “I had a talk with your colleague, but she looked differently, much younger, like someone I went to school with, perhaps a 13-year-old girl” (Schwartz & Maquet, 2002, p. 26). Or this one: “I recognize A’s sister . . . I am surprised by her beard, she looks much more like a man than a woman, with a big nose” (Schwartz & Maquet, 2002, p. 29). As Wilkinson (2015) notes, this bears significant similarity to delusional misidentification, which often occurs in association with first-episode psychotic disorders (Gupta et al., 2021; Jovic, 1992; Salvatore et al., 2014). Of course, another cause of delusional misidentification, and, one might suppose, of loss of access to relevant counterfactuals underpinning the generative model, is localized brain damage. Here the delusional individual may admit that the person they perceive looks just like a loved one but is not experienced this way because of a lack of the relevant counterfactuals (e.g., the individual is not experienced as huggable or as expected to behave in certain familiar ways).

4.3. Option 3: Poor counterfactual underpinning

Third, there is also a possibility that not enough counterfactual alternatives are generated in the presence of the correct winning hypothesis. On the perceptual level, this could lead to insufficient counterfactual depth (see, Seth, 2014). This would manifest itself experientially in things seeming flat, unreal, lacking in depth. This in turn might lead to delusions, for example, delusional misidentification of a loved one (e.g., they might be an android). On higher cognitive levels this could lead to impaired counterfactual reasoning and generation of hypothetical scenarios. In fact, counterfactual thinking is often impaired in patients diagnosed with schizophrenia precisely in the decreased ability to generate counterfactual scenarios (Albacete et al., 2017; Hooker et al., 2000). For example, Albacete et al. (2017) found that, although patients with schizophrenia do not differ from controls in their ability to identify an event most relevant for reversing a given scenario (their causal thinking is intact), they generate significantly fewer spontaneous

alternative and counterfactual scenarios, especially in cases of spatial and temporal “nearly happened” events. Interestingly, this is something that cannot be easily attributed to the traditional precision estimation breakdown/excessive prediction error accounts.⁷

4.4. Option 4: Subjective markers error

Fourth and finally, a problem could arise with the subjective markers of reality. This could happen both due to the problem with reality monitoring, or independently, for individual subjective markers. The alteration in the subjective signature of reality, in either case could lead to the experience of derealization. For example, the latter case could correspond to derealization in healthy individuals (which are rather common, see, e.g., Aderibigbe et al., 2001) induced, for example, by sensory deprivation (Reed & Sedman, 1964), extreme stress (Bernat et al., 1998) or drug/alcohol abuse (Melges et al., 1974). The former case, on the other hand, could be the cause of derealization commonly observed as an early symptom in patients with psychosis (Giersch & Mishara, 2017).

If we treat some cases of psychosis as resulting from counterfactual navigation impairment, it makes sense that one of the first symptoms of impaired actuality monitoring may be alteration of the reality signature of perception. Further, such alteration may on its own over time lead to precision redistribution in the generative model (without necessitating any consistent biases or precision estimation malfunction), resulting, for example, in different interpretation of the incoming sensory information/related higher-level causal inferences and, consecutively, perceptual hallucinations and delusions.

This may offer one explanation why hallucinations sometimes occur with or without accompanying sense-of-reality changes and can be judged by the suffering individual as either real or unreal. Depending on the individual's own differences in processing, including precision weightings assigned to certain parts of the model and specific types of evidence, as well as general proclivity toward more top-down/bottom-up processing, a malfunction in the sense of reality/actuality monitoring system may have stronger or weaker effect on the evaluation of the categorical reality status and the content of one's perception and vice versa. Although mutual reliance of the cognitive judgment, reality (meta-)monitoring, and the subjective signatures of reality on each other may occasionally lead to hypotheses selection errors, the functional redundancy in these processes is generally a helpful, rather than a hindering feature. Partial functional overlap among various types of evaluating the ontological nature of various submodels may make an individual less responsive to the processing errors in each subsystem making complex cognitive processing significantly more robust.

5. Consequences

5.1. Counterfactual depth and reality monitoring

We end up with a view that looks rather like the influential *reality monitoring* accounts introduced in the 1980's (Bentall & Slade, 1985). However, there are some important and illustrative differences that come with the counterfactually rich PPF interpretation. First, reality monitoring is task based: it gestures toward a task that an individual can do badly or well. We are delving beneath the success or failure of correctly ascertaining reality, to the mechanisms thanks to which this is possible.

Second, and most interestingly perhaps, reality monitoring was thought to be a subtype of *source monitoring*. Source monitoring is a notion borrowed from memory research, where the source could be defined, for example, as “the spatial, temporal, and contextual characteristics of an event as well as the sensory modalities through which it was perceived” (Vinogradov et al., 1997, p. 1530). In a classic review of source monitoring (in general, not in the context of psychosis), Johnson et al. (1993) claim that the term “source monitoring” subsumes at least three distinct abilities.

(1) *Internal source monitoring* – distinguishing one's real actions (verbal and bodily) from merely imagined ones.

(2) *External source monitoring* – distinguishing between outer sources (e.g., one third party from another).

(3) *Reality monitoring* – distinguishing between self-generated and outer events.

Bentall and Slade (1985) hypothesized that source monitoring could help to explain psychosis, and especially the third category of reality monitoring. In other words, the hypothesis at the center of reality monitoring accounts is that people with schizophrenia/psychosis are bad at distinguishing between self-generated and outer events. Not only that, but they have a bias in a particular, externalizing, direction: they have a *general propensity* to mistake self-produced events for external events. For example, a self-produced piece of imagery, either in the form of inner speech or episodic memory, is misattributed to an outside source. The basic idea is that if you misattribute something self-generated to the world (the non-self), then you will take fantasy (that which you made up), to be reality (that which is constrained by fact, by actuality). This basic logic is also what is behind a similar (but importantly different) approach to psychosis, namely, comparator-based self-monitoring (Frith, 1992).

This is very different to our understanding of reality monitoring. First of all, much of that which is self-generated is perfectly real. Since, according to the PPF, the world's contribution is so sparse and noisy, we have to construct our reality, albeit in a constrained manner. Furthermore, much of our cognition – both online and (more obviously) offline – is about inferring

what is the case. In an important respect these inferential processes are self-produced, but they aren't by that same token inaccurate fantasies. Conversely, external elements are very much capable of leading us astray, either because we draw inferences in directions we ought not to, or because we are genuinely misled through no fault of our own cognition. Stated most generally, then, we differ both from (task-referencing) reality monitoring and (mechanism-referencing) self-monitoring by insisting that the equations between “from me” and “not real”, and “not from me” and “real”, do not hold. Ultimately, then, on our view, distinguishing reality from non-reality is not about recognizing source: firstly, because it is more heterogeneous than that, but also because the relevant processes function at a lower level than experience: they help to generate the experience as the experience that it is, rather than characterizing the response to the experience. In other words, these processes will be baked into the experience, rather than a judgment we make based on the experience. This seems very much in keeping with the phenomenological complexity of psychosis, and its “location” within phenomenology. It is not like psychosis involves strange judgments based on relatively normal experiences: it involves alterations to experience (Berkovitch et al., 2021; Giersch & Mishara, 2017; Parnas & Henriksen, 2016).

5.2. Summary and implications

Taking the PPF as a starting point for thinking about psychosis has been a fruitful approach. Here we suggest that the innovation of counterfactual depth in the PPF should be similarly extended to our thinking about psychosis. We do not intend this as a critique of the more straightforward view, but as a potential addition to them that may help to account for a wider array of the many things that fall under the category “psychosis”.

Having said this, it does cast psychosis (or at least some forms of it) in a different light. It is no longer primarily cast as adopting a radically inaccurate hypothesis but rather as a subtle anomaly in something we all have, namely, mechanisms for distinguishing the actual, from the non-actual. This makes “accurate” cognition seem more fragile and delicate, but also closes the gap between psychosis and typical cognition. In other words, there is less difference between the brains of individuals in states of psychosis and the brains of those who are not in such states.

Our counterfactually enriched PPF account also allows for very heterogeneous forms of psychosis, and a richer understanding of its experience, beyond the presence of straightforward delusions and hallucinations, and into quasi-perceptual/uncanny/unreal etc. (see, Ratcliffe, 2017). Sometimes theorists talk about psychosis as if, against a backdrop of otherwise normal

experience, a voice is heard, or strange beliefs emerge, but the clinical and experiential reality of psychosis is often one of varied, pervasive, subtle, and unfamiliar changes to the basic fabric of experience.

5.3 Future directions

Our contribution is as speculative as it is modest. As we have clarified, we are not criticizing the existing views, but rather pointing in other unexplored directions. These speculations need to be tested and fleshed out through careful observation and experimentation. This requires a holistic, joined-up approach that examines everything from the neural and neurobiological underpinnings of counterfactual depth within the PPF, whether this be through imaging techniques, drug models (how might some drugs, for example, flatten counterfactual depth in ways that mimic certain form of psychotic experience?), etc., up to careful phenomenological investigation.

Another future direction is very straightforward. The counterfactually embellished PPF can be applied beyond psychosis, toward other conditions that have been given a more traditional PPF treatment, such as post-traumatic stress disorder (Wilkinson et al., 2017). It is worth noting that accounts within the related “Free Energy” framework have already cast depression in terms of changes to the experience of possibilities (Kiverstein et al., 2020), and to us this added subtlety seems very much in the right direction. Most generally, the appreciation that experience is not simply about perceiving sensory qualities of the here-and-now (hearing sounds, seeing colors and shapes), but experiencing a subtle patchwork of possibilities.

Notes

1. For an accessible introduction on predictive processing see, Wiese and Metzinger (2017). For more detailed treatments see, e.g., Clark (2015) and Hohwy (2013).
2. We thank one of the anonymous referees for clarifying this point.
3. As Jones and Wilkinson (2020) note, however, deliberate imaginative acts are a very specific form of personal-level counterfactual cognition, and the imaginative capacity is not fully exhausted by the ability to generate counterfactual alternatives.
4. There is, of course, an important difference between *agent's* free, conscious exploration of alternative hypotheses (conscious voluntary cognition) and constant counterfactual-model generation by the *entire cognitive system*, for example, such as those implicated in perception. In this paper we deliberately do not discuss agency and intentionality. The assumption is that some form of tracking is required for both conscious and intentional, and unconscious and involuntary simulation-based counterfactual exploration.
5. Importantly, there are a lot of cases where an agent may “misperceive” without anything being broken in the system. For example, simple auditory illusions like misrecognizing the sound of fresh snow under one's foot as a bird chirp can be

easily explained by selecting a wrong prediction on the basis of one's priors. Such erroneous selection does not signify any systematic problems; the bird noise may be simply more expected (although not corresponding to what is really going on). Other cases of misperception in healthy population populations, however, at least seemingly go beyond simple misperception. They are not rooted in perceptual signals and indicate that there has been a failure of (or rather "incorrect") integration of information somewhere at a higher level of inference that includes cognitive beliefs. A prime example of such a situation is the Third Man Syndrome. However, the phenomenon typically occurs in the situations of high stress where all kinds of one-time abnormalities seem plausible. Here, we ignore such one-time cases and focus only on the mechanisms of hallucinations, delusions, and derealization in psychosis.

6. A bottom-up processing bias, in fact, may serve as a compensatory mechanism in the situation where reality monitoring is somehow unreliable. Giving more "voice" to prediction errors is a helpful (although, perhaps, more effortful) strategy to keep grounded in the current, real world model when the monitoring mechanisms are less reliable.
7. As one of the reviewers pointed out, our treatment presents a strong deficit account in which the relevant counterfactual hypotheses are absent. Yet, another possibility would be that these hypotheses are in fact still generated but *lack in precision*, which in principle, could give rise to similar consequences.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Sofia Rappe is a Neurophilosophy Ph.D. Candidate at the Graduate School of Systemic Neurosciences at LMU Munich and a member of Cognition, Values & Behaviour research group. She investigates how predictive processing can be applied to the cognitive processes beyond sensory perception and give rise to the different kinds of agent-level experiences. More broadly, she is interested in the relationships between our linguistic capacity, conceptual thought, imagination, and constructive perception.

Sam Wilkinson is a Senior Lecturer in Philosophy in the Department of Sociology, Philosophy, and Anthropology at the University of Exeter. He works on hallucinations, delusions, psychosis, psychological trauma, brain injury, and the nature of illness and wellbeing. He also has a general interest in perception, action, and emotion as viewed from predictive processing and embodied perspectives, and especially in the way that the mind harnesses social and cultural context to enhance and shape cognition.

ORCID

Sofia Rappe  <http://orcid.org/0000-0003-3343-7025>

References

- Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4, 47. <https://doi.org/10.3389/fpsyg.2013.00047>
- Aderibigbe, Y. A., Bloch, R. M., & Walker, W. R. (2001). Prevalence of depersonalization and derealization experiences in a rural population. *Social Psychiatry and Psychiatric Epidemiology*, 36(2), 63–69. <https://doi.org/10.1007/s001270050291>
- Albacete, A., Contreras, F., Bosque, C., Gilabert, E., Albiach, Á., & Menchón, J. M. (2017). Symptomatic remission and counterfactual reasoning in schizophrenia. *Frontiers in Psychology*, 7, 2048. <https://doi.org/10.3389/fpsyg.2016.02048>
- Avila, M. T., Hong, L. E., Moates, A., Turano, K. A., & Thaker, G. K. (2006). Role of anticipation in schizophrenia-related pursuit initiation deficits. *Journal of Neurophysiology*, 95(2), 593–601. <https://doi.org/10.1152/jn.00369.2005>
- Barrett L Feldman, Quigley K S and Hamilton P. (2016). An active inference theory of allostasis and interoception in depression. *Phil. Trans. R. Soc. B*, 371(1708), 20160011 [10.1098/rstb.2016.0011](https://doi.org/10.1098/rstb.2016.0011)
- Bentall, R. P., & Slade, P. D. (1985). Reality testing and auditory hallucinations: A signal detection analysis. *The British Journal of Clinical Psychology*, 24(Pt. 3), 159–169. <https://doi.org/10.1111/j.2044-8260.1985.tb01331.x>
- Berkovitch, L., Charles, L., Del Cul, A., Hamdani, N., Delavest, M., Sarrazin, S., Mangin, J.-F., Guevara, P., Ji, E., d'Albis, M.-A., Gaillard, R., Bellivier, F., Poupon, C., Leboyer, M., Tamouza, R., Dehaene, S., & Houenou, J. (2021). Disruption of conscious access in psychosis is associated with altered structural brain connectivity. *Journal of Neuroscience*, 41(3), 513–523. <https://doi.org/10.1523/JNEUROSCI.0945-20.2020>
- Bernat, J. A., Ronfeldt, H. M., Calhoun, K. S., & Arias, I. (1998). Prevalence of traumatic events and peritraumatic predictors of posttraumatic stress symptoms in a nonclinical sample of college students. *Journal of Traumatic Stress*, 11(4), 645–664. <https://doi.org/10.1023/A:1024485130934>
- Burr, C., & Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology*, 29(4), 586–600. <https://doi.org/10.1080/09515089.2015.1135238>
- Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, 67(1), 135–157. <https://doi.org/10.1146/annurev-psych-122414-033249>
- Cassidy, C. M., Balsam, P. D., Weinstein, J. J., Rosengard, R. J., Slifstein, M., Daw, N. D., Abi-Dargham, A., & Horga, G. (2018). A perceptual inference mechanism for hallucinations linked to striatal dopamine. *Current Biology*, 28(4), 503–514. <https://doi.org/10.1016/j.cub.2017.12.059>
- Cella, M., Reeder, C., & Wykes, T. (2015). Lessons learnt? The importance of metacognition and its implications for cognitive remediation in schizophrenia. *Frontiers in Psychology*, 6 (September), 1259. <https://doi.org/10.3389/fpsyg.2015.01259>
- Chekroud, A. M. (2015). Unifying treatments for depression: An application of the free energy principle. *Frontiers in Psychology*, 6(February), 153. <https://doi.org/10.3389/fpsyg.2015.00153>
- Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

- Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, 26(9–10), 19–33. https://scholar.google.com/scholar_lookup?title=Bayesing+Qualia.+Consciousness+as+Inference,+Not+Raw+Datum&author=Clark,+A.&author=Friston,+K.&author=Wilkinson,+S.&publication_year=2019&journal=J.+Conscious.+Stud.&volume=26&pages=19%E2%80%9333
- Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, 35(3), 1–45. <https://doi.org/10.1007/s10539-020-09746-2>
- Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206(4), 515–530. <https://doi.org/10.1007/s00213-009-1561-0>
- Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R., III. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences*, 23(2), 114–127. <https://doi.org/10.1016/j.tics.2018.12.001>
- Deroy, O., & Rappe, S. (under review). The clear and not so clear signatures of perceptual reality in the Bayesian brain. Manuscript submitted for publication.
- Dima, D., Roiser, J. P., Dietrich, D. E., Bonnemann, C., Lanfermann, H., Emrich, H. M., & Dillo, W. (2009). Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. *Neuroimage*, 46(4), 1180–1186. <https://doi.org/10.1016/j.neuroimage.2009.03.033>
- Dokic, J., & Martin, J. R. (2017). Felt reality and the opacity of perception. *Topoi*, 36(2), 299–309. <https://doi.org/10.1007/s11245-015-9327-2>
- Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48. <https://doi.org/10.1038/nrn2536>
- Ford, J. M., Morris, S. E., Hoffman, R. E., Sommer, I., Waters, F., McCarthy-Jones, S., Thoma, R. J., Turner, J. A., Kedy, S. K., Badcock, J. C., & Cuthbert, B. N. (2014). Studying hallucinations within the NIMH RDoC framework. *Schizophrenia Bulletin*, 40(Suppl_4), S295–S304. <https://doi.org/10.1093/schbul/sbu011>
- Friston, K. J. (2005). Hallucinations and perceptual inference. *Behavioral and Brain Sciences*, 28(6), 764–766. <https://doi.org/10.1017/S0140525X05290131>
- Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301. <https://doi.org/10.1016/j.tics.2009.04.005>
- 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., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3(May), 151. <https://doi.org/10.3389/fpsyg.2012.00151>
- Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Psychology press.
- Giersch, A., & Mishara, A. L. (2017). Is schizophrenia a disorder of consciousness? Experimental and phenomenological support for anomalous unconscious processing. *Frontiers in Psychology*, 8(September), 1659. <https://doi.org/10.3389/fpsyg.2017.01659>
- Gupta, M., Gupta, N., Zubiari, F., & Ramar, D. (2021). Delusional misidentification syndromes: Untangling clinical quandary with the newer evidence-based approaches. *Cureus*, 13(12), . <https://doi.org/10.7759/cureus.20165>
- Hoerl, C., & McCormack, T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. *Behavioral and Brain Sciences* 42(1–69), . <https://doi.org/10.1017/S0140525X18002157>
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

- Hong, L. E., Avila, M. T., & Thaker, G. K. (2005). Response to unexpected target changes during sustained visual tracking in schizophrenic patients. *Experimental Brain Research*, 165(1), 125–131. <https://doi.org/10.1007/s00221-005-2276-z>
- Hooker, C., Roese, N. J., & Park, S. (2000). Impoverished counterfactual thinking is associated with schizophrenia. *Psychiatry*, 63(4), 326–335. <https://doi.org/10.1080/00332747.2000.11024925>
- Horga, G., Schatz, K. C., Abi-Dargham, A., & Peterson, B. S. (2014). Deficits in predictive coding underlie hallucinations in schizophrenia. *The Journal of Neuroscience*, 34(24), 8072–8082. <https://doi.org/10.1523/JNEUROSCI.0200-14.2014>
- Humpston, C. S., & Broome, M. R. (2016). The spectra of soundless voices and audible thoughts: Towards an integrative model of auditory verbal hallucinations and thought insertion. *Review of Philosophy and Psychology*, 7(3), 611–629. <https://doi.org/10.1007/s13164-015-0232-9>
- Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3), 404–413. <https://doi.org/10.1038/nn.4238>
- Jocic, M. D. (1992). Delusional misidentification syndromes. *Jefferson Journal of Psychiatry*, 10(1), 4. <https://doi.org/10.29046/JJP.010.1.001>
- Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114(1), 3. <https://doi.org/10.1037/0033-2909.114.1.3>
- Jones, S. R., & Fernyhough, C. (2007). Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Consciousness and Cognition*, 16(2), 391–399. <https://doi.org/10.1016/j.concog.2005.12.003>
- Jones, M., & Wilkinson, S. (2020). From prediction to imagination. In A. Abraham Ed., *The cambridge handbook of the imagination*. (Cambridge Handbooks in Psychology, pp. 94–110). Cambridge University Press.
- Kiverstein, J., Miller, M., & Rietveld, E. (2020). How mood tunes prediction: A neurophenomenological account of mood and its disturbance in major depression. *Neuroscience of Consciousness*, 2020(1), niaa003. <https://doi.org/10.1093/nc/niaa003>
- Knight, R. T., & Grabowecky, M. (1995). Escape from linear time: Prefrontal cortex and conscious experience. *The cognitive neurosciences*, (pp. 1357–1371). The MIT Press . <https://psycnet.apa.org/record/1994-98810-090>
- Larson, F. V., Wagner, A. P., Jones, P. B., Tantam, D., Lai, M. C., Baron-Cohen, S., & Holland, A. J. (2017). Psychosis in autism: Comparison of the features of both conditions in a dually affected cohort. *The British Journal of Psychiatry*, 210(4), 269–275. <https://doi.org/10.1192/bjp.bp.116.187682>
- Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, 8(May), 302. <https://doi.org/10.3389/fnhum.2014.00302>
- Lee, P. S., Foss-Feig, J., Henderson, J. G., Kenworthy, L. E., Gilotty, L., Gaillard, W. D., & Vaidya, C. J. (2007). Atypical neural substrates of embedded figures task performance in children with autism spectrum disorder. *Neuroimage*, 38(1), 184–193. <https://doi.org/10.1016/j.neuroimage.2007.07.013>
- Lysaker, P. H., Erickson, M., Ringer, J., Buck, K. D., Semerari, A., Carcione, A., & Dimaggio, G. (2011). Metacognition in schizophrenia: The relationship of mastery to coping, insight, self-esteem, social anxiety, and various facets of neurocognition. *British Journal of Clinical Psychology*, 50(4), 412–424. <https://doi.org/10.1111/j.2044-8260.2010.02003>

- Lysaker, P. H., Leonhardt, B. L., Pijnenborg, M., van Donkersgoed, R., de Jong, S., & Dimaggio, G. (2014). Metacognition in schizophrenia spectrum disorders: Methods of assessment and associations with neurocognition, symptoms, cognitive style and function. *Isr J Psychiatry Relat Sci*, 51(1), 54–62. https://d1wqtxts1xzle7.cloudfront.net/45610766/Metacognition_in_schizophrenia_spectrum_20160513-4642-eg2ljb-with-cover-page-v2.pdf?Expires=1647510986&Signature=ZthQiES~CmSFzRsOG2VJEALQmTkniaBk9FcYz~6eR-FHF4K2tStST72PMYVv4Vl1NTZRzy382iyZJgyMOvGZVFv0RiGB0UgepnnR2E3S7dzAhWQYZ1obn1QugxMBYm2D7CyawVQXi6cT~47Cs31IFT6F4yjWm1R~5RYGKpD0JzKTYCluNJ1b8zGwliaklCEFWsVDjxPTZml24520whXqH0y0i6tm9U1sc6oxKkpq8eOxRqvEsSL2sfgTkfdelsgs6OS6PWosvJpqWqI68IC9NaF0E~shKvwJW1dGlmoJwCpRtRyeraqzO67U1XUMRipzId3fQQQbKUnBY~vcrAsfw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA
- Manjaly, Z. M., Bruning, N., Neufang, S., Stephan, K. E., Brieber, S., Marshall, J. C., ... Fink, G. R. (2007). Neurophysiological correlates of relatively enhanced local visual search in autistic adolescents. *Neuroimage*, 35(1), 283–291. <https://doi.org/10.1016/j.neuroimage.2006.11.036>
- McGrath, J. J., Saha, S., Al-Hamzawi, A., Alonso, J., Bromet, E. J., Bruffaerts, R., Caldas-de-almeida, J. M., Chiu, W. T., de Jonge, P., Fayyad, J., Florescu, S., Gureje, O., Haro, J. M., Hu, C., Kovess-Masfety, V., Lepine, J. P., Lim, C. C. W., Mora, M. E. M., Navarro-Mateu, F., ... Kessler, R. C. (2015). Psychotic experiences in the general population: A cross-national analysis based on 31 261 respondents from 18 countries. *JAMA psychiatry*, 72(7), 697–705. <https://doi.org/10.1001/jamapsychiatry.2015.0575>
- Meehan, M., Razaque, S., Whitton, M. C., & Brooks, F. P., Jr. (2003). Effect of latency on presence in stressful virtual environments. *Proceedings of the IEEE Virtual Reality, 2003*, 141–148. <https://doi.org/10.1109/VR.2003.1191132>
- Melges, F. T., Tinklenberg, J. R., Deardorff, C. M., Davies, N. H., Anderson, R. E., & Owen, C. A. (1974). Temporal disorganization and delusional-like ideation: Processes induced by hashish and alcohol. *Archives of General Psychiatry*, 30(6), 855–861. <https://doi.org/10.1001/archpsyc.1974.01760120099014>
- Miller, M., & Clark, A. (2018). Happily entangled: Prediction, emotion, and the embodied mind. *Synthese*, 195(6), 2559–2575. <https://doi.org/10.1007/s11229-017-1399-7>
- Noë, A. (2006). Experience without the head. *Perceptual experience*, 1, 411–433.
- Parnas, J., & Henriksen, M. G. (2016). Mysticism and schizophrenia: A phenomenological exploration of the structure of consciousness in the schizophrenia spectrum disorders. *Consciousness and Cognition*, 43, 75–88. <https://doi.org/10.1016/j.concog.2016.05.010>
- Pellicano, E., & Burr, D. (2012). When the world becomes ‘too real’: A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10), 504–510. <https://doi.org/10.1016/j.tics.2012.08.009>
- Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351), 596–600. <https://doi.org/10.1126/science.aan3458>
- Ratcliffe, M. (2013). Phenomenology, naturalism and the sense of reality. *Royal Institute of Philosophy Supplement*, 72, 67–88. <https://doi.org/10.1017/S1358246113000052>
- Ratcliffe, M. (2017). *Real hallucinations: Psychiatric illness, intentionality, and the interpersonal world*. MIT Press.
- Reed, G. F., & Sedman, G. (1964). Personality and depersonalization under sensory deprivation conditions. *Perceptual and Motor Skills*, 18(2), 659–660. <https://doi.org/10.2466/pms.1964.18.2.659>

- Salvatore, P., Bhuvaneshwar, C., Tohen, M., Khalsa, H. M. K., Maggini, C., & Baldessarini, R. J. (2014). Capgras' syndrome in first-episode psychotic disorders. *Psychopathology*, 47(4), 26–269. <https://doi.org/10.1159/000357813>
- Schwartz S and Maquet P. (2002). Sleep imaging and the neuro-psychological assessment of dreams. *Trends in Cognitive Sciences*, 6(1), 23–30. [https://doi.org/10.1016/S1364-6613\(00\)01818-0](https://doi.org/10.1016/S1364-6613(00)01818-0)
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565–573. <https://doi.org/10.1016/j.tics.2013.09.007>
- Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5(2), 97–118. <https://doi.org/10.1080/17588928.2013.877880>
- Smith, R., Parr, T., & Friston, K. J. (2019). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *Frontiers in Psychology*, 10(December), 2844. <https://doi.org/10.3389/fpsyg.2019.02844>
- Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10(November), 550. <https://doi.org/10.3389/fnhum.2016.00550>
- 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>
- Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eysen, L., Boets, B., De-wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, 121(4), 649. <https://doi.org/10.1037/a0037665>
- Vinogradov, S., Willis-Shore, J., Poole, J. H., Marten, E., Ober, B. A., & Shenaut, G. K. (1997). Clinical and neurocognitive aspects of source monitoring errors in schizophrenia. *American Journal of Psychiatry*, 154(11), 1530–1537. <https://doi.org/10.1176/ajp.154.11.1530>
- White, T. P., Wigton, R. L., Joyce, D. W., Bobin, T., Ferragamo, C., Wasim, N., Lisk, S., & Shergill, S. S. (2014). Eluding the illusion? Schizophrenia, dopamine and the McGurk effect. *Frontiers in Human Neuroscience*, 8(August), 565. <https://doi.org/10.3389/fnhum.2014.00565>
- Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: a primer on predictive processing. In Metzinger, T., Wiese, W.(Eds). *Philosophy and Predictive Processing*. Frankfurt: MIND Group. <https://predictive-mind.net/papers/vanilla-pp-for-philosophers-a-primer-on-predictive-processing>
- Wilkinson, S. (2014). 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. (2015). Delusions, dreams, and the nature of identification. *Philosophical Psychology*, 28(2), 203–226. <https://doi.org/10.1080/09515089.2013.830351>
- Wilkinson, S. (2020). Distinguishing volumetric content from perceptual presence within a predictive processing framework. *Phenomenology and the Cognitive Sciences*, 19(4), 791–800. <https://doi.org/10.1007/s11097-019-09632-7>
- Wilkinson, S. (2021). What can predictive processing tell us about the contents of perceptual experience? In Logue & Richardson (Eds.), *Purpose and procedure in the philosophy of perception*(pp. 174–190). Oxford University Press.

- Wilkinson, S., Dodgson, G., & Meares, K. (2017). Predictive processing and the varieties of psychological trauma. *Frontiers in Psychology*, 8, 1840. <https://doi.org/10.3389/fpsyg.2017.01840>
- Wong, E. H., Yocca, F., Smith, M. A., & Lee, C. M. (2010). Challenges and opportunities for drug discovery in psychiatric disorders: The drug hunters' perspective. *International Journal of Neuropsychopharmacology*, 13(9), 1269–1284. <https://doi.org/10.1017/S1461145710000866>
- Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. *Neuron*, 99(5), 1083–1097. <https://doi.org/10.1016/j.neuron.2018.07.035>

Chapter 5

Discussion

In this thesis, I aimed to further the case of predictive processing as a framework for explaining cognition, perception, and action by investigating two specific examples where predictive processing explanations of the cognitive processes conflict with the first-person experiences we have when we draw on these processes—conceptual linguistic thought and reality monitoring. Applying the approach of bidirectional reconciliation to these examples, on the one hand, highlighted that the strategy may be successful in some cases that go beyond sensory perception and that may seem initially problematic from the predictive-processing perspective.

The subjective-to-objective direction consisted of stripping down the assumptions that draw on the first-person experience of mental life, such as those about compositionality of thought (Chapter 2), categorical nature of the sense of reality (Chapter 3), or the absence of the bizarre predictions manifesting in psychosis in the neurotypical brains (Chapter 4). Removing these assumptions was the crucial part of arriving at the solutions presented in all three papers.

On the other hand, the same examples highlight that the minimal computational principles of predictive processing cannot fully explain some mental phenomena without further detailing and augmentation of the framework (objective-to-subjective direction). For example, Chapters 2 and 3 highlight the explanatory necessity of specifying the interaction between different cognitive/perceptual modules for explaining the compositional appearance of thoughts, while Chapter 4 highlights the need to augment the traditional predictive-processing-based architectures with richer, more expansive generative models than are typically assumed in perceptual predictive processing.

The solutions discussed in this dissertation have significant limitations. First, my approach in all three papers has been aligned with representational

predictive processing. Hence, if the anti-representational ecological view is favored, the solutions might not easily transfer. Second, the solutions discussed in Chapters 2-4 are mutually compatible at the **current** level of discussion—describing computational mechanisms and broad architectural principles of individual phenomena. However, it remains to be seen whether these explanations will remain compatible when their implementations are specified more precisely, that is, whether they could function as part of a single cognitive system in an internally coherent way. The aspects requiring more specificity include, for example, a clarification regarding the relationship between concepts for perception, cognitive inference, and language. Are they similar or distinct? How do they relate to one another? This question taps into a broader concern regarding mental representations (and their neurobiological implementations) involved in various cognitive processes. For example, can they be similar for the perceptual and cognitive content (traditionally understood)? Further, Chapter 3 requires a proposal regarding how reality monitoring is realized in predictive processing, while Chapter 4—a proposal regarding the realization of simulation, which is necessary to support processes such as imagination and counterfactual reasoning. Finally, all three papers are missing detailed specifications of how different cognitive modules communicate and integrate information in predictive brains. The task of providing such specifications is primarily an empirical one, but some preliminary models are necessary to establish the starting hypotheses.

Further, the current work does not allow one to establish whether the entire cognitive system is subject to a single underlying principle such as free energy minimization. In terms of unification, a pluralistic approach appears most likely, either of the kind where predictive processes have different implementations or where disparate predictive processes work alongside the non-predictive processes. Overall, it should not come as a surprise if the brain implements various non-homogeneous predictive processing-based solutions. After all, the consensus in biology is that our body has evolved through local mutation and adaptive solutions, and some internal mechanisms are realized in multiple ways (for example, chemical and electrical communication in neurons). Yet, either one of the pluralistic pictures may be compatible with the free energy principle since predictive processing is just one among many possible ways of minimizing free energy.

Finally, although this dissertation is primarily concerned with conscious mental states, it has completely ignored one aspect often present in conscious cognition: cognitive control. Cognitive control may play an important role in conceptual linguistic thought and reality monitoring discussed in this dissertation. Yet, addressing the topic of cognitive control would require going far

outside the reasonable scope of this project. Notably, unlike motor control, extensively featured in the predictive processing literature (see, e.g., Burnston, 2021; Carls-Diamante, 2021; Friston, 2011; Kahl & Kopp, 2018; Kilner, Friston & Frith, 2007), cognitive control has received little attention in the predictive community. This omission presents both a future challenge and an exciting new direction of inquiry.

Overall, the field would benefit from philosophical clarification of the conceptual (epistemic) status of predictive processing, which has been previously described in the literature as a theory of brain functioning (Ficco et al., 2021; Millidge, Seth & Buckley, 2021), a research tradition (Litwin & Miłkowski, 2020), a research program (Sprevak, 2021), a paradigm (Swanson, 2016), and a research framework (Clark, 2016; Hohwy, 2013). More clearly articulated stances regarding the status of predictive processing would aid both the proponents of predictive processing and its critics. First, without specification, predictive processing presents an ever-escaping target that is hard to criticize since there is always an opportunity to claim a different understanding. This clarification is crucial in the debates regarding falsifiability and empirical evidence for and against predictive processing. While specific theories may fail, the broader research program (or framework) of which they were a part may survive and still prove to be heuristically and explanatory useful. Hence, it is necessary to identify the target explicitly. Further, as Colombo and Wright (2017) noted, without specification of both the conceptual status of predictive processing and the conditions for success regarding unification, it is impossible to judge (and confirm) whether predictive processing fulfills the criteria for a grand unifying framework. “Advocates of PTB [predictive processing theory of the brain] have left unspecified the conditions under which they would take their assertion that PTB is a grand unifying theory to be true” (Colombo & Wright, 2017, p. 5) which means leaving the interpretation of this project to the critics.

Another clarification required is the use of personal and subpersonal language in predictive processing research. The terms like “confidence,” “surprise,” and “inference” in predictive processing are often applied both to the cognitive processes and the agent’s experience, creating a false sense of unification through the common vocabulary (recall a similar point in section 1.3, see also Litwin & Miłkowski, 2020). Although such ambiguous terminology has been characteristic of predictive processing from the beginning, as the framework is increasingly applied to explain mental states and mental phenomenology, linguistics precision becomes an ever more vital requirement.

Beyond such clarifications and the future directions highlighted in Chapters 2-4 concerning the phenomena targeted in these chapters, the gen-

eral next steps fall into three categories. First, there is a need to flesh out and test neurobiological and computational models that implement predictive processing explanations of specific processes and phenomena sketched out in this thesis and beyond. As highlighted by Walsh and colleagues (2020), this may also require the development of new methodologies and neuroscientific tools. Second, a rigorous assessment of the mutual compatibility of promising explanations is required. Compatibility is of particular concern for elucidating the phenomena that either occupy the same level of explanation, co-occur or strongly correlate (see, for example, the discussion on autistic spectrum disorders and schizophrenia in chapter 4). Third, we must establish a more systematic view of the human cognitive repertoire from the predictive-processing perspective—a predictive processing-based cognitive ontology. Such an ontology would also include specifying the relationships between various online and offline processes, for example, those traditionally classified as perception, imagination, and cognition.

Significant undertakings lie ahead. However, it is safe to say that predictive processing provides a promising direction and an important perspective to keep in mind when considering cognitive processes outside sensory perception. A particularly notable feature of predictive processing is that it may be used not just for explaining cognitive mechanisms but, as the newly emerging literature on conscious experiences on predictive processing indicate, for explaining mental states and mental phenomenology. This seeming suitability puts predictive processing in a unique position to elucidate the Real Problem of Consciousness (Seth, 2021)—how and why cognitive mechanisms may give rise to conscious mental states. Some researchers, such as Clark, Friston, and Wilkinson (2019), even argue that the emergence of subjective qualia is a necessary consequence of having a predictive mind in a complex environment such as our own.

It remains to be seen whether the project of predictive processing as a grand unifying theory succeeds or whether even local predictive processing solutions prove to be the actual rather than potential solutions implemented in the brain. At this point, the framework has gone mainstream, and we can plausibly expect many exciting developments very soon. However, even if unification through predictive processing is ultimately a dead end, there does not seem to be any good reason to cast away the framework just yet, and certainly not on account of offline perception and cognition.

Bibliography

(Introduction and Discussion)

Aitchison, L., & Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current opinion in neurobiology*, 46, 219–227.

Asprem, E. (2019). Predictive processing and the problem of (massive) modularity. *Religion, Brain & Behavior*, 9(1), 84–86.

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The depressed brain: an evolutionary systems theory. *Trends in Cognitive Sciences*, 21(3), 182–194.

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

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, 22(5), 543–564.

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1755), 20170341.

Burnston, D. C. (2021). Bayes, predictive processing, and the cognitive architecture of motor control. *Consciousness and Cognition*, 96, 103218.

Carls-Diamante, S. (2021). Explanation Within Arm's Reach: A Predictive Processing Framework for Single Arm Use in Octopuses. *Erkenntnis*, 1–16.

Cavedon-Taylor, D. (2021). Untying the knot: imagination, perception and their neural substrates. *Synthese*, 199(3), 7203–7230.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

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

Clark, A. (2019). Consciousness as generative entanglement. *The Journal of Philosophy*, 116(12), 645–662.

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, 26(9–10), 19–33.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science*, 68(2), 451–484.

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

Deane, G. (2021). Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution. *Neuroscience of Consciousness*, 2021(2), niab024.

Deroy, O. (2015). Modularity of perception. In M. Matthen (Ed.), *The Oxford handbook of philosophy of perception*. Oxford: Oxford University Press.

Deroy, O. (2019). Predictions do not entail cognitive penetration: “Racial” biases in predictive models of perception. In C. Limbeck-Lilienau & F. Stadler (Eds.), *The Philosophy of Perception* (pp. 235–248). Berlin: De Gruyter.

Dobs, K., Martinez, J., Kell, A. J., & Kanwisher, N. (2021). Brain-like functional specialization emerges spontaneously in deep neural networks. *bioRxiv*.

Dolega, 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.

Ficco, L., Mancuso, L., Manuello, J., Teneggi, A., Liloia, D., Duca, S., ... & Cauda, F. (2021). Disentangling predictive processing in the brain: a meta-analytic study in favour of a predictive network. *Scientific Reports*, 11(1), 1–14.

Fletcher, P. C. & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48.

Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford: Oxford University Press on Demand.

Friston, K. J. (2005). Hallucinations and perceptual inference. *Behavioral and Brain Sciences*, 28(6), 764–766.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K. (2011). What is optimal about motor control? *Neuron*, 72(3), 488–498.

Friston, K. J., Daunizeau, J., Kilner, J., Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biological cybernetics*, 102(3), 227–260.

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

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3, 130.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. New York: Psychology Press.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193, 559–582.

Gregory, R. L. (1973). *The confounded eye. Illusion in nature and art*. New York: Scribner.

Heavey, C. L., & Hurlburt, R. T. (2008). The phenomena of inner experience. *Consciousness and cognition*, 17(3), 798–810.

Helmholtz, H. V. (1866/1911). *Treatise on physiological optics*. Rochester: Continuum.

Hill, H., & Johnston, A. (2007). The hollow-face illusion: Object-specific knowledge, general assumptions or properties of the stimulus? *Perception*, 36(2), 199–223.

Hoerl, C., & McCormack, T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. *Behavioral and Brain Sciences*, 42.

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

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, 35(2), 209–223.

Janssen, A., Klein, C., & Slors, M. (2017). What is a cognitive ontology, anyway? *Philosophical Explorations*, 20(2), 123–128.

Jones, M., & Wilkinson, S. (2020). From Prediction to Imagination. In A. Abraham (Ed.), *The Cambridge Handbook of the Imagination* (pp. 94–110). Cambridge University Press.

Kahl, S., & Kopp, S. (2018). A predictive processing model of perception and action for self-other distinction. *Frontiers in Psychology*, 9, 2421.

-
- Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, 195(6), 2387–2415.
- Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 8(3), 159–166.
- Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third-wave view*. Routledge.
- Kiverstein, J., Miller, M., & Rietveld, E. (2017). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese*, 1–23.
- Klein, C. (2012). Cognitive ontology and region-versus network-oriented analyses. *Philosophy of Science*, 79(5), 952–960.
- Kogo, N., Trengove, C. (2015). Is predictive coding theory articulated enough to be testable? *Frontiers in Computational Neuroscience*, 111.
- Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, 8, 302.
- Litwin, P., & Miłkowski, M. (2020). Unification by fiat: arrested development of predictive processing. *Cognitive Science*, 44(7), e12867.
- Machery, E. (2005). You don't know how you think: Introspection and language of thought. *The British Journal for the Philosophy of Science*, 56(3), 469–485.
- Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*, 47, 6–16.
- Millidge, B., Seth, A., & Buckley, C. L. (2021). Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*.
- Oudeyer, P.Y. and Smith. L. (2016). How Evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, 8(2),

492-502.

Palmer, C. J., Paton, B., Kirkovski, M., Enticott, P. G., & Hohwy, J. (2015). Context sensitivity in action decreases along the autism spectrum: a predictive processing perspective. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1802), 2014–1557.

Pellicano, E. & Burr, D. (2012). When the world becomes ‘too real’: a Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10), 504–510.

Penny, W. (2012). Bayesian models of brain and behaviour. *ISRN Biomathematics*, 2012.

Piekarski, M. (2021). Understanding predictive processing. A review. *Avant*, 12(1).

Poth, N. (2022). Schema-Centred Unity and Process-Centred Pluralism of the Predictive Mind. *Minds and Machines*, 1–27.

Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*, 22(3–4), 262–275.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79.

Rappe, S. (2019). Now, never, or coming soon? Prediction and efficient language processing. *Pragmatics & Cognition*, 26(2–3), 357-385.

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

Seth, A. K. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*: 35. Frankfurt am Main: MIND Group.

Seth, A. (2021). *Being you: A new science of consciousness*. London: Faber & Faber.

Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2, 395.

Siegel, S. (2016). *The rationality of perception*. Oxford University Press.

Sims, A. (2017). The problems with prediction: The dark room problem and the scope dispute. In T. Metzinger W. Wiese (Eds.), *Philosophy and Predictive Processing: 23*. Frankfurt am Main: MIND Group.

Solms, M., & Friston, K. (2018). How and why consciousness arises: some considerations from physics and physiology. *Journal of Consciousness Studies*, 25(5-6), 202-238.

Sprevak, M. (2021). *Predictive coding II: The computational level*. PhilSci-Archive. <http://philsci-archive.pitt.edu/id/eprint/20641>.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., ... & Petzschner, F. H. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10, 550.

Swanson, L. R. (2016). The predictive processing paradigm has roots in Kant. *Frontiers in Systems Neuroscience*, 10, 79.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.

Thagard, P. (2019). *Brain-Mind: From Neurons to Consciousness and Creativity (Treatise on Mind and Society)*. Oxford: Oxford University Press.

Van de Cruys, S. (2017). Affective value in the predictive mind. In T. Metzinger W. Wiese (Eds.), *Philosophy and Predictive Processing: 24*. Frankfurt am Main: MIND Group.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eysen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, 121(4), 649.

Varela, F., Maturana, H. & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*, 5(4), 187–196.

Varela, F., Thompson, E. & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA, London, UK: The MIT Press.

Venter, E. (2021). Toward an embodied, embedded predictive processing account. *Frontiers in Psychology*, 137.

Viola, M. (2017). Carving mind at brain's joints. the debate on cognitive ontology. *Phenomenology and Mind*, 12, 162–172.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, 1464(1), 242–268.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 16, 715–736.

Wiese, W. (2018). *Experienced Wholeness: Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*. MIT Press.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger W. Wiese (Eds.), *Philosophy and Predictive Processing: 1*. Frankfurt am Main: MIND Group.

Wilkinson, S., Dodgson, G., Meares, K. (2017). Predictive processing and the varieties of psychological trauma. *Frontiers in Psychology*, 8, 1840.

Wilkinson, S., & Fernyhough, C. (2018). When Inner Speech Misleads. In P. Langland-Hassan A. Vicente (Eds.), *Inner Speech: New Voices*. Oxford: Oxford University Press.

Williams, D. (2020). Predictive coding and thought. *Synthese*, 197(4), 1749–1775.

Williford, K., Bennequin, D., Friston, K., & Rudrauf, D. (2018). The projective consciousness model and phenomenal selfhood. *Frontiers in Psychology*, 2571.

Eidesstattliche Versicherung / Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation "Mental states and cognitive mechanisms in predictive agents" selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation "Mental states and cognitive mechanisms in predictive agents" is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, den 1.06.2022
Munich, date 1.06.2022

.....
Sofia Rappe

Author contributions

Paper I

Rappe, S. (2022). Predictive minds can think: Addressing generality and surface compositionality of thought. *Synthese*, 200(1), 1-22. doi: 10.1007/s11229-022-03502-7.

S.R. is the first and sole author of this manuscript.

Paper II

Deroy, O.* & **Rappe, S.*** (*under review after "minor revisions"*). The clear and not so clear signatures of perceptual reality in the Bayesian brain.

S.R. conducted the initial literature review. S.R. and O.D. conceived of the research idea, wrote the original draft, and revised the manuscript for publication. Both authors recognize equal contributions to the current paper.

Paper III

Rappe, S. & Wilkinson S. (2022). Counterfactuals and psychosis: Adding complexity to predictive processing accounts. *Philosophical Psychology*, 1-24. doi: 10.1080/09515089.2022.2054789.

S.R. and S.W. conceived of the research idea and developed the arguments presented in the paper. S.R. drafted most of the manuscript and revised it for publication with the help of S.W.

.....
Sofia Rappe
Munich, 1 June 2022

.....
Prof. Dr. Ophelia Deroy