This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Predictive Embodied Concepts
An exploration of higher cognition within the predictive processing paradigm

Christian Michel

PhD in Philosophy
University of Edinburgh
2023
Declaration of authorship

I, Christian Michel, declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Christian Michel

11 June 2023
Note on publications

Six chapters of this thesis have been published in the following peer-reviewed journals:¹

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 8:

Chapter 9:

¹ Where required, permission has been obtained for the inclusion of the articles in this dissertation.
In the chapters that are based on published journal articles, I have intentionally maintained the original structure (including the abstracts). Given that the papers are self-contained, each includes a short introduction to PP, which might feel repetitive when reading the whole dissertation. However, the advantage of maintaining those introductions is that each chapter can be read independently. For instance, a linguist interested in copredication, needs not to read the whole dissertation, and can go straight to Chapter 9. I have corrected or modified minor orthographic issues in the article-based chapters for reasons of overall consistency (e.g., ensuring a consistent use of the following: capital letters, like in "Liar Paradox"—used as a proper name, acronyms, small caps for the names of concepts, and different types of hyphens and dashes).

One relevant terminological issue should be highlighted here. I am using the notion "format" of a concept in two different senses. In Chapter 4 it refers to the question of whether a concept has an amodal or modal (sensorimotor based) nature. In Chapter 5, "format" refers to the types of knowledge accommodated by different theories of concepts (prototypes, exemplars, theories, etc.). The use of "format" with two different senses is unfortunate; however, it is the consequence of the cumulative nature of this dissertation. To avoid any confusion, I clearly indicate how I use the term in those chapters.
Acknowledgements

I am extremely grateful to my supervisors Nick Treanor and Mark Sprevak for their invaluable support, guidance, and encouragement during this journey. Many thanks to Andy Clark who supervised my MSc thesis (out of which this PhD thesis grew) and who was a supervisor during the first year of my PhD.

A very special thanks to Guido Löhr for our fruitful and enjoyable collaboration. I have profited from and am grateful to the audiences from various conference, including the SPP and ESPP, and from conversations with Agustin Vicente, Brian Rabern and Ofra Magidor. I am indebted to the many anonymous reviewers of the journals to which I have submitted my manuscripts that form the backbone of this thesis. Their comments helped me so much to improve my ideas.

The roots of this thesis date back many years ago to an online short course by the Oxford University Department for Continuing Education (OUDCE) in Philosophy of Mind which brought me back to philosophy again, after very brief glimpses of Wittgenstein and Kant during high school times. I still remember a very long and inspiring exchange with my peers on the nature of thought. I defended the idea that we think in sort of pictures and movies against those who defended the idea that we think in language. Back then I had not the slightest clue about the existing work and debates that we had naively started to reinvent. Not much later, I was deeply impressed when I learned about George Lakoff's idea that we think in metaphors. Since then, I have embraced this long journey, exploring the nature of thought, passing through a BA and an MSc in philosophy, and discovering the exciting predictive processing framework through Andy Clark. In the early years I received a lot of encouragement to keep going from many of the members of the Philosophical Society associated with the OUDCE, and especially their then President, Marianne Talbot, as well as Eileen Walker, Frank Brierley, Bob Stone, Peter Gibson, and Mike Arnautov. I owe much to them. Thanks also to Geoffrey Klempner who has accompanied me as a tutor in the early BA studies.

This work is dedicated to Sefa and Juan.
Abstract

Predictive processing, an increasingly popular paradigm in cognitive sciences, has focused primarily on giving accounts of perception, motor control and a host of psychological phenomena, including consciousness. But higher cognitive processes, like conceptual thought, language, and logic, have received only limited attention to date and PP still stands disconnected from a huge body of research in those areas. In this thesis, I aim to address this gap and I attempt to go some way towards developing and defending a cognitive-computational approach to higher cognition within the predictive processing paradigm. To test its explanatory potential, I apply it to a range of linguistic and conceptual phenomena. I proceed in three steps. First, I lay out an account of concepts and suggest how concepts are represented, how they can be context-sensitively processed, and how the apparent diversity of formats arise. Secondly, I propose how paradigmatic higher cognitive competencies, like language and logical reasoning, could fit into the PP picture. Thirdly, I apply the PP account of concepts and language to a range of linguistic-conceptual phenomena as test cases, namely: metaphor, the semantic paradox (specifically the Liar Paradox) and copredication. Finally, I discuss some challenges and objections to the PP framework as applied to higher cognition and in general.
Lay summary

The brain is probably the most complex and mysterious thing that exists. It consists of billions of neurons, each interconnected with up to ten thousand other neurons. Everything we do, we do with the help of our brain: perceiving, feeling, acting, and thinking. However, we do not yet understand how exactly the brain achieves all this.

According to predictive processing, the brain is an organ that does not merely register sense impressions and interpret them passively. Rather it also works in the opposite direction. The brain tries to predict and anticipate everything. For that purpose, it maintains a mental model of the world. If it makes a prediction error, the model is changed so that next time the prediction improves.

Predictive processing has primarily been used to explain aspects of vision and movement. It also has turned out to be useful for explaining psychopathology, for instance autism and schizophrenia, as well as consciousness. However, so far, the theory has not been applied extensively to higher cognition, including thought and language.

There have been doubts whether predictive processing can be applied to thought and language. In this thesis, I try to show that those doubts are not as serious as they might seem. Drawing on the ideas of predictive processing, I provide a model of how we categorize everything in the world into concepts and how we use those concepts in thought. I also propose a model of how language might work.

To test my model, I apply it to some interesting problems that occupy linguists, psychologists, and philosophers and which have no accepted solution (among them, the famous Liar Paradox). While I do not claim to provide conclusive solutions, my model allows one to approach these problems from a new perspective. This suggests that predictive processing holds promise in terms of being fruitfully applied to improve our understanding of how high-level thought works.
Table of contents

Declaration of authorship ................................................................. 2
Note on publications ............................................................................. 3
Acknowledgements ................................................................................ 5
Abstract .................................................................................................. 6
Lay summary .......................................................................................... 7
Table of contents ................................................................................... 8
List of tables and figures ......................................................................... 12

Part 1 - Introduction................................................................................. 13

Chapter 1. Overview .............................................................................. 13
  1.1. Aims and motivation ....................................................................... 13
  1.2. Thesis overview ............................................................................ 15

Chapter 2. Setting the stage: Predictive processing and higher cognition .... 21
  2.1. The PP paradigm ........................................................................... 21
    2.1.1. PP as a paradigm, not a theory ............................................... 21
    2.1.2. PP as a package deal ............................................................... 23
        a) Perception as inference ......................................................... 24
        b) The Bayesian brain hypothesis .............................................. 24
        c) A hierarchical Bayesian brain .............................................. 25
        d) A generative model: the centrality of anticipation and prediction 25
        e) Active inference ................................................................ 26
    2.1.3. A characterization of PP via Marr's three levels ...................... 29
        1) Computational level ............................................................... 29
        2) Algorithmic level ................................................................ 30
        3) Implementational level .......................................................... 32
    2.1.4. The scope of PP applications .................................................. 34
    2.2. The challenges from higher cognition .......................................... 37
        2.2.1. The state of the art: PP and higher cognition ...................... 37
            a) Embodied cognition ......................................................... 37
            b) Bayesian cognition ......................................................... 39
            c) The language faculty and psycholinguistic research .......... 41
        2.2.2. Potential reasons for the gap ................................................. 44
            1) The prima facie challenge from higher cognition for PP .... 45
            2) The theoretical challenge from higher cognition for PP . .. 48
            3) The instrumental challenge from higher cognition for PP .... 49

Part 2 - A PP account of concepts ......................................................... 52

Introduction to Part 2 ........................................................................... 52

Chapter 3. Concept contextualism through the lens of predictive processing .... 54
  3.1. Introduction ................................................................................... 54
  3.2. Concept contextualism and concept invariantism ............................. 58
  3.3. The PP framework ........................................................................ 61
  3.4. Concepts in a PP model ................................................................. 65
    3.4.1. The role of concepts in PP ................................................... 65
    3.4.2. Characterizing conceptual representations in PP .................... 67
        3.4.2.1. Concepts encompassing multimodal, non-lexicalized, non-consciously accessible representations ............................ 67
List of tables and figures

Table 2.1: Simplified and idealized cumulative "genealogy" of core ideas leading to the PP framework. ........................................................................................................... 28

Table 6.1: Comparison of key features of LOTH and PP as cognitive-computational paradigms. ........................................................................................................... 153

Figure 3.1: Proposed “concept unit” in predictive processing................................. 66

Figure 3.2: Schematic representation of the error weighting mechanism for context dependent feature selection................................................................. 75

Figure 5.1: A schematic toy example of a concept unit network for the concept PIZZA and modes of processing................................................................. 130

Figure 10.1: Toy example for the "information package" in the form of an "expectation hierarchy" associated with the word form 'book'. ................................. 244

Figure 10.2: Expectation hierarchies for the 'city' (A) and 'White House' (B) examples from statements (8 a,b) and (9 a,b). ......................................................... 263
Part 1 - Introduction

Chapter 1. Overview

The challenges (empirical, conceptual, and methodological) are many and profound. But the potential payoff is huge. What is on offer is a multilevel account of some of the deepest natural principles underlying learning and inference, and one that may be capable of bringing perception, action, and attention under a single umbrella. The ensuing exchanges between neuroscience, computational theorizing, psychology, philosophy, rational decision theory, and embodied cognitive science promise to be among the major intellectual events of the early twenty-first century.

Andy Clark (2013, p. 20)

1.1. Aims and motivation

The aim of this dissertation is to explore how higher cognition (especially thought and language) could be addressed within the predictive processing framework. Predictive processing is an influential paradigm from cognitive neuroscience that has been gaining momentum and shape over the last twenty years and has seen a dramatic increase in attention in many disciplines. Predictive processing (hereafter PP) might have the potential to provide a long sought-after unifying neuro-mechanistic paradigm of concepts, principles, and mechanisms, that could guide and constrain theories of how our brain works and how perception, cognition, emotion, and action interact and contribute to survival in an uncertain world.

Accounting for higher cognition within the PP paradigm is an open problem. So far, the PP literature has thrived in the domain of perceptual and motor phenomena, and psychological phenomena including psychopathologies (e.g., autism and schizophrenia) and consciousness. But only a few incipient treatments of conceptual thought, logical competencies like deduction and induction, as well as language, are available within the PP framework. For instance, Hohwy’s (2020) recent overview of philosophically oriented work in PP includes only few examples of such treatments. Higher cognition today is not a domain in which PP is flourishing. Quite on the contrary, some critical voices (most notably Daniel Williams, 2019, 2020) have expressed concerns about the potential of PP to account for conceptual thought at
all. Also, Litwin & Miłkowski (2020) think that one should restrict the scope of PP to perceptual and motor phenomena.

My ambition in this thesis is limited to providing a sketch of a model of some core aspects of higher cognition within the PP paradigm. I am specifically interested in those aspects where Fodor's highly influential Language of Thought Hypothesis (LOTH) (e.g., 1975, 2008) has shown its strength, namely in compositional conceptual thought, logic, and natural language. Models of higher cognition describe the structure of mental representations used, and the processes carried out over them when we exercise those intelligent, higher-level cognitive competencies. This thesis is an exploration of how conceptual thought could arise in a brain-body system operating according to the PP principles.

If we develop accounts of certain higher cognition phenomena within PP, we can hope to achieve various things. Firstly, we would support the plausibility of PP as a cognitive-computational model of embodied minds because we would have shown that higher cognition could be within its scope. It would therefore, of course, offer support to those who see in PP a grand unified theory of the mind and agency in general (though I do not endorse this extreme view). Secondly, we may provide novel insights into interesting higher cognitive phenomena that, so far, have not been fully understood. A fruitful model of how higher cognition works should not only redescribe phenomena but also provide an advancement in their understanding. I have chosen some challenging conceptual-linguistic phenomena from philosophy of language and linguistics as test cases. These include the semantic paradox ("Liar Paradox"), metaphor and copredication. These phenomena pose deep and interesting problems, and we still have no common view on them. The hope is that a new way to think about thinking under the PP framework can shed new light on those phenomena. Thirdly, such a model of higher cognition might also gesture towards new approaches in artificial intelligence. Despite impressive advances in AI and specifically machine learning, artificial intelligent systems still lack general and flexible higher cognitive capabilities. What we need, many AI researchers think, is an account of human common-sense reasoning in AI terms. One approach is to focus more on the embodied nature of human cognitive agents, i.e., the fact that the brain is part of a body that interacts constantly with the world in its struggle for survival. It might be wrong to conceptualize intelligence as some abstract competency that can
be implemented as a set of rules that operate on symbols within a computer standing passively in some corner.

A central part of my thesis is a PP-specific account of the structure of concepts, understood in a cognitive psychological sense, that are used in higher cognition. Concepts, as I understand them, are bodies of knowledge (Machery 2009) or information packages that are posited to explain thought. Once such an account of concepts is in place, we can tackle issues like the representation of world knowledge, language and logic and then provide examples and apply the account to higher cognitive phenomena.

1.2. Thesis overview

This thesis is organized in three parts with a total of twelve chapters. Six of the chapters are standalone papers that have been published in peer reviewed journals. Each part is preceded by an introduction that briefly sets out some context and the main ideas and closes with some conclusions with a summary of the main take-aways.

Part 2 contains Chapters 3 to 5, which lay out a PP account of concepts. I focus on the psychological notion of concepts, i.e., concepts understood as cognitive-psychologically significant bodies of knowledge. I do not discuss the more philosophical aspects of concepts, like, for example, how concepts can refer, their possession conditions and/or criteria for their correct application.

Chapter 3 focuses on a mechanistic account of concept contextualism, the view that concepts are highly dynamic entities. Increasingly, scholars consider that concepts are not invariable cores of knowledge, but rather contain information structures from which parts are recruited depending on the context. We currently lack a cognitive computational account of how this context sensitivity works. I suggest that PP is well suited to provide a mechanism for context sensitivity. The PP-specific, so-called precision weighting apparatus supplies the means of how concept features or information pieces can be switched on and off depending on the context. I argue that the knowledge of which feature is relevant in which context is stored as more knowledge in the same PP model, namely as second-order knowledge. So, we can
use the same mechanism to account for relevance. Precision and relevance are faces of the same error weighting coin.

Chapter 4 tackles the current debate about the format of concepts. "Format" is understood here as whether concepts are represented as amodal symbols – like linguistic signs on which rule-like operations are carried out (following Fodor's LOTH) – or whether they are sensorimotor representations – like images or movies albeit not limited to the visual sense modality and also potentially in abstracted forms. I argue that there are two problems with regard to drawing the purported distinction between the modal and amodal. First, evidence can be interpreted as supporting both modalism and amodalism. Secondly, there is no agreed-upon definition as to what the notion of amodal or modal actually amounts to. I propose a solution by avoiding the dichotomy in the first place. The picture proposed is that concepts are hierarchically structured networks, identified by the "root node" of a complex tree-like structure of nodes (realized, e.g., as neural assemblies). The higher in the structure the concept is, the more abstract/compressed the information is to which this node is sensitive. The idea is that concepts can be instantiated with different degrees of abstraction. In its more abstract instantiations, the concept becomes increasingly "amodal".

Chapter 5 takes up a second long-standing debate about concepts, namely how the information inside the concept is structured. Traditionally, concepts were considered to be definitions represented in some propositional form. However, this view has long been abandoned in favour of views in which concepts are prototypes, sets of exemplars, or little theories that encode common-sense knowledge about the category that the concept denotes. However, it turns out that no single theory can accommodate the available empirical data. Consequently, some have endorsed eliminativism in regard to concepts, i.e., the view that we should eliminate the notion of a “concept” from scientific vocabulary. Others defend pluralism (each format corresponds to a different concept) or hybridism (a concept is a complex structure that contains the different formats). The current hybrids turn out to be problematic and I propose an improvement of Vicente & Martinez Manrique's "coactivation package" hybrid. I argue that each of the three classical formats (prototype, exemplar, and theory) arises as different ways of processing the same unified representational structure.
Part 3 contains Chapters 6 and 7 and builds on the account of concepts developed in Part 2. I propose how we can account within PP for language and logic and in particular for the compositionality of conceptual thought (under some modified Language of Thought paradigm). Language and logic are two flagship higher cognitive competencies of the human mind and if PP could accommodate those competencies, it would underscore that it is a fruitful cognitive computational paradigm.

Chapter 6 tackles a *prima facie* challenge that higher cognition poses to PP. Our intuitions about compositionality of thought and language—so I suggest—are (often tacitly) based on a structure of language understood as consisting of two independent parts: a lexicon and a set of formal syntactical rules. This corresponds to a Chomskian-style generative grammar (GxG), which in turn has led to Fodor's LOTH for conceptual thought. As PP cannot easily be mapped with generative grammar and LOTH, one might think that PP cannot account for compositionality of language and thought. I propose that if we change the language paradigm to Construction Grammar, arguably the main rival of GxG, then this prima facie concern loses force.

Chapter 7 builds on the Construction Grammar approach from Chapter 6 and suggests how formal logical thought might arise, namely as the processing of modal but highly abstracted/compressed representations. The idea I defend is that logical rules do not exist as truths in some Platonic world, rather, they are dialogical patterns represented as "constructions", like words or grammatical patterns. In other words, logical rules are representations one step above the level in the hierarchy on which sentences (situations) are represented. Logical rules are hence just more conceptual representations, though they are more abstract/compressed than words and sentences, and hence higher in the hierarchy. This idea closely follows Dutilh Novaes’ (2012, 2020) proposal that logic has a dialogical origin.

Part 4 attempts a series of applications of the view of concepts and language developed in the two previous parts. The objective is to test the PP account of higher cognition by trying to shed light on conceptual-linguistic problems. Specifically, I deal with the semantic paradox (Liar Paradox), metaphor and copredication.
Chapter 8 tackles one of the most famous and widely discussed paradoxes in the history of philosophy and logic, the Liar Paradox. The Liar Paradox (or variations of it called generally "semantic paradoxes") are normally treated as formal logical puzzles. Many solutions to the paradox imply the revision of classical logic. I pursue a different and novel avenue, by claiming that the paradox arises from the way we cognitively process the Liar sentence.

Chapter 9 fleshes out an account of metaphor within PP. There is an ongoing debate as to how metaphors are cognitively represented and processed. The two main contenders are the "Category Inclusion View" and the "Implicit Comparison View". Some scholars argue that the former is slightly better supported, but a hybrid is probably needed. I argue that PP better supports the Category Inclusion View and makes a hybrid account unnecessary. This leads to a conditional claim: if PP were on the right track, it would support the Category Inclusion View. But, I argue, this also works the other way round. If the Category Inclusion View were to receive more confirmation, this would support PP, because it can supply a cognitive computational underpinning.

Chapter 10 deals with copredication. Copredications are statements where the same nominal is used with two predicates invoking two incompatible senses. For instance, in "The book is heavy and informative," 'book' has a physical and an abstract (content) sense with 'heavy' and 'informative' respectively. Copredication raises a puzzle for traditional truth-conditional semantics. In truth conditional semantics the truth value of a sentence is a function of the semantic values of the components of the sentence. The question that arises for copredication is then how composition works if we have one nominal that refers to two different and incompatible entities at the same time. We (Guido Löhr and I) suggest going down a cognitive rather than a truth-conditional route, based on the account of concepts and language from the previous chapters. We focus on the question of how the linguistic intuitions about the degree of acceptability of copredication sentences arise in the first place. The view we propose could be considered a further development of an existing psychological account, the "coactivation package" account by Vicente & Ortega Andrés. On this account copredication sentences can be felicitous because the two incompatible senses form part of a unified package and are both immediately available when the sentence is processed. Such a further development is necessary, because the
coactivation account cannot deal with certain asymmetries and the context-sensitivity of acceptability judgements. For instance, there are copredication sentences that have the same two senses in the same order, but different degrees of acceptability like:

"The newspaper has been attacked by the opposition [institution] and was publicly burned by the demonstrators [physical object]."

"The newspaper has been attacked by the opposition [institution] and fell off the table [physical object]."

Part 5 concludes the thesis. I engage with some critics of PP approaches to higher cognition and PP theorizing in general. I provide some responses and highlight where further work is necessary.

Chapter 11 discusses Williams' 2019 and 2020 papers, which cast doubt on the suitability of PP as an approach to higher cognition on theoretical grounds. I will focus on three problems Williams sees with PP regarding higher cognition. Williams points out as a first problem that conceptual thought is "richly compositional" and that the PP model cannot account for this. He also takes two issues with the idea of a homogenous hierarchical structure of the PP model: for him it is unclear what exactly the hierarchy tracks, and he states that, in any case, a more modular structure is needed. I synthesize the main arguments and provide some tentative responses.

Chapter 12 deals with some objections from Litwin & Milkowski (2020) to theorizing within the PP paradigm, as well as two additional concerns affecting core commitments of the PP paradigm. Litwin & Milkowski object that PP in general has been applied too quickly to too many domains. Rather, the PP community should focus on first resolving various fundamental issues of the PP framework. Specifically, they believe that the application of PP should be limited to the sensorimotor domain and should not be extended to higher cognition. Another fundamental issue that affects core commitments of the PP framework in general has been mentioned by Clark (personal conversation) and is related to the role of the so-called "prediction error weighting mechanism" (see Section 2.1.3), which might be seen as a "magic modulator" that can be used to explain everything. A final issue I discuss is the concept of "active inference" (see Section 2.1.2 e) and the problem with fleshing out
the difference between a desire and a belief within PP. I summarize all those objections and provide tentative responses. I also suggest areas for further work.
Chapter 2. Setting the stage: Predictive processing and higher cognition

2.1. The PP paradigm

In this section I lay out what I take predictive processing (PP) to be, namely an emerging paradigm (as opposed to a theory) of how brain-body systems work in uncertain environments. Because PP—as understood here—is an amalgam of concepts, principles, and mechanisms, not all of which are original, I describe the main precursor ideas that found their way into this paradigm. Then I characterize PP's core commitments using Marr's three levels, following Sprevak (2021a-d). Finally, I provide a very brief overview of the breadth of domains in which the PP framework has already been used to do explanatory work.

A key conclusion from this section is that despite the apparent applicability of PP to many types of higher-level phenomena, higher cognition (especially conceptual thought, language, and logic) is not, so far, a thriving field for PP. It is also striking that no detailed discussion or account of concepts is available within the PP framework. Given that concepts are among the most central posits in cognitive science, the lack of such an account might explain why higher cognition has not been tackled to a larger extent within the PP community so far.

2.1.1. PP as a paradigm, not a theory

In this thesis, I take predictive processing to be an emerging paradigm or research program in cognitive neuroscience. In a nutshell, PP pictures the brain as an embodied prediction machine, that tries to get better and better at anticipating its sensory input such that it favors its survival in an uncertain environment. This is achieved by maintaining a mental model of the world (a model about the causes that give rise to the sensory input stream). The model is being constantly improved to minimize prediction error on average and in the long run. This prediction error

2 Note that PP is referred to by different authors using different terms (e.g., "prediction error minimization" by Hohwy, "predictive coding" by Sprevak, "predictive processing" by Clark, "free energy minimization" (FEM) and "active inference" by Friston). I stay with Clark's nomenclature for the overall paradigm: predictive processing. It is not always clear what the exact relation is between PP and the other concepts. For instance, PP is called by Friston (2018) and Clark a "process theory" of FEM. Clark (2022) also speaks of an "active interference formulation of predictive processing". For this reason, I will later make clear what specific commitments I take to be associated with PP.
minimization approximates Bayesian inference from perceptual input to its most likely causes. Here "embodied" can best be understood as meaning that perception, cognition, and action are closely integrated making use of the same representational and processing principles. Cognition is, therefore, essentially shaped and constrained by the body and its sensorimotor apparatus interacting with the world.

A paradigm, as I define the term for the current purposes, is a set of concepts and principles that guide and constrain the building of specific theories of some cognitive phenomenon. This definition follows the notion of a paradigm in science coined by Kuhn in its broader sense and is also referred to by him as a "disciplinary matrix" (e.g., Kuhn & Hacking, 2012, p. 181). One such element of the disciplinary matrix, is a paradigm in a narrower sense, namely an exemplar of good science (see also Bird, 2018). The important point is that—in this dissertation—a paradigm is not to be understood as an empirically verifiable theory itself. What is empirically verifiable is a theory (of some phenomenon) within the PP paradigm. The reasons why PP should not yet be seen at this stage as a mature theory of brain-body systems are manifold. Let me stress three of them.

Firstly, PP is not well-defined but is rather used as an umbrella term for a wide range of approaches. Allen & Friston (2018), for instance, point out the large variety of approaches under the PP label, which range from cognitive approaches based on "modular internalistic mental representations", to moderate versions with "body representations" and "radically enactive, embodied and dynamic" (p.2459) approaches.

Secondly, there is no agreement about the ambition or scope of PP. In a more ambitious expression, PP aspires to be a unified theory for perception, cognition, and action: a grand unified theory of the brain (e.g., Friston et al., 2010). In its less ambitious version, PP is just a part of a more complex story of how the brain works, and covers only some aspect, for instance vision and maybe other lower-level perceptual processes (e.g., Rao & Ballard, 1999; Jiang & Rao, 2021). It has also been suggested that PP can reach beyond the agent bounded by brain and body and provide a model of "extended cognition" (e.g., Kirchhoff & Kiverstein, 2019), in which external parts of the world form part of a cognitive system (see also Clark & Chalmers, 1998). PP principles are even suggested to govern whole agent-evolutionary niche systems, or even larger systems. The whole of evolution itself can
be seen as a "PP engine" (Allen & Friston, 2018, p. 2467; Fernando et al, 2012; Harper 2011).

Thirdly, PP is still significantly underspecified in many aspects on the algorithmic and implementational levels (see Sprevak, 2021a-d). For instance, even on the highest level of description, the computational level, not all theorists agree that prediction error minimization is the only computational task the brain is carrying out. But especially on the algorithmic and implementational levels the number of lacunae is large (Jiang & Rao, 2021; Millidge et al., 2021; Sprevak, 2021a-d).

That PP is not well-defined, is underspecified, and that there is a lack of agreement on its ambition and scope should not prevent us from trying to extract a largest common denominator version of a PP paradigm, articulated as a set of minimal commitments. I will do this in a moment.

Much like theories, paradigms are also subject to certain desiderata. Here is a tentative list. Firstly, a paradigm should be fruitful. Fruitfulness should be understood as the capacity to generate, in a first step, new and interesting hypotheses and perspectives on target phenomena. In a second step, the paradigm should lead to the development of specific theories that can be exhaustively empirically scrutinized. Secondly, a paradigm should have unifying power. This implies that it should consist of a limited and coherent set of concepts, principles and mechanisms that have a large scope of application. This goes hand in hand with the third desideratum, simplicity. The framework should be simple and integrated and not, for instance, a patchwork of many ad-hoc principles. Another desiderata is coherence. The pieces that constitute the paradigm should not be disjunctive ad-hoc elements. They should fit together in some coherent way. Finally, the constraints it provides should be consistent with existing empirical data. I hope that the present work contributes to making the case that the PP paradigm promises to meet those key desiderata.

2.1.2. PP as a package deal

It is crucial to understand that PP is not entirely original where its constituting individual concepts, principles, and mechanisms are concerned (Sprevak, 2021a).

---

3 See the discussion in Chapter 12 regarding whether consistency is enough to support the confirmation of a theory.
Rather, it is the specific combination of various of those elements that constitutes the PP paradigm. Hence, PP can be seen as a package of a range of ideas combined (in, I think, a conceptually exceptionally neat, elegant, and unifying way).

Among the key ideas that have found their way into the PP paradigm (as understood here) are the following:

a) Perception as inference

The idea that perception is an unconscious inferential process was first proposed by von Helmholtz (1867, p.430). Stimuli that arrive at our sensory apparatus are ambiguous and can be interpreted in many ways. This leads to the "Problem of Perception" (Hohwy, 2018), or the "Inverse Problem": sensory stimuli underdetermine their causes in the external world. For instance, consider a round shape projected onto the retina. It could have been produced by a sphere or a round plate, or a cylinder seen from below, just to mention some possibilities. How can the brain solve this problem of perception, given this massive underdetermination of sensory information? Merely bottom-up processing seems insufficient. A possible solution, however, is to complement stimuli with prior knowledge, beliefs, or biases to produce good enough hypotheses about the causes of the round shape. Those go beyond the immediately given, i.e., what is projected on the retina. This implies that often what we perceive is the result of the influence of prior beliefs or biases.

b) The Bayesian brain hypothesis

The idea that perception is unconscious inference has been generalized to the idea that cognition more generally is inference. Within a simple, elegant, and increasingly influential framework, the mind is a Bayesian brain, i.e., cognition consists largely in carrying out Bayesian inference (see, e.g., Chater et al., 2010). The Bayesian approach provides a formalization of how beliefs (broadly understood and including, e.g., perceptual hypotheses/representations) are updated based on new incoming evidence.

There is much evidence that we are (at least approximate) Bayesian perceivers and thinkers. Bayesian theory prescribes a rational way to update beliefs if we have incomplete and uncertain information. This is especially suitable for real agents with noisy and incomplete information about their environment arriving from their sensory
organs. The Bayesian framework describes how we need to change a belief (represented as some probability distribution over the hypothesis space) given some new evidence. Much empirical behavioural data that has been interpreted as being irrational turns out to be perfectly rational under the Bayesian paradigm (see Gershman, 2021, for a very detailed and nice discussion).

c) A hierarchical Bayesian brain

However, it is not clear how a simple Bayesian approach could give rise to the complex behaviour and knowledge structures we seem to be able to represent and process in the mind. An extremely fruitful idea, championed especially by Griffiths, Kemp, Lake, Tenenbaum and others (e.g., Griffiths et al., 2007; Kemp & Tenenbaum, 2009; Lake et al., 2015; Tenenbaum et al., 2011) is to stack Bayesian models into a hierarchy. In this way, the brain can represent complex structures at different scales and levels of abstraction. This also implies that we can "meta-learn". For instance, we represent the world as composed of objects. This can be seen as a higher-level belief (hyperprior) that conditions/constrains what we then perceive and how we conceptualize the world (namely mainly in the form of objects). So, there is a whole hierarchy of inductive biases on different levels of abstraction that structure the brain’s model in an extremely powerful way. The conceptualization of the model as a hierarchical Bayesian model sheds an entirely new light onto how we perceive, think, and (as we will see in a moment) act. The idea of a model consisting of representations on an abstraction gradient is extremely central throughout this thesis.

d) A generative model: the centrality of anticipation and prediction

A hierarchical model can be used in different ways. For instance, it may be used in a bottom-up/passive way: the brain just reconstructs by feature aggregation what it perceives. But another way in which the model can be used is to anticipate its perceptions. This leads to the idea that the brain is a prediction machine that uses a hierarchical Bayesian model.

Prediction/anticipation has various advantages and might be very plausible as a fundamental operating principle for how the brain works. Firstly, if the brain proactively predicts its sensory input, it can compare it to the real input and then it
needs only correct deviations—if there are such deviations. In this way, the brain avoids processing the whole stimulus every time bottom-up from scratch. Only prediction errors need to be processed and passed on to higher levels in the model. In other words, the model needs only to react to the unexplained parts of stimuli. This is a very efficient way of representing the world if there is sufficient regularity in the input. This is the idea of how, for instance, video compression works. Instead of representing whole frames, compression techniques allow for storing only the changes from frame to frame. Secondly, neural transmission introduces a lot of latencies which might slow down our ability to react with the necessary speed. For instance, by anticipating expected proprioceptive signals when executing motor commands we can overcome issues with latencies, provided, of course, that we predict correctly sufficiently often. The speed argument is especially relevant in action, which I bring into the picture in a moment.

The centrality of prediction or anticipation in cognition, is, however, not original to PP. The idea has many precursors and can be found in writings of early cognitive psychologists, for instance in the form of "analysis-by-synthesis"⁴ (e.g., Halle & Stevens, 1959, 1964; Neisser, 1967), "hypothesis testing" (Bruner, 1951, Gregory et al., 1980), "prediction making" (Craik, 1943), “trial and check” (Solley & Murphy, 1960), and so on. Note also that other writers claim to have thought in terms of predictive minds as early as 1986 (e.g., Hawkins, 2021).

What might be original to PP is to declare that prediction error minimization is a central, if not the only, computational principle that drives the brain's activity, and to rely on generative hierarchical Bayesian models.

e) Active inference

One of the most beautiful and unifying ideas in the PP paradigm is that it pictures perception and action as complementary aspects of cognition, implying a very broad and inclusive notion of cognition.⁵ This means that perception and action are represented and processed in essentially the same way, and only differ in the direction of fit of the prediction between the brain-body system and the world.

⁴ See also Bever & Poeppel, 2010.
⁵ But it is also controversial, see Chapter 12 for the brief discussion of an objection.
According to the embodied cognition paradigm, our conceptual system is shaped by our physiology, our sensors, and our body and how they interact with the world. This idea was initially spearheaded by writers like Johnson, Lakoff, Varela, etc. (e.g., Lakoff & Johnson, 1980a-b, 1999; Lakoff & Núñez, 2000; Varela et al., 1992). Embodied cognition in this generic sense is fully endorsed by PP because its focus is on the brain-body system, which aims to survive in an uncertain world.

The novelty contributed by PP (specifically by Friston), which spells out what embodied cognition amounts to computationally, is the simple idea that motor commands are also inferred, much like perception is inferred. In this kind of motor inference, also called "active inference" (in contrast to "passive" or "perceptual inference" in the case of perception)\(^6\), the model predicts the proprioceptive signals that would result from the intended action. What is also unique to PP is that it is based on the idea that the brain is driven by a constant process of prediction error minimization. Prediction error can be minimized by either updating the model of the world (perceptual inference), or by changing the world such that it fits the prediction (active inference). In the case of active inference, to minimize the error of the anticipated proprioceptive signals, the body must carry out the intended action. For instance, to grasp an apple, the mind would predict how it feels to move the arm towards the apple, grasp it with the fingers and hold it in the hand. Before this action sequence, the brain does not receive the corresponding proprioceptive signals and a large prediction error is the consequence. To suppress the prediction error, the body must move the muscles in just the right way. In effect, the mind-body system is carrying out a self-fulfilling prediction. To make this work, the top-down prediction of the target proprioceptive states needs to be kept fixed in this case.

In this view, action and perception are tightly coupled dual aspects of the predictive brain-body system (e.g., Friston et al., 2010; Friston et al., 2011; Friston et al., 2017; Fountas et al., 2020). This tight integration of perception and action converts PP into a paradigm of potentially very broad scope.

---

\(^6\) The notion "active inference" is used increasingly in a more inclusive sense including both perceptual inference and the inference of actions/policies (often in connection with Friston's free energy framework). However, I use it here in the narrow sense of inference of actions/policies.
An idealized "genealogy" of core ideas in PP

Table 2.1 summarizes in a simplified way the core ideas just discussed and the logic of how more and more layers of novel and radical ideas have been added on von Helmholtz's initial idea of perception as inference. Those lead us to the PP paradigm, as we will now see.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Key idea</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Perception as inference</td>
<td>Solution to the underdetermination problem of perception</td>
</tr>
<tr>
<td>b) Cognition as Bayesian inference</td>
<td>Bayesian inference is an evolutionarily rational way to perceive and think</td>
</tr>
<tr>
<td>c) Cognition as hierarchical Bayesian inference</td>
<td>Allows for the most complex structures to be inferred and represented; can implement meta-learning</td>
</tr>
<tr>
<td>d) Cognition as prediction using hierarchical Bayesian inference</td>
<td>Anticipation in cognition (economy, speed)</td>
</tr>
<tr>
<td>e) Cognition and action as dual aspects (&quot;active inference&quot;) in predictive inference with hierarchical Bayesian models</td>
<td>Perception and action are both inferred</td>
</tr>
</tbody>
</table>

Table 2. 1: Simplified and idealized cumulative "genealogy" of core ideas leading to the PP framework.

I will characterize the PP paradigm in terms of what I take to be its largest common-denominator commitments. I have relied on the following sources. Arguably the locus classicus of a large part of the PP paradigm is Rao and Ballard's (1999) treatment of vision. The first book-length treatment by Hohwy 2013, as well as Clark's 2013 target article and 2016 book, are crucial for the establishment of PP as a more general framework for brain-body systems. Furthermore, Friston has played a central role, through countless papers, and is doubtlessly among the most influential and productive proponent of PP (in its active inference or free energy formulation). Furthermore, recently there have been longer critical or encyclopedic treatments of the framework on which I also rely, like Sprevak (2021a-d) and Jiang & Rao (2021), as well as Millidge et al. (2021). Additionally, there is now a huge number of
introductions to the PP framework in most PP related papers (many of them relying on Hohwy, Clark and Friston). Especially useful are, e.g., Williams (2019, 2020, 2022).

2.1.3. A characterization of PP via Marr's three levels

A useful way to characterize the minimum set of commitments that constitute PP, as I understand it in this dissertation, is by appealing to Marr's three levels of description (see Sprevak 2021a-d). The three levels respond to three types of questions: What is the problem the system/device is designed (or has evolved) to solve (= computational level)? How does it do it (= algorithmic level)? How is the system physically implemented (= implementational level)? As I take PP to be a paradigm, a full specification on the algorithmic level and the implementation level is not necessary. Doing so would be developing a specific theory. A paradigm provides constraints, in the form of concepts, principles and (schematic) mechanisms, not a full specification. The commitments cut across all three levels, and those commitments and their implications set up guiding and constraining factors for the development of more specific theories of aspects of higher cognition. Therefore, the level of discourse in the rest of the dissertation is conceptual. Here, then, is a characterization of the PP paradigm in terms of Marr's three levels according to Sprevak (2021a-d). As we will see, the core ideas from Table 2.1 are all included. But there are some PP specific additions, most notably the idea of precision weighted prediction error minimization.

1) Computational level

According to the PP framework, the computational problem that the brain-body system must solve is long-term error minimization of the predictions of the system's own sensory input. Sensory input needs to be understood in the broadest sense, including exteroceptive but also interoceptive and affective signals. One might wonder why the brain has evolved to solve precisely that problem. How could such a simple principle give rise to the very complex brain anatomy and behaviour of humans? Wouldn't "facilitating survival" or something similar be a more appropriate task description? That in turn would plausibly require the brain to solve many different computational problems. Friston (e.g., 2010) has proposed that survival—understood as the avoidance of entropic disintegration—is closely linked to the
minimization of some information theoretic quantity, the so called "free energy" (FE). It is by minimizing prediction errors in a model of the environment that free energy is minimized. Intuitively, in this way "surprisal" (i.e., unexpected states of the organism) is minimized, which allows the systems to keep their parameters in a corridor of survival-feasible values. For instance, a too high or too low environmental temperature is bad, too much or too little blood sugar is bad, etc. Being outside of a feasible range of values of certain external and internal parameters leads to a phase transition that dissolves the identity of the individual. So, the idea is that the brain-body system is tuned to act in such a way as to (correctly) predict that it will be in such feasible windows of those parameters that are critical for survival.

While this is a fascinating proposal, for current purposes I will not include the FE principle in the list of commitments of PP. The reason for this is twofold. Firstly, not all PP theorists endorse it whole-heartedly (e.g., Clark does not put it at the centre of his PP pitch). Secondly, the suitability of the FE formulation has recently been questioned (Williams, 2022), so it is not uncontroversial. A defence of the FE principle would clearly exceed the scope of the present work and is also not strictly necessary for the current purposes.

2) Algorithmic level

The algorithmic level provides a mathematical description of how to solve the computational task. Within PP, the brain-body system achieves long term prediction error minimization of its sensory input by means of inferences and continuously improving a specific prediction model, namely a so-called hierarchical generative probabilistic model, combined with a prediction error weighting mechanism. This is a model of the environment and of the organism itself.

a) Hierarchical generative probabilistic model

The model is *generative* because it generates hypothesis (or predictions). Those predictions flow top down in a *hierarchical* structure. The model is hierarchical in the sense that it consists of multiple layers of (predictive) representations that are interconnected. Hypotheses/predictions are encoded by variables representing different spatiotemporal scales. The model tries to replicate a causal model of the environment, i.e., the way the environment changes over time. Higher level representations serve as constraining "priors" for lower levels. The system attempts
to bring itself into an overall error minimizing state on all levels in the hierarchy at the same time. Notice that a prior is not something that is fixed. In (approximate) Bayesian inference, a prior is updated into a posterior which then serves as a new prior for subsequent inference. The model is *probabilistic* because the hypotheses/predictions are represented as sufficient statistics or probability distributions. Notice that this probabilistic aspect per se will play a less prominent role in this thesis.

b) Precision-weighted prediction error minimization

The top-down influence/flow of information is central to PP, but it is just half of the story. In the PP model, there is still crucial bottom-up flow of information. However, the twist in PP is that only *error signals*, not the full information, are passed upwards (while predictions are flowing top down). This is a very economical way of operating the brain and is also sometimes called "predictive coding". As briefly mentioned already, we can get an intuition for it by considering data compression of videos. In the transition from one frame to another not too many changes occur, so it makes sense not to store each full frame, only the differences. One can focus on processing the new, unexpected information rather than everything the model knows already. Errors-only processing allows for dramatically compressing the size of storage needed and speeds up data transmission.\(^7\)

The error minimization process ultimately approximates Bayesian inference. As mentioned already, Bayesian inference is a rational way to update prior beliefs in the face of new evidence (or prediction errors). It is well known that exact Bayesian inference, however, is computationally intractable, so instead the brain carries out an approximation to Bayesian inference (Hohwy, 2020, p.211).

This view leads to a crucial question: How much prior knowledge should be used versus the bottom-up input? How can the influence of top-down and bottom-up information be correctly balanced in case of discrepancies? Apart from having a bad model, error could also be due to the unreliability of the signal. Therefore, the model should not be updated in all cases of prediction errors. If there is fog, the visual input of a dog-like shape may not be very reliable, because the input may also be

\(^7\) Sprevak (2021c) correctly points out that not in all cases is it advantageous to use only difference or error signals.
compatible with a large cat. So, the error signal needs to be reduced when the visual signal is not reliable, and we need to trust the prior knowledge more (which hypothesis has a higher prior probability given the circumstances, the cat, or the dog?). In terms of probability distribution, the reliability can be measured by the inverse variance (the spread) of the distribution. To accommodate those considerations, PP posits a modulation mechanism of the error signals. The less reliable a signal, the more the induced prediction error is attenuated before being passed upwards for further processing.

At this stage, a methodological note is in order. I have described the algorithmic level in relatively informal terms. That should suffice for a characterization as a paradigm. The architecture can be formalized mathematically in many ways, and the PP community has not converged on any universal detailed algorithmic description so far (Sprevak 2021b-d). In this thesis I will not provide formally specific algorithms for the higher cognitive capacities considered. The purpose of the present thesis is to dispel more fundamental concerns about whether PP can be scaled up to higher cognition. For that purpose, it does not matter, for instance, by which exact algorithm prediction errors are minimized.

3) Implementational level

According to PP, the brain's world model is implemented neurally by a hierarchical structure of pairs of representation and error units in the form of neural assemblies in the brain. I will call those pairs "prediction units". These are the physical locations where predictions are made, and errors calculated. The error signal is fed upwards, and the prediction signal downwards for further processing. The core idea is to posit certain "canonical neural microcircuits", which are the building blocks of the PP brain circuitry. But the commitment is not necessarily that the whole model is composed of one single (physical) type of microcircuits.

Again, it is still largely unsolved how exactly the prediction units are implemented. Also, the exact nature of the connectivity between layers remains open (as well as the exact inhibitory and excitatory nature of the connections). One specific relevant question is whether there is a cleanly layered hierarchy, in which each level is connected only to the adjacent layers one up and one down. Various PP theorists seem to assume such an idealized structure. However, evidence from the structure
of neural connections in real brains points to a much more complicated connectivity with many connections skipping layers. Also, insights from machine learning, even if not straightforwardly transferable to a real brain, might suggest that there could be advantages of connections that shortcut layers (the so-called residual deep neural networks with skipping connections can increase the performance of deep networks, e.g., He et al., 2016). Rather than a clean hierarchy the brain might, rather, be a messy "heterarchy" (see Millidge et al., 2021).

The PP paradigm, which I have tried to articulate here on a largest common denominator basis, is both contentious and leaves many features open and unspecified. However, this is not a problem for the type of conceptual exploration I want to undertake in the rest of this work. I will discuss in more depth some of the objections and concerns brought forward regarding PP in Chapters 11 and 12. It should suffice here to point out one complaint that is often heard. Some critics take issue with the epistemic status of PP. What evidence do we have that PP is correct? Let me just insist again that empirical confirmation of a paradigm is somewhat of a misnomer. To understand my point, consider other cognitive computational paradigms, like Fodor's LOTH again. Clearly, Fodor did not put forward a very specific cognitive theory. His description of how the mind works left many features open, and a notorious problem was its psychological plausibility (Schneider, 2011). However, was LOTH an unfruitful paradigm for that reason? Not at all. Much work in cognitive science tacitly still adopts the LOTH-paradigm. Many, if not most, attempts to formally model cognition somehow assume that it can be described in amodal-symbols-plus-rules terms. Similar points apply to the fruitfulness of many other cognitive oriented paradigms, like Relevance Theory (Sperber & Wilson, 1995). Their fruitfulness did not hang on an immediate decisive empirical proof that they were correct.

Therefore, I will not undertake a detailed defence of PP based on a review of empirical evidence. I will assume that it is in the right ballpark and then see what we can do with it in the realm of higher cognition. A lot of successful work has been done under the PP paradigm and there is a lot of evidence which shows that the principles of PP seem in fact to operate at least in parts of the brain (especially the visual or motor areas) (see, e.g., Clark, 2016; Jiang & Rao, 2021; Walsh et al., 2020 for comprehensive overviews). But when analysing the range of applications of the
PP framework, one observes that one domain is strikingly underrepresented, namely higher cognition. Let me therefore turn to examining in which domains PP has already been applied to.

2.1.4. The scope of PP applications

In this subsection I briefly survey the scope of the domains or problems where the PP paradigm has been used for explanatory work. I will not be exhaustive; rather the purpose of this subsection is to make two points. Firstly, as just mentioned, I suggest that PP is indeed scalable as its concepts, principles and mechanisms can be explanatorily deployed on different levels of description: from small scale perceptual phenomena up to speaker communities as coupled systems. This I find quite stunning, as to the best of my knowledge, there is no other cognitive paradigm with that capacity. (There is a flip side to it. Some critical authors (e.g., Litwin & Miłkowski, 2020) suggest that a proliferation of all sorts of rather verbal and conceptual theories under the PP umbrella is not helpful. This is discussed more in Chapter 12).

Secondly, despite some existing work, surprisingly, higher cognition is not thriving so far as a domain of application of PP. This lack of treatments of higher cognition under the PP banner is precisely the gap I want to examine and address in this thesis.

In a recent review, Hohwy (2020) provides a useful overview of philosophically oriented research output related to PP, which has grown substantially in the last few years, and this growth appears to have accelerated even more, including in the literature in the special sciences. The PP literature manifests an impressive scope of domains of application. Here there is no space for a full review, but it is worth mentioning a few examples to show its thematic breadth.

Areas of application of PP concepts and principles cover specific perceptual phenomena, for instance, binocular rivalry (Tong et al., 2006; Hohwy et al., 2008), the Cornsweet effect (Brown & Friston, 2012), and visual art perception (e.g., Van de Cruys & Wagemans, 2011) and action/motor phenomena (e.g., Friston et al., 2010; Wiese, 2017a). PP has also been used to explain pain (Fardo et al., 2017), as well as certain psychological and psychopathological phenomena, like autism (e.g., Pellicano & Burr, 2012) and delusions (e.g., Bortolotti & Miyazono, 2015). Recently there has also been an increasing interest in consciousness (e.g., Hohwy & Seth,
2020; Marvan & Havlík, 2021; Schlicht & Dolega, 2021; Seth & Hohwy, 2021; Whyte, 2019; Wiese, 2018). Those examples only touch the surface, and Hohwy (2020) provides many more. But beyond philosophically oriented research, the literature in special science that use PP approaches is growing strongly. PP approaches can be found in neuroscience, linguistics, even literature.\(^8\)

However, despite this impressive scope, some phenomena of higher cognition, namely language, logic, and conceptual thought have played only a marginal role in the literature so far. I will give a slightly more detailed overview in Section 2.2. Here it suffices to make the point that there is, of course, some incipient work close to PP or directly under the banner of PP, e.g., related to language (Pickering & Garrod, 2013; Pickering & Clark, 2014; Lupyan, 2012; Lupyan & Clark, 2015; Rappe, 2019, 2022; Murphy et al., 2021), irony (Fabry, 2021), and literature/poetics (e.g., Kukkonen, 2020, 2021). There is also, of course, and to a much larger extent, work with approaches that share some commitments (but not all) of the PP paradigm, like Bayesian approaches to higher cognition. However, the point I want to make is that treatments strictly under the PP banner that endorse most of PP's commitments (its "package") are limited. Also, more importantly, increasingly there are sceptical voices with regard to the outlook that PP is suitable at all to account for higher cognition (as already mentioned, e.g., Williams, 2019, 2020, 2022; or Litwin & Miłkowski, 2020)— I address those in Chapters 11 and 12.

Almost completely absent is any more detailed discussion of concepts under the PP paradigm (and even under Bayesian or other PP-affine approaches) that goes beyond identifying them simply with nodes or variables (or some specific representational format like visual patterns). This is quite surprising as concepts are generally considered to be among the most central explanatory entities in cognitive science. In the PP literature, sometimes concepts are just considered to be the latent variables represented by nodes in a generative model (e.g., Smith et al., 2020), and no more detail is provided, for instance, on whether they have an internal structure.

\(^8\) Note that much work appeals to "predictive processing/coding" but brings to bear only a few of the PP commitments, most prominently the predictive nature of cognition. That some phenomenon manifests that something is anticipated/predicted does not make it a "PP account" under the notion of PP as I use it here.
and which one, what a node exactly represents and how the enormous wealth of existing empirical concept research can be accommodated.

The hypothesis motivating this thesis is that one reason why higher cognition in all its richness has not been tackled so far more broadly might be precisely that there is a lack of an account of concepts within the PP framework. Once concepts are characterized in some more detail, it should be possible to unleash a fruitful and novel examination of many phenomena in higher cognition under the PP paradigm.

Conclusion of Section 2.1.

PP is considered here to be a paradigm, not a theory. PP should be seen as a mechanistic neurocognitive paradigm, in contrast to connectionism or classical computationalism. This implies that it is committed to obeying constraints from neuroscience because its commitments cut across Marr’s three levels, including the neural implementational level.

PP provides a conceptual picture that elegantly integrates perception, cognition, and action, i.e., the key aspects of fully embodied cognitive agency. The potential explanatory scope is the broadest possible. PP is a framework that has already been applied to phenomena on different scales, from visual phenomena to irony. On this count, PP should be able to say interesting things about other aspects of higher cognition, including compositional conceptual thought, language, and logic.

As already mentioned, the reason why accounts of conceptual thought, language, and logic under the PP banner are scarce might have to do with the lack of an account of concepts within PP; concepts after all are central posits in cognitive science. The traditional concept literature in philosophy, psychology, neuroscience, and linguistics is enormous and rich and almost no connections have been made so far between PP and that tradition.

But there might be another, or a complementary, explanation for why higher cognition research under PP is not blooming. PP might not be suitable to account for higher cognition, especially for those competencies where compositional conceptual thought is relevant. There are intuitive or prima facie, but also deeper theoretical reasons, as we will see. This would be a complication for any project tackling higher
cognition within PP, so let us turn to analysing those purported reasons why research of higher cognition under the PP framework has not taken off so far.

2.2. The challenges from higher cognition

In this subsection, first I briefly review the available work addressing higher cognition within the PP paradigm, as well as other work, close to or relevant for PP (Section 2.2.1). This brief review provides the context for the proposals I will develop here, and points to the gap I want to tackle in this dissertation. I conclude that while the range of areas to which the PP paradigm has already been applied is broad, higher cognition, especially conceptual thought, language, and logic, is largely underrepresented. Especially striking is that there is no detailed account for, and little discussion of, concepts within the PP framework.

I then formulate three challenges from higher cognition for PP that might be hypothetical reasons for this underdevelopment: the prima facie challenge, the theoretical challenge, and the instrumental challenge (Section 2.2.2). By articulating those challenges, I set the stage for the rest of the thesis, in which I want to develop some key elements of an account of higher cognition within PP. If such an account can be put forward, then there are no strong reasons to think that those challenges cannot be met. After all, PP promises to be a fruitful framework for higher cognition.

2.2.1. The state of the art: PP and higher cognition

Some of components that found their way into the PP paradigm (see Section 2.1.2) are active fields of research on their own and a lot of work about higher cognition has already been carried out. We, therefore, need to differentiate between work done explicitly under the PP banner and work that is in the vicinity of PP, in the sense that it is compatible with PP because it shares one or various core commitments. I will structure the following review of relevant previous work on higher cognition into three blocks: a) embodied cognition, b) Bayesian cognition and c) work about the nature of language and psycholinguistics.

a) Embodied cognition

As pointed out, PP is located within the broader embodied cognition tradition. Embodied cognition is an umbrella term for an extremely large and diverse field of
proposals that is united by the general idea that the body and environment play a crucial role in cognition and cognition is not some disembodied capacity. The now often used term "4E" (embodied, embedded, enactive, and extended) cognition (Newen et al., 2018) makes clear how many different interests and perspectives are pursued. One of the Es is "extended", i.e., the idea that the boundaries of cognition might extend beyond the brain or even the brain-body system (Clark & Chalmers, 1998; Clark, 2022). Another E is "enactive", which roughly refers to the importance of action for cognition (meaning is brought about by interacting with the environment). A further E is "embedded" to emphasize the importance of the interaction of the brain-body system with the environment.

The sort of questions that contemporary 4E cognition focuses on (boundaries of cognition, perception-action loops, cultural influence, etc.) relate only in a limited way directly to the questions I am interested in here, namely the representational nature and processes of compositional conceptual thought, including language and logic. Methodologically speaking, I will focus—in a more classical and traditional manner—on the brain as a representational and computational organ, and only indirectly (namely via modal sensorimotor representations) on the aspects of the body and its interaction with the world. In other words, I take 4E seriously, but I focus on how body, action and environment are mirrored in the brain by sensorimotor-based representations and how those are processed. This brain-focused perspective is compatible with the view that higher cognition is extended. For instance, I will suggest (following Dutilh Novaes, e.g., 2012, 2020) that logical thought is grounded in dialogue (and hence in some sort of action). My interest then is to flesh out what is going on in the brain in terms of representations and the processing of such representations.

The most relevant feature of 4E cognition for my purposes is the early work from the 1980s by Lakoff and Johnson, and especially the Conceptual Metaphor Theory. Note that this is a substantial cognitive theory of conceptual thought, rather than a theory about metaphor understood as a narrow poetic-linguistic phenomenon. According to the Conceptual Metaphor Theory, our concepts are anchored in image schemas, which are a sort of modal (i.e., sensorimotor-based) representations. Abstract concepts (i.e., concepts that cannot be perceptually represented) are represented and understood as mappings with sensorimotor grounded concepts. For instance,
the abstract concept of time is often conceptualized in concrete spatial terms. Though there are also proposals on the neural organization of Conceptual Metaphor Theory (e.g., Feldman, 2008; Lakoff, 2009), those are mainly connectionist in a traditional spirit. An interesting question is whether PP and Conceptual Metaphor Theory could be combined in some way, or whether PP could serve as more detailed cognitive computational underpinning. I will discuss some aspects of those questions in Chapter 9.

In the area of formal languages and logic (as opposed to natural language), there is also interesting work under the banner of embodied cognition that could be relevant for a PP account of higher cognition. Lakoff & Nuñez (2000) have written a book-length treatment of mathematics and logic in which they flesh out what sort of conceptual metaphors mathematical concepts are. Furthermore, Dutilh Novaes' (2012, 2020) approach pictures logic as rooted in dialogical interaction, i.e., as joint action, rather than as an abstract disembodied system. Therefore, any work on higher cognition (and specifically logical thought) within the embodied cognition paradigm needs to consider her work.

Note, however, that neither in Lakoff et al.'s nor Dutilh Novaes' work do the ideas of prediction, Bayesian inference, or hierarchical models play any role for cognition. In other words, the common denominator is limited to the commitment to embodiment more broadly. "Embodiment" stresses the central role in cognition of sensorimotor representations, as opposed to amodal symbolic representations. However, it is of interest to examine how this work can be related to, or is compatible with, the PP paradigm. PP might, after all, be a cognitive-computational underpinning of Lakoff's and Dutilh Novaes' approaches to concepts and logic/formal languages. Therefore, in this thesis I will also engage with this work (see Chapters 7 and 9).

Finally, I shall highlight the very extensive work on the nature of concepts under the embodied cognition paradigm by Laurence Barsalou. Barsalou is a fierce proponent of the modal nature of concepts (conceived as sensorimotor "simulators") in opposition to LOTH-like amodal models (see e.g., Barsalou, 2016). Barsalou often comes very close to other PP commitments, without strictly providing a PP account. My proposals in this thesis are deeply influenced by his work.

b) Bayesian cognition
Bayesian approaches have become increasingly popular in cognitive science (e.g., Jones & Love, 2011; Colombo & Hartmann, 2017). Especially hierarchical models are gaining influence (e.g., Griffiths et al., 2007; Kemp & Tenenbaum, 2009; Lake et al., 2015; Tenenbaum et al., 2011). Recent theoretical and modelling work on the Bayesian brain has often relied on so-called "Hierarchical Bayesian Models". These are very powerful because they allow for representing and inferring complex structured knowledge. All of this work has had an enormous influence and is taken on board by PP (see, e.g., Clark, 2016, pp.171-176).

Note that Bayesian inference within the PP framework is only being approximated, namely through the prediction error minimization process. Bayesian approaches are often presented as computational level accounts (see Marr's (1982) three levels of description) and expressed with the help of mathematical equations, rather than neurological models (Colombo & Hartmann, 2017, p. 455; Jones & Love, 2011, p.170; Tenenbaum et al., 2011, p.1284). But the PP paradigm, as we have seen in the previous section, covers all three of Marr's levels, i.e., it includes computational, algorithmic, and implementational (neural) level commitments (Sprevak, 2021a-d).

Concepts are sometimes simply identified with variables in the models (e.g., Smith et al., 2020), or, as in work by Lake at al. (2015), visual patterns (handwritten letters) are simply called "concepts".⁹ Therefore, here, the notion of a concept is relatively thin and limited to the computational level: those accounts do not connect to, and engage with, the vast field of concept research (see Chapter 5) where concepts are the explanatory posits of a wide range of complex behavioural phenomena. In this sense, Bayesian approaches do not supply us with a more detailed and explicit account of concepts that explains, e.g., why we appear to have concepts in different formats, like exemplars, prototypes or little theories. Nor do they provide us with algorithmic and implementational level details about concept representation and processing.

Gopnik & Wellman (e.g., 2012), who have been extremely influential in developmental psychology, also take a Bayesian approach that is in the vicinity of the PP paradigm, leveraging Hierarchical Bayesian Models to explain conceptual development and world knowledge representation. Their approach can be

---

⁹ Those are examples of Bayesian models of concept learning.
interpreted as a "theory-theory" model of concepts on the computational level, i.e., they consider conceptual knowledge to be inferred by hypothesis testing. However, as I will argue in Chapter 5, the theory format of concepts is plausibly just one of many formats in which concepts manifest themselves and a more inclusive account is necessary.

In the realm of language, there are many approaches that leverage the Bayesian framework (see, e.g., Cohen, 2016; Watanabe & Chien, 2015) to explain, for instance, sentence processing. Though, again, these are not glossed as PP accounts, this work is very useful for any project to account for concepts and language under the PP paradigm.

c) The language faculty and psycholinguistic research

In psycholinguistics, empirical evidence and theoretical support for the importance of a predictive mechanism and top-down influences (neural feed-forward and feedback loops) in language production and comprehension is accumulating. This rich and growing literature should certainly be considered by PP theorists who aim at an account of language under the PP paradigm. The literature often appeals to the notion of "predictive processing". However, I want to emphasize that such research mostly stresses the predictive aspect of language processing, and not the rest of the commitments of PP. Strictly speaking, it is not work carried out under the PP paradigm, as I have characterized it here, but it is potentially compatible with it. While there is no space for an exhaustive review and discussion, in what follows I summarize some representative and relevant work.

The empirical behavioural and neurophysiological evidence covers different levels of the linguistic hierarchy: phonemes, words, sentences, and discourse both for language production and language comprehension (see Kuperberg & Jaeger, 2016; also: Chow et al., 2018, p.804). Regarding sentence understanding, the evidence points to a mechanism of word-wise sentence prediction (e.g., Kutas et al., 2011, or Kuperberg & Jaeger, 2016). At each step, the word chain produced so far, together with knowledge of grammar and contextual information constitute the priors for the
next word prediction. Usually, the processing time increases when we encounter unexpected words. This seems to be evidence for predictions at least in the form of pre-activation that facilitates further processing. Also, it has been discovered that a certain brain signal, the so-called N400 event-related potential, correlates differently with unexpected versus expected words (see Kuperberg & Jaeger, 2016, p.33; Szewcyk & Schriefers, 2018). Prediction also seems to happen at higher and lower levels. For instance, at discourse level, quick and precise turn-taking in conversations appears to require predictions about when the interlocutor will finish his or her intervention (Casillas & Frank, 2013). Baus et al. (2014) provide evidence that listeners generate predictions in their language production system regarding others' verbal actions. Lastly, there is also evidence for prediction at the lowest phonological and orthographic levels (Allopenna et al., 1998; Kuperberg & Jaeger, 2016, p.41, and Gagnepain et al., 2012, for predictive completion of spoken words).

A predictive framework best explains the speed of language processing and the ease with which we disambiguate speech in noisy contexts and adapt to the variability across speakers (see Lau, 2009, p.2; Garrod & Pickering, 2007). Strijkers et al. (2019) put forward evidence from an MEG study that supports the hypothesis that the observed speed of syntactic unification (binding) operations of different representations (e.g., the binding of a possessive pronoun and a noun into a noun-phrase structure) is driven by top-down predictive activations stemming from the prefrontal cortex. A model in which linguistic representations are being pre-activated through predictive mechanisms would be the best explanation for those cognitive feats.

Pickering & Garrod (e.g., 2004, 2007, 2009, 2013), however, might be coming much closer to what could be considered a PP account of language. They have developed a model of communication (i.e., of speech comprehension and production, as well as dialogue) that eliminates the production/comprehension dichotomy or modular model of comprehension/production. Both modes operate on the same (hierarchical) model of representations, but in different directions (top-down—production—versus bottom-up—comprehension).

[...] we define a production process as a process that maps from a “higher” to a “lower” linguistic level (e.g., syntax to phonology) and a comprehension process
as a process that maps from a “lower” to a “higher” level. (Pickering & Garrod, 2013, p.331) [my emphasis].

We assume that instances of both production and comprehension involve extensive use of prediction – determining what you yourself or your interlocutor is likely to say next. (Pickering & Garrod, 2013, p.332) [my emphasis].

[...] production is a form of action, and comprehension is a form of perception. More specifically, comprehension is a form of action perception – perception of other people performing actions. (Pickering & Garrod, 2013, p.332) [my emphasis].

Pickering & Garrod endorse the central role of prediction (and a form of analysis by synthesis, i.e., speech comprehension by a speech production approach) and, hence the close integration of action (= speech production) and perception (= speech comprehension). While they do not explicitly gloss their account as a PP account, it is still very close in spirit and highly compatible with PP.

Finally, there are some specific proposals explicitly within the PP paradigm regarding language representation and processing. For instance, Lupyan & Clark (2015) discuss the function of language through the lens of PP. They emphasize the role of linguistic labels as "artificial contexts". So, they reject the idea that linguistic structures are parallel representations that merely "latch on pre-existing concepts" for communication purposes. Rather, they endorse the notion that language is much more intertwined with concepts and knowledge and influences the unfolding of thought. Named categories, for example, are easier to learn and verbal labels make their categories more salient. Verbal cues also act as "artificial contexts" or priors, prompting expectations or predictions, for instance for the perception of noisy inputs. For example, a visual image that cannot be easily recognized is more easily perceived when the subject has been primed by a word clue (2015, pp.282–283). Rappe (2019) has also started to apply the PP framework on a conceptual level to sentence processing. She proposes that the PP framework can provide an account for "rapid and efficient processing of linguistic information and integration of contextual cues". She argues that this account is better able to accommodate evidence about the role of context in language comprehension then other available accounts. Recently Murphy, Holmes & Friston (2021) have related syntax and free energy and suggest that (Chomskian style generative) grammar is compatible with
the free energy principle. In the domain of communication more generally, Friston & Frith (2015) and Vasil et al. (2020) have proposed PP accounts of communication where agents are seen as coupled generative models who try to align their world models.

Overall, while there is some emerging and limited work within the PP framework that addresses language and communication, as I have insisted already, an account of concepts is entirely missing. Further, no work is available that addresses logical thought. What has been increasing recently, however, are critical voices. Some recent papers call into doubt the suitability of PP to account for higher cognition, like Williams (2019, 2020, 2022) and Litwin & Miłkowski (2020) (but see, e.g., Rappe, 2022, for a response, which builds on my proposed view of concepts (Chapter 3), as well as on ideas about the (dual processing) interface between language and concepts, similar to the one proposed in Chapter 8). In Chapters 11 and 12 I will discuss those objections in more detail.

2.2.2. Potential reasons for the gap

The main conclusion of the previous section is that the treatment of higher cognition, especially compositional conceptual thought, language, and logical thinking within the PP framework is still largely underdeveloped. How can this (sociological) fact be explained? I suggest that there are at least three reasons that I want to spell out and formulate as challenges from higher cognition for PP in this section. Firstly, there might be a prima facie concern: PP is not couched in terms of amodal symbols, like LOTH. The concern might mirror the well-known LOTH versus connectionism debate. PP is not suited to the more traditional higher cognition where LOTH-like approaches have been dominating. Secondly, there are some theoretical considerations that speak against PP being suitable for higher cognition (notably those put forward by Daniel Williams). As we will see in point 2) below (and discuss in more detail in Chapter 11), those theoretical concerns are numerous and directly attack some core commitments of PP with respect to the structure of the generative model. Thirdly, I suggest that the fact that there is still no account of concepts within PP—one that also connects to the vast body of traditional concept research—might be the cause of the underdevelopment of higher cognition accounts within the PP framework. While the purpose of this dissertation is not to respond one to one and
exhaustively to those challenges, but to develop a positive account of higher
cognition, it is nevertheless worthwhile to quickly walk through those three possible
reasons. They will provide a useful motivational backdrop for my enterprise.

1) The prima facie challenge from higher cognition for PP

The first challenge has to do with the notions of productivity, systematicity, and
compositionality (which I call here "PSC"\(^{11}\)). PSC is considered a central
characteristic of thought and language. Productivity captures the fact that, in
principle, we can entertain an unlimited number of thoughts or produce an unlimited
number of sentences, with a finite inventory of concepts and words respectively.
Systematicity is the property that when we can produce certain thoughts or
expressions, we can produce and comprehend systematically related thoughts or
expressions. For instance, if we can produce or understand “Peter kisses Mary” we
can do so for "Mary kisses Peter". Finally, compositionality is the phenomenon that
the meaning of a thought or expression is determined only by the meaning of the
components (concepts or words) and the way they are composed.

To account for PSC, we seem to need discrete symbols that can be flexibly
combined following certain rules. That thought and language have the PSC property
is famously articulated by the Language of Thought Hypothesis (LOTH) by Fodor
(e.g., 1975). The basic argument for LOTH is the following. As language observably
has the PSC property, and language expresses thought, thought must be language-
like. Now, it is not easy to see how we get symbols and compositionality out of the
PP model, as it is couched in completely different terms, terms that resemble much
more a distributed, connectionist paradigm. Indeed, the PP model certainly has
connectionist features (though it does not reduce to connectionism). Hence, those
concerns that have been put forward in the context of the influential
connectionism/classical computationalism debate that initiated in the 1980s and
1990s in relation to symbolic representations and the PSC property (e.g., Fodor &
Pylyshyn, 1988; Hawthorne, 1989; Smolensky, 1990; Fodor & McLaughlin, 1990;

\(^{11}\) Often the notion "compositionality" is used to refer to all three properties. I prefer to keep them
separate and use "PSC" for the combination. It is an interesting question—that I will not address
here—how the three properties are related, whether some of them are implied by others, etc.
Chalmers, 1990, 1993; Antony, 1991; Butler, 1993, 1995; McLaughlin, 1993, etc.) might carry over to PP.

What seems to crystalize out of more recent research is that the schism is not as dichotomic as pictured in the origin of the debate. For instance, Feldman (2012) argues that "both worlds are connected by the concept of modality, as formally captured by mixture models." (p.78) His point is that symbolic representations are warranted as good approximations in probabilistic worlds that are "spiky", i.e., they can be captured by a combination of multiple—narrow—probability distributions (i.e., "mixture models"). In this way we can account for symbolic cognitive processes by probabilistic means. Another interesting way to account for symbolic thought and language has already been discussed by Bechtel (1994). Deduction is not a process based on internal symbolic representations. Rather, we can carry out deductions because we have learned to use an external symbol system—natural language. This allows for a division of labour: the mind itself does not necessarily work based on the LOTH principle (mental rules operating on mental language-like symbols), but in virtue of properly manipulating an external symbols system it has the capacity to carry out symbolic processes. In other words, symbols and rules are in the world, not in the head. If the cognitive system has learned to interact with those external symbols systems, it also manifests the PSC property. While this seems a clever move (that also anticipates later ideas from embodied and extended cognition), it still does not address in detail the question of how we do learn and internally represent and process such external symbols systems.

The main point I want to make is that any cognitive model that deviates from LOTH is under pressure to provide an account for symbolic capabilities. PP is just an account that deviates from LOTH and it is not obvious how we can get the PSC property out of a PP model.

A similar concern for other non-classical cognitive paradigms has been dubbed the "scale-up problem" by Silva & Ferreira (2021):

Furthermore, a more radical approach to cognition faces the so-called “scale-up objection”, namely, the challenge of proving itself relevant for the investigation of traditional problems related to higher level cognition involving concepts such as contentful information, representational states, symbolic thought, logic, mathematics etc. (p.52).
By "more radical approaches to cognition", Silva & Ferreira refer here to 4E cognition generally, not specifically the PP paradigm. The problem they point out seems to afflict all cognitive accounts that deviate from traditional symbolic computationalism. Often the interest of 4E approaches lies not in the internal representational structure and its processing in the brain—which have been the focus of the classical approach—but in more high-level and holistic theories of mind-body-world interaction. In some cases, representationalism is even rejected, for instance in system dynamics approaches (e.g., Thelen et al., 2001). Therefore, issues around the PSC property of thought, or the fine-grained representational structure, or how exactly language is represented in the mind, as examples, are not addressed at all. In the case of system dynamics, for instance, the language in which theories are coached is differential equations. But more general, deviant cognitive approaches often use concepts, principles, and methods for which it is not clear how they could explain higher cognition of the conceptual and compositional form.

While PP is located squarely in the 4E space, its commitments emphasize neural structures corresponding to mental representations and processing in the brain. Therefore, PP is closer to the classical paradigm than more radical versions of 4E cognition, like the system dynamics approach. Nevertheless, the problem pointed out by Silva & Ferreira also seems pressing for PP. In his influential target paper from 2013, Clark pointed out that it is still unclear how to extend the PP account to higher cognition (2013, p.201):

> Questions also remain concerning the proper scope of the basic predictive processing account itself. Can that account really illuminate reason, imagination, and action-selection in all its diversity? What do the local approximations to Bayesian reasoning look like as we depart further and further from the safe shores of basic perception and motor control? What new forms of representation are then required, and how do they behave in the context of the hierarchical predictive coding regime?

Here, Clark seems to be concerned that the central posits of PP, like approximate Bayesian inference and hierarchical models, do not have an obvious connection to many higher cognitive competencies, like reasoning and decision making. The classical approach—with the explicit notions of symbols and rules—fits the bill much better for higher cognition (see also Piccinini, 2020, pp.125–126).
In sum, the prima facie challenge from higher cognition for PP can be summarized as follows:

**Prima facie challenge from higher cognition for PP:** Given that PP is couched in terms very different from amodal symbols and their formal manipulation, how can PP account for aspects of higher cognition that are paradigmatically symbolic and compositional ways of thought, including language and logic?

The main objective of this dissertation is to provide a positive account of conceptual higher cognition, which can then be seen as a response to the prima facie challenge.

2) The theoretical challenge from higher cognition for PP

The second concern relates to theoretical considerations about some core commitments of PP. If those considerations are correct, then PP might be unsuitable as a framework for higher cognition. Williams (2020) recently restated the general scale-up concern of the PP community in the following way:

As even its most enthusiastic proponents acknowledge, one of the most important challenges for predictive processing is whether the mechanisms it posits can be extended to capture and explain thought (Clark, 2016, p.299; Hohwy 2013, p.3; see also Roskies & Wood, 2017). (p.1750)

Williams has gone much further than just putting forward a prima facie concern. He has provided the most comprehensive and nuanced objections so far against PP and the role it can play in accounting for higher cognition. Here, I provide only the gist of those objections. I discuss them in more detail in Chapter 11:

a. *Insufficient expressive power of the mathematical model (namely Probabilistic Graphical Models—PGMs) to which many PP theorist are committed.* Williams' concern is that PGMs have an expressive power equivalent to propositional logic, which allows only for the representation of whole propositions. However, compositional thought requires more fine-grained representations, namely at least first order logic, i.e., the possibility to represent objects and relations.

b. *Incoherence of what PP proponents claim the model hierarchy tracks.* Many PP theorists claim that the PP model hierarchy tracks a spatiotemporal gradient. According
to Williams, this leads to trouble because then very small things, like electrons, need to be represented on lower levels and very large things, like the universe, on higher levels in the hierarchy. Given the specific architecture of the PGMs just mentioned, thoughts that contain the concepts ELECTRON and UNIVERSE would not be possible (because we cannot connect in thoughts concepts that are at a distance of more than one level in the hierarchy). But obviously we can have such thoughts.

c. Need for a modular structure. According to Williams, our mental model must be more modular than PP suggests because we have very different domain specific intuitive theories, like folk psychology or folk physics. PP posits a single unified hierarchy, not different multiple structures, so it cannot accommodate the evidence for modularity.

d. Incoherence of the free energy formulation (FE) of PP as driving the self-organization of the brain. Williams argues that the FE principle is not a causal principle; therefore, FE minimization cannot support prediction error minimization as a unifying mechanism that explains brain activity.

We can summarize the second—theoretical—challenge from higher cognition for PP as follows:

**Theoretical challenge from higher cognition for PP:** There are fundamental theoretical considerations against some core commitments that make the PP framework unsuitable to be extended to higher cognition.

The first three theoretical objections (a-c) are relevant for my purposes. They affect the core commitments of PP, namely the structure of the generative model. While it is not the aim of this dissertation to give a detailed response to, and defence against, Williams’ objections, I will nevertheless engage in some more detail with them in Chapter 11, to emphasize that these objections are worth taking seriously but are also far from being obviously decisive. As I take a neutral stand towards the free energy formulation of PP, and the last objection affects PP (in its FE formulation) more generally and not only higher cognition specifically, I do not deal further with objection d in this dissertation.

3) The instrumental challenge from higher cognition for PP

A third possible explanation for why higher cognition is not yet a thriving field of PP research, is that no PP account of concepts is available, on which more specific
Theories of higher cognitive phenomena could be built. Concepts are generally considered to be the core posits required to explain most competencies that fall under higher cognition. Concepts, as I understand them here, are mental representations of categories or classes and are hence crucial in categorization and classification tasks. Higher cognitive tasks, like analogical thinking, mathematical thinking, planning, etc. are difficult to characterize and understand without the notion of concept. Furthermore, concepts are also crucial in language. Sentences, which have a propositional form and express thoughts, consist of words and word meanings are concepts, i.e., semantics and concepts are intimately tied together. It is, therefore, unclear how an account of thought and language could get off the ground without an account of concepts.

The notion of concept relevant for the current purposes is the psychological as characterized by Machery (2009), not the philosophical, one. The psychological notion focuses on the cognitive significance, i.e., the cognitive-computational work done by concepts. The philosophical notion is concerned with reference, possession conditions and conditions for their correct application. It is debated whether those two notions relate to the same thing or whether psychologists and philosopher talk about different things when referring to "concepts" (see Machery, 2009; Löhr, 2020). I will sidestep this discussion by focusing exclusively on the psychological notion. In this context, a concept is treated as a body of knowledge that is used in higher cognitive activities like those mentioned above. What interests a psychologically focused researcher is how concepts are represented and processed (and all derived questions like whether they have an amodal or modality-specific/multimodal format, whether they are stable or flexible and context-dependent, etc.).

As already briefly highlighted, some of the PP literature mentions concepts, however, without spelling out in detail how they are structured and operate as representational devices. At most, concepts are equated with variables or nodes in the network structure (e.g., Smith et al., 2020) or representations on higher levels of the hierarchy—that are considered to be more stable (e.g., Hohwy, 2013, p.72: concepts as "longer-term expectations"). There is no work either that connects the notion of concept with the vast literature on concepts in psychology, philosophy, and neuroscience (see Chapter 5 for a review of some of the key literature).
In sum, the third—instrumental—challenge from higher cognition can be summarized as follows:

**Instrumental challenge from higher cognition for PP:** Higher cognition cannot be approached more broadly within the PP framework without an account of concepts. Ideally this account would connect with the vast body of results from traditional concept research.

Meeting this challenge should be the starting point of any account of higher cognition. This I set out to do first, in Part 2. With an appropriate account of concepts in hand, one should be able to address a wide range of higher cognitive phenomena, including language, within the PP framework.

Conclusion of Section 2.2.

The review of the available work reveals that higher cognition is not yet a thriving field of research within PP. The aim of this dissertation is to contribute to closing this gap by providing a positive sketch of the key aspects of higher cognition.

I have speculated that there might be three possible reasons for this underdevelopment, which I have formulated as three challenges from higher cognition for PP. Note that these three challenges only serve as a backdrop and motivation for this dissertation; the primary aim is not to respond in an exhaustive and rigorous way to each of the challenges. Nevertheless, the instrumental challenge is a logical starting point, and I will respond to this challenge by sketching a positive account of concepts (Part 2), extending it to language and logic, i.e., to higher cognition more generally (Part 3), and test its explanatory potential (Part 4). With this positive account of concepts and higher cognition in hand, I will then revisit Williams’ theoretical challenges (and other possible objections) in Part 5. The prima facie challenge (and specifically the concerns about the PSC property) are explicitly addressed in Chapter 6.

With this I hope to make the case that the PP paradigm can ultimately be a fruitful paradigm for higher cognition, and that it is worthwhile exploring and developing higher cognition under the PP framework.
Part 2 - A PP account of concepts

Introduction to Part 2

My aim is to use PP as a neurocognitive-computational paradigm to flesh out a sketch of how psychologically understood (see 2.2., point 3) concepts might be represented and processed in the mind in a neurally plausible way.

In the following three chapters I propose a positive account of concepts within PP. First, I tackle the very general question of whether concepts are stable or dynamic entities. Then I address the more specific question of whether concepts are amodal symbols or modal. Finally, I turn to the issue of what a more fine-grained internal structure of concepts might be.

Those three chapters together paint a picture of concepts as richly structured, context-sensitive, dynamic, sensorimotor grounded bodies of information. My aim is to connect the proposed PP account of concepts with the existing core issues in concept research: the dynamic or stable nature of concepts, whether concepts are amodal symbol systems or perceptual/embodied representations, and what representational structures do concepts have among those widely discussed in psychology and philosophy: definition-like representations, exemplars, prototypes, theory-like structures, or others.

Please note that, as already pointed out in "Note on publications", when I speak of the "format" of a concept, I refer to two things. In Chapter 4, with "format" I refer to the question of whether a concept is amodal (symbolic or language-like) or modal. In Chapter 5, with "format" I refer to the representational structure of a concept and whether it represents exemplars, prototypes, theories, etc. I will make clear in which sense I use "format" in each of the chapters.

In Chapter 3, I take a position within the invariantism/contextualism debate. This debate addresses a very high-level question about the stability of concepts across contexts. Invariantism is the view that concepts correspond to a fixed core of information activated by default each time a concept is tokened. Contextualism says that concepts activate different parts of information depending on the context. I endorse contextualism and flesh out the structure of a concept as a network of nodes where the nodes can be switched on and off depending on the context. What
is crucial here is the precision weighting apparatus that PP supplies. The knowledge of what is relevant in each context is encoded as more (second order) knowledge in the same PP model.

In Chapter 4, I then take up the debate about the format of concepts, where format is understood in a dichotomic way: as either amodal symbols or modal representations. This mirrors the divide between the paradigms of symbolic cognition and modality-specific (embodied) cognition. My idea is to undermine this dichotomy. I argue that there is no agreement on what the modal/amodal distinction amounts to. I briefly analyse different understandings of modal/amodal and show that none of them is totally clear. I then argue that the evidence can be used to support both formats, so there is no discriminatory evidence. My suggestion is that all concepts are modal and that amodality is just an extreme case of modality, namely the highest level of compression/abstraction of modal representations. A crucial element to make this proposal plausible is the hierarchical structure of the PP model, which contains an abstraction gradient.

Finally, in Chapter 5, I connect the PP account of concepts with literature concerning concepts in the second sense of “format”, internal structure. Concepts in this sense have been characterized variously as definitions, prototypes, exemplars, perceptual symbols, and theories. There is an increasing consensus that no single type of format can account for all of the empirical—specifically behavioural and neuroimaging—data associated with concept use. Psychology is, therefore, moving towards hybrid solutions. The PP model of concepts is uniquely suited to providing a hybrid that improves on other hybrid proposals on the market.

Let me stress that I do not develop a complete theory of concepts here. Any theory of concepts ideally needs to address a long list of desiderata (see, e.g., Prinz, 2002). Those desiderata often mix psychological and philosophical oriented criteria. As I have insisted previously, I am concerned here mainly with cognitive-psychological desiderata. I do not focus on the more philosophically relevant desiderata like reference and correctness and possession conditions. However, I will return to those desiderata and briefly evaluate my account with regards to them in the conclusion of Part 2.
Chapter 3. Concept contextualism through the lens of predictive processing

Abstract

Concept contextualism is the view that the information associated with a concept is dependent on the context in which it is tokened. This view is gaining support in recent years. The received and contrary view is concept invariantism, according to which a concept is instantiated by a core of information that is stable across all contexts of use. While psychologists, cognitive scientists, philosophers, and linguists have proposed concept contextualism from different perspectives, no specific cognitive-computational model that provides a mechanism for the dynamics of context-sensitive concepts is available so far. In this chapter, I make the case that an emerging cognitive paradigm, predictive processing (PP), has the resources to provide a plausible cognitive-computational model for concept contextualism and hence increase the plausibility of this view.

Keywords: predictive processing; concepts; concept contextualism; concept invariantism; precision-weighting

3.1. Introduction

Traditionally, psychologists and cognitive scientists have more or less tacitly presupposed concept invariantism, the view that concepts – once we have acquired them and attained competence in their use – are stable mental representations in the form of bodies of knowledge that refer to a class or category. I will assume throughout the remainder of this chapter that a “representation” is a mental state realized by some neural code that stands in for or refers to something in the world. The notion of mental representations is a central one in cognitive science.

---

12 I assume here that concepts are mental representations. There are alternative views on this, for example, that concepts are abilities or abstract objects (see Margolis & Laurence, 2014).

13 See, for example, Bechtel (1998, pp.297–299) for a discussion of options of how one could spell out the notion ‘standing-in.’

14 See, for example, Gladziejewski (2016), who argues that cognition in the predictive processing framework – with which I will be concerned here – is representational, that is, mental computations rely on internal representational structures. However, see, for example, Hutto and Myin (2013) for the view that representations are not involved in basic cognition, but only in the most complex linguistic
and psychology because they are taken to be entities that are important for explaining how perception gives rise to behavior through internal cognitive processes. Cognitive processes involve and operate on representations which take the form of “bundles of information available for use” (Billman, 1998, p.649). This information bundle “in the head” is sometimes called the cognitive content (as opposed to the semantic content, which is the referent of the representation). Psychological theories of concepts focus on cognitive content, that is, they describe the format of the knowledge represented by concepts as well as computational processes for concept acquisition and revision. They also explain, with the help of mental representations, the mechanisms for a range of other, higher cognitive competencies, such as categorization or classification, analogy-making, and property inference (see Piccinini & Scott, 2006; Machery, 2009, pp.8–10).

For example, the concept ‘cat’ is a mental representation that refers to the category of cats and stores specific, relatively stable information about cats that is activated when a concept is tokened. Different psychological theories of concepts posit different formats and processes that operate on those formats and implement cognitive competencies. For example, in regard to the definitional account of concepts, ‘cat’ is a definition of a cat in some propositional-symbolic form, and the classification of an object as a cat consists in checking whether it fulfils the definition. Alternatively, according to prototype theory, ‘cat’ is a list of typical features of cats with a typicality weight for each. The classification is carried out through the calculation of feature-distances between the representation of the thing to be classified and the prototype. According to exemplar theory, ‘cat’ is a set of representations of specific cats, and we categorize objects by calculating the feature-distance to an exemplar or a set of them. Another example is theory-theory, and perceptual processes. More radically, Chemero (2009) claims that our best theories of cognition should not invoke representations (“anti-representationalism”).

---

15 See Prinz (2002, pp.6–8, pp.263–282) for a discussion of the notion of cognitive content.

16 Philosophical accounts normally focus on explaining how concepts can refer to things in the world or on concept-possession conditions. See Machery (2009, pp.7–51) and Löhr (2017) for a discussion of the psychological versus the philosophical notion of ‘concept.’

17 See for example, Margolis & Laurence (1999, pp.43–51), Machery (2009, pp.100–108), or Prinz (2002, pp.75–89) for a more nuanced presentation of theory theory of concepts. Note that the term ‘theory theory’ is also used to refer to (a) an account of mind reading or folk psychology (see, e.g., Morton, 1980; Margolis & Laurence, 1999, p.43) or (b) an account of cognitive and semantic development in general (e.g., Gopnik & Meltzoff, 1997; Gopnik & Wellman, 2012).
according to which ‘cat’ is a mini-folk-theory that comprises a body of common-sense knowledge about cats, and classification is achieved via an inferential process within this theory.

Those traditional psychological accounts of concepts\textsuperscript{18} need not necessarily be invariantist, but they seem biased toward invariantism, given their relatively rigid data structures. In any case, those accounts do not supply a specific mechanism for highly dynamic, so-called context effects in the use of concepts. The information retrieved or made available when a concept is tokened seems highly flexible and depends on the context and situation in which the cognitive agent carries out the cognitive task that involves concepts. Invariantists might be able to deal with such context effects, but not within their theories of concepts. Rather, context effects are typically relegated to a theory of “background knowledge” (e.g., Machery, 2009) and do not form part of the core of a theory of concepts. However, invariantists have not developed such a supplementary theory.

Because more and more evidence for context effects is accruing, the view that we should not characterize concepts in the invariantist manner – as stable bodies of knowledge – is gaining momentum. An increasing number of philosophers, psychologist, linguists, and cognitive scientists (e.g., Barsalou, 2009, 2011; Casasanto & Lupyan, 2015; Ludlow, 2014; Rice, 2016) view concepts as highly flexible, dynamic, and context-dependent representations. Here, we should understand ‘context’ in a broad sense: as a set of parameters that describe the environment in which a cognitive agent is embedded, and which influence the exercise of cognitive competencies. Those parameters might include the neuro-chemical history of the cognitive agent, bodily conditions (like the position of the body or arms), physical environmental conditions (like temperature or light conditions), social context (e.g., the agent in the role of a participant in an experiment), cultural context, language spoken, recent memories, priming

\textsuperscript{18} Here I have just briefly mentioned theories of concepts that have been relatively influential in recent decades, especially in psychology. Other recent proposals include Gärdenfors’ theory of “conceptual spaces” (2014), which has elements in common with prototype theory, or Prinz’s “proxytype theory” (2002) – which I will mention again later in connection with Barsalou’s account. Fodor proposed that concepts are structureless symbols with his doctrine of “conceptual atomism,” but some authors think that this is not a psychologically plausible account. For a discussion, see Prinz (2002) or Schneider (2011). Schneider has reformulated Fodor’s “conceptual atomism” into “pragmatic conceptualism,” and she emphasizes the functional role of concepts, thereby providing an account that is psychologically more plausible.
experiences, or the purpose of the cognitive task (see also Casasanto & Lupyan, 2015).

Although I have thus far characterized the invariantism–contextualism distinction as a dichotomy, it should be better understood as being one of degree. On the one hand, invariantists do not deny that concepts can change, at least in the long run. On the other hand, contextualists would not deny that many concepts might have stable informational cores across a wide range of contexts. My aim, however, is not to come up with a robust delimitation criterion. Rather, I want to focus on contextualism as the view that conceptual representations might change in a very short time frame (e.g., within a conversation), and the relevant knowledge retrieved might – at least in principle – change each time a concept is used.

Now, while the contextualist view is increasingly being endorsed, it is unclear what the underlying cognitive-computational model for such highly flexible dynamics of context-dependent concepts could look like. By ‘cognitive-computational model’ I understand here a description of the format of conceptual representations and a mechanism that operates on those representations to generate context-effects. The model should be sufficiently specific and detailed, and plausible as a neural mechanism and implementable as a computer simulation. It would be a strength of a cognitive-psychological view like concept contextualism to have the support of such a model.

I suggest that the framework from an emerging cognitive paradigm, predictive processing (PP), can provide such a model. According to PP, our brain maintains a probabilistic generative model that continually predicts the brain’s sensory input. The system compares the prediction with the perceptual input and tunes the model or generates bodily action to minimize prediction error. I will argue that concepts in such a model have the role of prediction units: they are the mental representations in terms of which predictions are made. For instance, if we predict a cat on a mat, the model generates a prediction that involves the concepts ‘cat,’ ‘mat,’ and ‘being-on.’ I will argue that PP – and especially the version fleshed out by Clark (2016) – can

---

19 The “format” of a conceptual representations is its structural description – that is, the parts which it consists of and how they are related. Of course, a concept might be non-structured or atomic, as Fodor holds.
supply a mechanism for dynamic ad hoc modulations of concept meanings and information retrieval.

My plan is as follows: in Section 3.2, I describe concept invariantism and contextualism in more detail and provide some examples. In Section 3.3, I explain the PP framework, and in Section 3.4, I provide an account of concepts in PP. In Section 3.5, I sketch a mechanism for the context sensitivity of concepts and illustrate it with some examples. In Section 3.6, I discuss some objections, and in Section 3.7, I conclude.

3.2. Concept contextualism and concept invariantism

Machery (2009, 2015) has recently defended invariantism and shall serve as an example to explain the view in more detail. He characterizes the psychological notion of ‘concept,’ following what he takes to be a commonly accepted and plausible characterization of ‘concept’ in psychology:

A concept of x is a body of knowledge about x that is stored in long-term memory and that is used by default in the processes underlying most, if not all, higher cognitive competences when these processes result in judgments about x. (2009, p.12)

Central to this characterization is the notion of ‘used by default.’ Machery (2015, p.570) clarifies this notion as quick, automatic, and context-independent retrieval of information from long-term memory. He therefore takes the general view about concepts to be concept invariantism, the view that each concept is instantiated by its stable core of information that is retrieved independently from the context each time it is used. Indeed, the traditional psychological theories of concepts are compatible with Machery’s characterization of the mainstream view on concepts. For example, each time the concept denoted by ‘cat’ is instantiated, the body of knowledge retrieved could be a list of dominant features of a prototype of cats, such as [has whiskers; is furry; meows].

Note that Machery himself defends invariantism by discussing a range of neuroscientific and behavioural experiments and concludes that “thus, contextualism … seems to mischaracterize the nature of knowledge retrieval from long-term memory” (2015, p.585). However, Machery is a “concept eliminativist” (2009). He observes
that there is ample empirical evidence for the existence of prototypes, exemplars, and theories. He then argues that those fundamental types of concepts have very little in common in terms of their format and how they are processed, and, hence, we should not expect to find rich generalizations regarding an inclusive notion of concept. Machery recommends, therefore, that psychologists eliminate the notion of concept from their scientific vocabulary. This implies that we do not have a single concept CAT but, rather, at least three different ones (see also Machery, 2015, p. 568): CAT\text{exemplar}, CAT\text{theory} and CAT\text{prototype}. For the purpose of this chapter, however, it does not matter whether Machery is right in “splitting” concepts. We only need to be aware that what Machery has in mind when talking about invariantism is that for each tokening of one of those fundamental types of concepts, a stable core of knowledge is retrieved.

Concept contextualism is the contrary view. Concepts are dynamically adapted to contextual parameters or created on the fly by a situated cognitive agent. Contextualism is motivated by the observation that the information retrieved by a concept depends on the specific situation in which it is used. As already pointed out, concept contextualism is becoming increasingly popular across a wide range of disciplines. Let me briefly illustrate the view with four examples. For instance, Ludlow (2014, 2017) proposes – from a linguistic and philosophy of language perspective – a “theory of dynamic lexicon,” where he develops the idea that word meanings are built on the fly in the form of “micro-languages” – that is, languages built for the occasion. According to Ludlow, all meanings are modulated and might change, even within a conversation. He explicitly asserts that meanings have no stable core:

Some theorists think that there is a core meaning for a term that is the absolute sense of the term ... I will argue that the “absolute” sense of a term (if it even exists) is not privileged but is simply one modulation among many – there is no core or privileged modulation. (2014, p.6)

Ludlow is not concerned with cognitive-computational models but, rather, with questions about the formal semantics of dynamic meanings; he has not put forward

\begin{footnote}{There are other accounts with an arguably contextualist spirit, like that of Gärdenfors (e.g., 2014, 2018) or Evans (2009).}

\end{footnote}
a cognitive-computational model. Another example comes from Casasanto and Lupyan (2015), who have recently developed an “ad hoc cognition” framework from a psychological and cognitive-sciences angle. According to those authors, there is no stable body of information that is retrieved by default when a concept is tokened; rather, there is a dynamic process of context-dependent conceptualizing. For them, all concepts are ad hoc and created on the fly, and they might change, even in micro-second time scales:

We will use the term ‘concept’ to mean a dynamic pattern of information that is made active in memory transiently, as needed, in response to internally generated or external cues. ... Rather than a process of accessing a preformed package of knowledge, instantiating a concept is always a process of activating an ad hoc network of stored information in response to cues in context. (2015, p. 546)

As in the case of Ludlow, Casasanto and Lupyan do not provide any cognitive-computational model for concept-contextualism. Barsalou (e.g., 2009, 2011) should serve as the last example. His theory of “situated conceptualizations” pictures concepts as dynamic representations (“simulators”) with highly context-dependent and multimodal content:

For example, one situated conceptualization for ‘bicycle’ might support riding a bicycle, whereas others might support locking a bicycle, repairing a bicycle, and so forth. On this view, the concept for ‘bicycle’ is not a single generic representation of the category. Instead, the concept is the skill or ability to produce a wide variety of situated conceptualizations that support goal achievement in specific contexts. (2009, p. 1283)

According to Barsalou, concepts are implemented as a multimodal network structure in a neural net. The selective activation of sub-nets depends on the situation in which the cognitive agents act. Those activations correspond to simulations or “modal re-enactments” of perceptual, motor, and introspective states (see Barsalou, 2009). Barsalou takes “situated conceptualization and pattern completion inference on situated conceptualizations” as the fundamental cognitive mechanism (2009,

21 Another contextualist account of concepts, which is very close to Barsalou’s account, is Prinz’s “proxytype theory”. Here, concepts are mental representations of categories – in the form of multimodal knowledge networks – in long-term memory that are or can be activated context dependently in working memory (e.g., Prinz, 2002, p.149).
p.1284); however, he does not supply a detailed cognitive-computational model. In his 2011 article, he moves one step further and proposes that the simulations are the result of Bayesian inference, but he does not flesh this out further.

Those contextualist positions are motivated and supported by a range of behavioural and cognitive-neuroscientific findings that indicate that the information that is processed when a concept is tokened varies with the situational parameters (e.g., Hoenig et al., 2008; Lebois et al., 2015). However, as I mentioned in the introduction, invariantists could also account for context effects by postulating “background knowledge” that is context-sensitively retrieved in addition to the stable core of knowledge (Machery, 2009, 2015). Therefore, the bulk of the evidence seems compatible with both contextualism and invariantism. There have also been some theoretical arguments put forward against invariantism. For instance, Löhr (2017) has claimed that the invariantist cannot account for abstract concepts or for concept composition. On the other hand, as already mentioned, Machery (2015) has defended invariantism on empirical grounds. However, again, it has been suggested that this evidence is compatible with both positions (e.g., Löhr, 2017). It is not the purpose of this chapter to evaluate those arguments. Instead, I want to draw attention to the fact that, so far, we have no specific cognitive-computational model for concept contextualism. One desideratum of a cognitive-psychological theory of concepts is that a model exists that can underpin it. Therefore, in the next section I develop a proposal with the aim of strengthening the case for concept contextualism.

3.3. The PP framework

You arrive at home and see a cat on the mat. What, according to a common view of cognition, 22 is going on in your brain? The story runs as follows: light patterns are hitting the retina of your eyes. Those patterns (e.g., encoded in a two-dimensional pixel-field) are processed bottom-up, composing features, like oriented edges, which are then arranged as parts of a representation of a three-dimensional object. The object is finally classified as a cat based on the recognition of the features of a cat (such as the right size and form and the possession of whiskers).

22 For example, see Marr (1982) and the discussion by Tye (1991, pp. 77–83).
Predictive processing (PP) (e.g., Clark, 2013, 2016; Hohwy, 2013; Friston, 2010) provides a different story of how the perception of the cat works. PP has already been successfully applied to a wide range of psychological phenomena, like certain effects in perception (e.g., hollow-mask and rubber-arm illusions), mental disorders (e.g., autism, schizophrenia), and attention and consciousness, among others. What characterizes PP in contrast to traditional cognitive approaches is the emphasis on the importance of top-down predictions based on expectations, in addition to bottom-up processing. One’s perception of the cat on the mat is not exclusively driven by bottom-up feature-aggregation from sensory input, but also through top-down influences from prior beliefs. The main tenet of PP is that the brain is a prediction machine that continually anticipates its own sensory input, and the predictions are the result of a complex interplay of bottom-up information-flow and top-down influences. The predictions are inferences carried out mostly unconsciously in a causal model of the world. The prediction model has the following main features: (a) it is probabilistic, (b) it has a hierarchical structure, and (c) it is driven by a constant minimization of prediction error. Let me explain those features in turn.

Regarding feature A, proponents of PP usually take the prediction model to have the form of a generative probabilistic model in which (at least approximately) Bayesian inference is carried out23 (see, e.g., Clark, 2013, pp.188–189; Hohwy, 2013, pp.15–39; for Bayesian causal models, see Glymour, 2001). Such a model can be represented as a network of nodes over which a probability distribution is defined. The nodes represent random variables that stand in for, for example, events, states, or properties. The connections between the nodes represent probabilistic causal influences. Cognition, including perception and the control of action, is grounded in inference in such a Bayesian net. For instance, the stable perception of a cat on a mat occurs because the “hypothesis” of a cat being on the mat has achieved the highest so-called “posterior probability.” The posterior probability is the product of two other probabilities: the probability of the event that a cat on the mat causes the sensory input which the retina actually receives (“likelihood of the hypothesis”) and the prior probability that a cat is on the mat. The hypothesis that a cat is on the mat might have been the winner in a competition between various hypotheses like “a cat

---

23 However, see Aitchison & Lengyel (2017), who suggest that predictive coding and Bayesian inference need not necessarily go together.
is on the mat,” “a dog is on the mat,” and “a bag is on the mat.” In this way, the predictive model infers hidden causes from observed effects (visual, auditory, or other sensory stimuli). The current probabilistic knowledge represented in the Bayesian network (the “priors”) is updated as new evidence accumulates. An initial version of such a model can be built from relatively sparse input (see also Kemp & Tenenbaum, 2009, p.21) and then increasingly refined by adding nodes, changing connections, and adjusting the probability distributions.24

Regarding feature B, in the predictive processing framework, the causal model has a hierarchical structure and represents prior knowledge on many levels of abstraction (e.g., Clark, 2013, p.25; Lupyan & Clark, 2015). For instance, the lowest level might predict edges with different orientations, and a higher level might represent shapes formed from those edges. At even higher levels, we might have representations of cat parts, cats, animals, and so on. In the top-down prediction cascade, the predictions of higher-level layers serve as priors for the lower-level predictions and, in this way, constrain the hypothesis space on the lower level. For example, on a higher level in the hierarchical model, the implicit knowledge about the size ranges of various objects provides a constraint for the sorts of things that can be on mats.

Regarding feature C, key mechanism in PP, which will be central to the account of concept contextualism that follows, is prediction-error minimization with an associated precision-weighting of prediction errors. The predictions are constantly compared to the actual sensory input in so-called “error units,” and the “residual error” of the predictions is calculated. This error signal then flows laterally and upwards in the hierarchical Bayesian net and may lead to updates of the model at different levels of the hierarchy. The model is constantly tuned and is supposed to converge toward a version that minimizes the overall prediction-error in the long run. There are two fundamental and interrelated ways to minimize prediction-error: firstly, as already mentioned, by updating the internal Bayesian model to fit predictions to the sensory flow. However, the error can also be minimized by issuing actions to fit the sensory flow to predictions (“active inference”), that is, the brain with its body can change the world to fit its prediction.

24 Those processes are carried out mostly unconsciously. The description with terms like 'hypothesis,' 'evidence,' or 'inference' should not be considered an "over-intellectualization." See also Hohwy (2013, p.23) for a similar point.
The prediction-error minimizing process is supplemented by a mechanism of precision-weighting of prediction errors (Clark, 2016, pp.53–83). The brain needs to discriminate noise from useful signals because noise should not force an update of the model. The brain must therefore predict the reliability of the sensory input, assign weights to the error signals, and thereby determine the influence of top-down predictions versus that of updates driven by bottom-up error-sIGNALS of the model. An example from Clark (2013, p.198) might serve to illustrate this point. In a situation with thick fog, visual sensory information about the shape of an object is less reliable than, for instance, tactile or auditory information. In such a context, the precision-weighting mechanism predicts that the bottom-up visual signal has low precision. Bottom-up error signals related to the shape are then tuned down to avoid an update of the brain’s prediction model, and the influence of other sensory modalities or top-down predictions increase.

Empirical evidence supports the existence and relevance of the precision-weighting mechanism in the brain. Kanai, Komura, Shipp, and Friston (2015), for instance, provide a brain-anatomical model for PP as a prediction or inference organ consisting of a hierarchical cortical structure with two classes of neural connections: (a) first-order connections that encode content represented in the neural network; and (b) second-order connections representing context in the form of salience, precision, or confidence in regard to first-order content. It is those second-order connections that implement the mechanism of precision-weighting through which they modulate the gains of neural populations that pass prediction-error information upward in the neural hierarchy. Kanai et al. (2015) also provide empirical evidence for precision-weighting in the pulvinar (a part of the thalamus that is supposed to be involved in, among other things, visual attention). This evidence supports the thesis that a significant number of the neurons of the pulvinar encode expected precision or confidence, and the pulvinar has different specific neural mechanisms to modulate the gain of the error-sIGNAL flow.

To wrap up the description of the PP framework, let me emphasize that there are specific algorithms and a mathematical apparatus that describe the PP model dynamics (e.g., Spratling, 2017; Clark, 2013). However, it is not the purpose of this chapter to propose a specific and detailed implementation-level algorithm of the workings of dynamic, context-sensitive concepts. Instead, I want to focus on making
the case that the PP framework, with the features just described, could provide a plausible cognitive-computational model for concept contextualism, and it is the just described error-minimization and precision-weighting apparatus that provides the necessary mechanism.

3.4. Concepts in a PP model

So far, the PP literature has not addressed in detail the nature of concepts. Therefore, in this section, let me propose what role concepts could play in a PP model. Then, I will flesh out what kind of representations concepts could be. Finally, I will describe a possible mechanism of context-sensitive concept-modulation.

3.4.1. The role of concepts in PP

As we have seen, PP is a complex mechanism of continual predictions and error-minimizing adjustments in a distributed, hierarchically organized neural network representing a generative model. The neural network is composed of “error units” and “representational units” that are connected via top-down and lateral prediction signals, and bottom-up and lateral error signals that have to be minimized (see Kanai, Komura, Shipp, & Friston, 2015; Seth & Friston, 2016). Error units receive a modulatory signal that weights the signals they generate. This modulatory signal drives the precision-weighting mechanism described in Section 3.3 (c).

How might concepts fit into the PP model, and what role might they play? If cognition consists in the generation of predictions, then, plausibly, concepts are those representations in terms of which our brain makes predictions. In other words, I suggest that concepts are prediction units, and thus, they must be computational structures with which the error units are associated. I speculate, therefore, that we can associate concepts with pairs of error and representational units that form “conceptual units” (see Figure 3.1), which are heavily interconnected in a hierarchical network.

25 I assume, as pointed out in the introduction, that concepts are representations.
Given the high metabolic cost of brain activity, it seems plausible to assume that the brain evolved in such a way that predictions are carried out in a way that secures a good enough trade-off between the processing costs and the accuracy and relevance of the predictive content. Concepts, as I have suggested, being those representations that constitute a prediction outcome, would then be instrumental to secure a good enough trade-off. How can concepts achieve that?

I suggest that concepts as prediction units have a crucial role to play in modulating the prediction detail. It would not be efficient to always predict a situation with the maximum level of detail. For example, imagine the situation where you have to step over a cat. In order to infer a motor program, it is not necessary to predict and represent all of the attributes of a specific cat or the exact pixel pattern of the cat as it impinges on the retina. The rough shape and size, as well as some information about cat behaviour in general, would probably suffice. More generally, to engage with the world, it would not be efficient – or even possible – to always operate with representations with a maximum level of detail. Rather, the brain needs to extract regularities and patterns from the environment and to abstract away unnecessary details. In this sense, concepts – I suggest – are crucial devices through which data compression and context-sensitive modulation of the prediction detail is achieved.
This modulation shall depend on the relevance of the features to be taken into account for the cognitive task at hand.

3.4.2. Characterizing conceptual representations in PP

In the previous subsection, I proposed that concepts are representations that serve as prediction units and secure efficient predictions by modulating the prediction detail. I turn now to the question of what sort of representations concepts and the resulting network of concepts need to be in order to fulfil this role. How could we characterize concepts individually as representations along various dimensions, such as conscious accessibility, modality specificity, or degree of stability (which will be explained in a moment), and how do concepts hang together to form larger networks of knowledge? Once those questions are clarified, the next challenge will be to provide a mechanism for the context-sensitive modulation of prediction detail.

3.4.2.1. Concepts encompassing multimodal, non-lexicalized, non-consciously accessible representations

Many theories of concepts assume that conceptual representations are bodies of knowledge that can be expressed in sentential or mathematical terms containing amodal symbols. This characterization is compatible with the idea that concepts might be non-consciously accessible or non-lexicalized representations (e.g., the knowledge of the grammar of a language could be represented in this way). However, there is also evidence for modal representations that cannot be expressed like language (e.g., representations of proprioceptive or motor states) and that exercise the functions of prediction units. Firstly, there is mounting empirical evidence for conceptual representations involving sensorimotor areas in the cortex in neural assemblies forming complex multimodal functional webs (see Hoenig et al., 2008; Pulvermüller, 2011; Kiefer & Pulvermüller, 2012). Secondly, in the PP story, representations of, for example, proprioceptive or motor states are also predictive outcomes, namely those serving as programs to bring about action. Therefore, I

26 “Amodal” symbols are not linked to any perceptual mode and are purely language-like.
propose taking a more liberal view of concepts in PP that also includes multimodal representations\textsuperscript{27} in addition to non-consciously accessible and non-lexicalized ones.

3.4.2.2. Concepts including ad-hoc representations

PP theorists have emphasized the value of flexible ad-hoc representations for efficiency reasons. For example, consider Clark’s “outfielder’s problem” (2016, p. 247). In baseball, to catch the ball, the outfielder uses a strategy called “optical acceleration cancelation.” He runs in such a way that the ball moves at a constant speed through the visual field. In this way, he will end up catching the ball. This problem-solving strategy seems to rely on thin ad-hoc representations that are suitable for efficient predictions for the purpose at hand. Richly detailed three-dimensional conceptual representations of the ball, the trajectory, and the environment are replaced by much thinner representations created on the fly for the purpose. There are two ways to account for such frugal representations. One way is to claim that entirely new concepts (prediction units) are created on the fly (let’s say \textbf{VISUAL POINT} and \textbf{CONSTANT SPEED}). However, one might object on the basis that this implies too broad of a notion of concept. Representations, apart from their component role in predictions, must fulfil other conditions to count as concepts, such as conscious accessibility, stability, informational richness, level of abstraction, and cognitive promiscuity (i.e., general applicability across a wide range of domains). However, a positive argument for an inclusive notion of concepts is that it allows for a view of concept development that seems plausible, especially within the PP framework, and it also avoids complications to establish principled cut-off points for each criterion. We could consider concepts to be on a continuum from thin and fleeting to rich and stable representations. 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.”\textsuperscript{28} A second way to account for frugal representations is through a modulation mechanism applied to

\textsuperscript{27} See Wajnerman (2018) for an overview of the debate between modal and amodal theorists of cognition. Thagard and Findlay (2012) also defend multimodal representations and provide an account for, for example, abductive reasoning with multimodal concepts.

\textsuperscript{28} For example, Dehaene (2009) argues that letter recognition and reading piggy-backs on evolutionarily developed visual representations and mechanisms for recognizing objects in the environment, which one might think to be quite specialized and even "encapsulated."
already existing concepts. For instance, VISUAL POINT might not be a new concept, but a version of an existing concept (e.g., OBJECT or BALL) with dynamically suppressed features (e.g., “round shape” or “three-dimensional extension”). For the purpose of this chapter, the option we choose here is inconsequential. Both options can be accommodated by the account of context-sensitive concept modulation that has been proposed here.

3.4.2.3. Intermeshing of domains

The degree of modularity of the brain, that is, the extent to which its architecture is organized into relatively independent functional units, is the subject of much debate (see, e.g., Robbins, 2017, for an overview). However, inter-modal neural assemblies (see above in Section 3.4.2.1) as well as re-wiring experiments (Newton & Sur, 2005) and “neural re-usage” or “recycling” phenomena (e.g., Anderson, 2010; Dehaene, 2009) suggest a high degree of flexibility and interconnectivity. Inter-domain-connectivity is also plausible because some higher cognitive competencies, like analogy-making, rely on processes that cut across domains. The PP picture can naturally accommodate the idea of significant flexibility and multimodal and interdomain-connectivity. The idea of multimodal conceptual representations is getting increasingly popular. For instance, Gardenförs (2014) characterizes concepts as convex areas in spaces spanned by basic “quality dimensions.” Different modalities can be combined and form “product spaces.” Thagard and Findlay (2012) also endorse multimodal concepts: neuronally realized representations of different modalities can be combined through so-called “convolution” operations. The thesis that the brain maintains a hierarchical generative model is compatible with a certain degree of functional and structural modularity. Functional and structural cognitive units and their type and degree of independence might arise from innate constraints, which we could take to be hard-wired priors. Alternatively, a modular structure might simply crystalize as the overall model because it minimizes prediction-error in the long run (see also Drayson, 2017).

3.4.2.4. Coherence and consistency

Some theorists of concepts – for instance, those in the camp of theory theory – insist that conceptual systems are organized as coherent structures, much like propositionally expressed scientific theories (e.g., Gopnik & Meltzoff, 1997, pp.32–
Wellman, 2014, pp.119–120). However, given the complexity of the brain’s PP model, its multimodal representations and its continual adjustments and tuning, it is unlikely that the representations form a coherent and consistent body of knowledge that we can fully formalize in a propositional or language-like format. Some authors have even suggested that our mind is necessarily inconsistent and that we are forced to believe infinitely many contradictions (Sorensen, 2004). If this is right, then the demand for consistency could not even be fulfilled in principle. However, inconsistencies (or irrational or biased thinking or behaviour) can also be of practical, evolutionary advantage in an uncertain world, and they might, therefore, be a desirable feature in the brain’s model (see, e.g., Bortolotti & Sullivan-Bissett, 2017).

In any case, it is unclear whether we could evaluate the consistency of a model consisting of a mixture of lexicalized and non-lexicalized, consciously and non-consciously accessible, and modal and amodal concepts. Consistency and coherence are requirements of certain formal systems, such as those modelled on mathematical axiomatic systems or first-order logic. We should view formal systems like those as cultural artifacts that contribute to shaping the mind rather than constitute it (see Dutilh Novaes, 2012, p.161).

In sum, I suggest that concepts are representations that include multimodal, non-consciously accessible, and dynamic ad hoc representations. Those are realized as functional neural assemblies with inter-modal and cross-domain connections. The body of knowledge represented in a PP model is a web of interconnected concepts, and it might not be fully characterizable in terms of coherence and consistency. The purpose of the PP model implemented in the brain is to make efficient predictions. The PP model, therefore, might also contain pragmatic “miss-representations,” which take the form of simplifications and abstractions that are adequate given the limited human cognitive capacities and need for efficiency. Only what passes a cognitive cost-benefit test should be represented in the model. The flexibility of the representations in the PP model leads to the notion of a concept as a pragmatic

29 As an anonymous reviewer correctly pointed out, the admission of inconsistencies allows for systems with increased expressive power. See, for example, Priest, Tanaka, and Weber (2018).

30 Maybe we can find a way to approximately “formalize” the workings of the mind. However, it is not clear whether the mind implements language-like formalism, nor is it clear whether notions like ‘consistency’ and ‘coherence’ would be applicable. My speculative view is that the functioning of the brain can probably be formalized on a sufficiently low level of description (e.g., in terms of neural spike patterns), but (language-like) formalization at higher levels are idealizations or approximations.
prediction-unit. Context-sensitive modulation is a way to make predictions more efficient by regulating the level of detail of the predictions. The PP model can, as I will argue in the next section, provide a mechanism for context-sensitivity.

3.5. A mechanism for context-sensitivity

3.5.1. The role of prediction granularity in prediction-error minimization

I have already mentioned two fundamental ways to reduce the prediction error in a PP model: adjusting the model to fit the world better and adapting the world to fit the model better. To explain how concepts are context-sensitively modulated, we now need to have a closer look at more specific ways to reduce prediction errors in a PP model. Kwisthout, Bekkering, and van Rooij (2017) have recently discussed in detail the six specific methods that are available in principle. The last two are relevant for our purposes. Here is a summary:

- Revision of the probability distribution over the hypotheses, done by calculating the posterior probability distribution over the hypothesis space and using it to assign new priors. As a result, a different hypothesis can be selected that generates a lower prediction error.
- Revision of the causal model, done by updating the stochastic dependencies between the nodes and/or the introduction of new nodes.
- Gathering of evidence. The prediction error is minimized by seeking observations of unobserved intermediate variables in the model and in this way forcing a value upon those variables.
- Interventions in the world, called “active inference,” done by carrying out some actions in the world and in this way setting the values of some variables in the model. The prediction becomes a “self-fulfilling prophecy.”
- Modulation of the network toward a lower level of detail of the predictions. This can be achieved, for instance, by using a representation in which fewer nodes (i.e., fewer features) are activated.
- Modulation of the network toward a higher level of detail of the predictions. This can be achieved, for instance, by using a representation in which more nodes (i.e., more features) are activated.
Kwisthout et al. place a particular emphasis on the importance of methods 5 and 6. The reason is that the prediction-grain adjustment is relevant for error minimization in the case of categorical (discrete) probability distributions (2017, pp.84–85).

Concepts, as I have suggested, are networks of (discrete) prediction units; hence, the relevant probability distributions are plausibly categorical. I suggest that those two ways of error minimization correspond to context-sensitive modulations of concepts. Observe that methods 1 to 4 do require structural changes in the model or a special effort in the form of action. Methods 5 and 6, however, are merely transient changes of the activation patterns. It seems plausible that it is an efficient cognitive strategy to first try to adjust predictions in a way that does not require structural changes nor energy-consuming actions. The prediction granularity can be increased by limiting the number of features or the amount of knowledge activated with the tokening of a concept, and it can be decreased by increasing the number of features or the amount of knowledge.

3.5.2. Granularity modulation with the error-weighting mechanism

I suggest that the mechanism which modulates the level of detail by switching on and off features could operate through the same mechanism that is used for prediction-error weighting based on precision estimations of sensory signals. The only thing we need in addition is a representation of the “knowledge” associated with a concept; then, we can predict the relevance of its features depending on the context. We could also characterize this knowledge about relevance – as is the case with precision – as second-order knowledge represented as priors at higher levels of the hierarchy of the model. It contributes to determining the weights of the error signals that are related to concepts at the level of first-order knowledge. This knowledge is just more knowledge in the same PP model that drives error-signal weighting; it encodes information about what features should be salient and computationally effective depending on the context. The error signal of features that are judged to be irrelevant for a specific context should be reduced and, if it reaches a certain threshold, not processed further, which amounts to switching off those features.

31 I use ‘knowledge’ in a loose way here, roughly as a synonym of ‘information.’
At this point, one might be concerned that I am stretching the precision-weighting mechanism too far. It is one thing to predict the accuracy of sensory information but a different thing to determine the right level of detail for a prediction.\textsuperscript{32} Therefore, goes the complaint, we need to posit a separate mechanism—hence multiplying mechanisms. However, the mechanism is only one, namely, the tuning of prediction-error signals in the error units. The knowledge about which level of grain might be suitable for predictions in different contexts – knowledge which we plausibly have and apply when we use concepts – is just part of all of the encoded knowledge that feeds into the mechanism that tunes error signals. There are, thus, two complementary sorts of encoded knowledge that drive error-weighting: precision and prediction grain. Adding this second driver to the error-weighting mechanism is not ad hoc because Kwisthout et al. (2017) have shown that precision and prediction grain are two faces of the same error-minimization coin. Let me now explain how the mechanism which modulates the granularity of concepts would work, using two examples of context effects from the literature.

3.5.3. Examples of context-dependent concept-modulation

There is increasing evidence for the existence and pervasiveness of context-sensitive modulations of multimodal and distributed conceptual representations (for overviews, see, e.g., Yee & Thompson-Schill, 2006; Yang, 2013). I will focus on two examples of context effects, one classical and another more recent, and explain for each the mechanism of context-sensitive feature-modulation. Both examples show that we can explain feature selection in conceptual representations by using the same error-signal weighting mechanisms as we used for precision-weighting.

3.5.3.1. Context sensitivity of semantic recall

A classical experiment that shows context effects for conceptual representations of object words was carried out by Barclays, Bransford, Franks, McCarrel, and Nitsch (1974). In the experiment, participants were presented with a set of sentence-stimuli in the following form:

The man lifted the piano.

\textsuperscript{32} I thank an anonymous reviewer for this objection.
The man tuned the piano.

The participants were then provided with a cue word to recall the target noun (‘piano’). There were two sorts of cues, both describing the properties of a piano. However, one cue word was closely related to the property highlighted by the sentence (i.e., “heavy” in the case of sentence 1 and “with a nice sound” in the case of sentence 2). The result was that related cues correlated with better recall of the target nouns. A context effect can explain this: when reading sentence 1, the information retrieved for ‘piano’ is influenced by the situation described by the sentence in which ‘piano’ appears, namely a situation where the weight is salient. Hence, the weight feature for the concept ‘piano’ is primed. The PP account of concepts can explain this in the following way. The sentence “the man lifted the piano” is processed, and conceptual representations in the PP model are generated. Among other concepts, ‘piano’ is represented. It is represented with features relevant to the situation that the sentences evoke. Given that the situation is one in which the piano is moved as a heavy object, the weight feature is relevant and salient. Expressed in terms of the PP framework, this amounts to the following: the physical features, such as the weight, are given special cognitive “attention” (by tuning the bottom-up direction of information-flow in the hierarchical model). This is implemented in the PP model by augmenting the error sensitivity toward, for instance, the feature “heavy.” At the same time, error weights are reduced for other features associated with ‘piano,’ for example, that it “has a nice sound.” The tokening of the concept ‘piano’ with the activated feature “heavy” now works as a prime for the following recall task: when asked to recall the object related to the cue “heavy,” that is, to match the property “heavy” with the objects remembered from the phase where the sentences were presented, ‘piano’ is recalled more easily. As the memory of the sentence fades away, and ‘piano’ is used in other contexts, the error weights will be changed to activate features relevant to the new situation.
3.5.3.2. Context sensitivity of modality-specific features in conceptual representations

Van Dam, van Dongen, Bekkering, and Rueschemeyer (2012) and van Dam, van Dijk, Bekkering, and Rueschemeyer (2012) provide evidence that when object concepts are being processed, the modality-specific regions activated in the brain vary with the context. The authors present concept words to participants for which both visual and motor properties are relevant (e.g., ‘tennis ball,’ ‘boxing glove’). Before the presentation of the word-stimuli, the participants were focusing on either action or colour. This was achieved by asking them to decide for each word whether the denoted object had a certain property (i.e., “is green” in the colour condition and “is an action done by feet” in the action condition). It turned out that for a specific word, the brains of participants in the action condition showed stronger connectivity between the semantic-linguistic and motor areas than did the brains of participants in the colour condition.
In the PP framework, this can be explained again as context-dependent feature selection via the error weighting mechanism. By inducing the participants to focus on a context of action, the error-weighting mechanism tuned up the error signals of action features (represented in the motor cortex), while the weights at error units linked to representations of colour features are tuned down. This reduces the influence of error signals from colour features, effectively switching them off and increases the influence of motor features. If the error signals of colour features are blocked, then one can expect reduced connectivity with the areas where those features are represented, while increased motor connectivity should be expected as the related error signals are tuned up.

3.6. Some objections

There might be a fundamental concern, namely about what makes those dynamic and probabilistic concepts real entities. From the computational point of view, however, there is nothing problematic with assuming that dynamic functional webs do all of the work. On the psychological and phenomenological side, one’s introspective grasp of consciously accessible concepts may work similarly to one’s perception of the environment. Contrary to introspective evidence, and due to bandwidth constraints, one perceives in a detailed way only a small area of one’s visual field at each moment. The rest of the environment is represented as rough “summary statistics” (Cohen et al., 2016). We could apply this to concepts too. By attending to the word ‘cat,’ the entire functional web of ‘cat’ might be grasped in summary form. In the absence of vocabulary (as in the case of a feral child), the focus for ‘wolf’ might be, for instance, an exemplar or maybe some “mental location” associated with the error unit. The summary-statistics approach might also explain the stability illusion of word meanings (see also Casasanto & Lupyan, 2015). The summary representation might not be granular enough to capture the context-dependent modulations of the concept, giving rise to the impression of a stable meaning. Clark (2018) has discussed the following related puzzle: if the brain is driven probabilistically, then why does perceptual experience appear to be univocal and determinate? The answer that Clark gives is that cognition in the PP paradigm is there to drive action, and action requires at each moment the selection of a single, best model. The same idea might apply to the introspective experience of concepts.
To make the decision to run away from a tiger requires a clear-cut categorization of “that object over there” as a tiger.

Another important question that still needs further clarification is whether we should consider concepts not as constituents or components of complex thoughts, but as skills (see Clark, 1996, p.20, personal conversation; Barsalou, 2011; Glock, 2010; Margolis & Laurence, 2014). To possess the concept ‘cat’ is, in this view, to have the skill to produce appropriate cat thoughts or to successfully engage with cats. Such a skill would probably comprise sub-skills such as making available (context-dependently) a suitable subset of cat knowledge or adjusting the whole functional web corresponding to ‘cat.’ PP seems to be compatible with this view, but also with the constituent view, assuming that concepts are dynamic and flexible entities. The debate over whether concepts are abilities is closely related to the question of whether concepts are representations at all, which I mentioned in the introduction and in footnote 2. The question is how can such highly dynamic entities represent anything at all? I think that representations play an important explanatory role as entities with the help of which the agent successfully engages with the world, not necessarily with which he builds a veridical model of the world (see also Clark, 2015). However, a deeper discussion needs to be carried out elsewhere.

3.7. Conclusion

Within the predictive processing paradigm, concepts could be dynamic prediction units in the form of distributed, flexible, and cross-modal webs in a generative hierarchical model. Concepts are those representations in terms of which predictions are made, and they are therefore associated with error units that determine the prediction errors of the model. I have suggested that the PP framework could provide a cognitive-computational model for concepts, including a mechanism to account for concept contextualism. We could explain the modulation of the information that is retrieved when a concept is tokened by the PP-specific error-signal weighting mechanism that underlies precision-weighting. The precision and granularity of predictions go together. Second-order knowledge of both sorts is encoded in the overall model to allow for the tuning of error signals which, in effect, is a mechanism for switching concept features on and off.
Chapter 4. Overcoming the modal/amodal dichotomy of concepts

Abstract

The debate about the nature of the representational format of concepts seems to have reached an impasse. The debate faces two fundamental problems. Firstly, amodalists (i.e., those who argue that concepts are represented by amodal symbols) and modalists (i.e., those who see concepts as involving crucially representations including sensorimotor information) claim that the same empirical evidence is compatible with their views. Secondly, there is no shared understanding of what a modal or amodal format amounts to. Both camps recognize that the two formats play essential roles in higher cognition, leading to an increasing number of hybrid proposals. In this chapter, I argue that the existence of those fundamental problems should make us suspicious about a modal/amodal dichotomy. Also, I suggest that hybrid approaches, as they currently stand, do not provide suitable solutions to the impasse. Instead, we should overcome the dichotomy and treat the modal/amodal distinction as a graded phenomenon. I illustrate this hypothesis with an example of a cognitive-computational model of concepts based on the predictive processing framework.

Keywords: amodal representation; concept; concept empiricism; modal representation; predictive processing; representational format of concepts

4.1. Introduction

Concepts are considered by psychologists, philosophers, and cognitive scientists to be central building blocks for thought and cognition more generally. I presuppose here what is arguably the mainstream view about concepts, namely that they are mental representations of categories and associated bodies of knowledge or information (see e.g., Machery, 2009). Conceptual representations have semantic content, as they refer to some category in the world, and cognitive content, which consists of cognitively or psychologically significant information used for mental processing.

Many open questions surround the notion of concepts, like to what extent they are inborn, how it is possible that they can refer at all, and whether they have stable
cores that are instantiated with each tokening. A central and active debate in concept research concerns their format. The question is whether conceptual representations are amodal or modality-specific (short: modal). Modalism has traditionally been the common-sense view and is rooted in empiricism. Amodalism has then been a recent dominant view, in connection with the surge of the computational view of the mind, and especially with Fodor's (e.g., 1975) "Language of Thought" (LOT) hypothesis. Recently, modalist positions have also resurfaced strongly (some call it "neo-empiricism" \(^{33}\)) in the context of the embodied cognition paradigm, which stresses that our conceptual apparatus is being shaped by the constraints of our body and sensory apparatus (e.g., Clark, 2017; Lakoff & Johnson, 1999).

Recent modal views characterize conceptual representations as states corresponding to "re-enactments" or "simulations" of sensory or motor states involving the sensorimotor areas of the brain. To token the concept **DOORKNOB** is to token a mental representation similar to those mental representations tokened when a doorknob is perceived. In the case of motor states, representations are in an "off-line simulation" mode, i.e., they do not lead to the final execution of motor commands. For instance, to token an action verb involves the activation of parts of the motor brain area, suppressing efferent signals to muscles. Barsalou's "grounded cognition" has as a central tenet the modality of conceptual representations:

> [...] a diverse collection of simulation mechanisms, sharing a common representational system, supports the spectrum of cognitive activities. The presence of simulation mechanisms across different cognitive processes suggests that simulation provides a core form of computation in the brain (2008, p.619).

It is essential to point out that although the modalist view might have its roots in empiricism, it differs from traditional empiricism in some crucial aspects. Firstly, the modalist need not necessarily reject nativism (e.g., Barsalou, 2008, p.620, 2016, p.1123); the questions of concept format and nativism are orthogonal. Secondly, modal representations should not be confused with literal conscious mental images. Also, modalists have moved away from extreme simulation views and now allow for schematic, unconscious representations, as well as representations where various modalities are "convolved" (see Section 4.4.2) into multimodal representations.

---

33 E.g., Machery (2006).
Thirdly, modalists (e.g., Barsalou, 2008, p. 620; 2016) do not necessarily need to deny that in addition to modality-specific representations there are also representations that are not "grounded" in external experience, i.e., some positions are a hybrid, though biased towards modalism:

From the perspective of grounded cognition, it is unlikely that the brain contains amodal symbols; if it does, they work together with modal representations to create cognition. (Barsalou, 2008, p.618)

Amodal representational systems (e.g., Fodor, 1975; Pylyshyn, 1984), in turn, are not associated with a specific modality. They are formal, language-like and "abstract" and their symbols are processed syntactically, i.e., in virtue of some formal aspects (not meaning or content). Such representational systems work roughly like a formal calculus of symbols, much like a natural or formal language consisting of syntax/grammar and word forms. The motivation for amodalism is based on two key observations. Firstly, it is generally recognized that conceptual representations must be able to account for systematicity and productivity of thought. This requires amodal symbols that can be freely recombined to form novel concepts or propositions. Secondly, the existence of abstract concepts, like *democracy* or *truth*, purportedly requires amodal representations. For amodalists, it seems to be a contradiction in terms to have abstract concepts grounded in perceptual or motor representations.

While it seems intuitively clear what the two positions distinguish along the above-sketched lines, it turns out to be difficult to further characterize the difference between the representational formats. Authors define modality versus amodality in different ways, and none of the proposals available seems to survive more in-depth scrutiny (see Haimovici, 2018). A second fundamental problem concerns the available empirically support. From the debate it becomes apparent that the same evidence can be interpreted in ways that are compatible with each view, and it is still unsettled as to which view provides a better explanation of the phenomena.

In this chapter, I argue that in the face of those problems, we should be suspicious about the usefulness of the modal/amodal dichotomy. I suggest that we should overcome and reconceptualize it as a graded notion. For that purpose, first, I expand on the two fundamental problems that the dichotomy faces: the difficulty of fleshing out the distinction in precise and agreed on terms (Section 4.2) and the problem of
what evidence would count as support for the different views (Section 4.3). In Section 4.4, I deny that abstract concepts are more of a problem for modalists than for amodalists. I then review and reject recent hybrid approaches as an alternative (Section 4.5). In Section 4.6, I illustrate the reconceptualized notion of a graded distinction between modal and amodal formats using the example of a cognitive-computational model of concepts within the predictive processing framework, a relatively recent, but already well-established cognitive paradigm.

4.2. The problem of telling apart modal and amodal representations

The first fundamental problem concerns the very distinction between modal and amodal formats. Though one might have some intuitive grasp of such a distinction, there is no generally accepted criterion to tell apart modal from amodal representations. The point is not that there is no agreed upon conceptual analysis available for the notions of "modal" and "amodal", which would be too demanding, but that there is not even a rough criterion or a working definition that most authors share. In the absence of such common ground, one might worry that maybe the whole format debate is ill-conceived. Now, there are no generally agreed criteria for a distinction, but there are, of course, different working definitions that are used by various authors. The problem is that all of the characterizations suggested have issues (see Haimovici, 2018, for a more detailed discussion) and there is no suitable candidate to converge on.

Fodor, for instance, suggested a mereological criterion based on a distinction between icons and symbols. Every part of an iconic representation represents a part of the content, whereas this is not the case for symbolic representations. Similarly, some authors (e.g., Mahon & Hickok, 2016) appeal to the fact that amodal symbols are arbitrarily related to their content, while modality-specific representations have some isomorphic aspects between content and their vehicle. However, both approaches have a similar problem. The criterion might work well for visual-spatial representations, but it is far from clear how to generalize it to the many other sensory modes, e.g., to olfactory, auditory, proprioceptive, or interoceptive representations. For instance, the "parts" of (the projection of) a scene could stand in a one-to-one relation mirroring the retinal pixel arrangement, or some other neural activation patterns in some higher-level brain areas. This works well as the images and the cell
arrangement are both extended in space. However, how would this work for the other modes in which spatial extension is not essential?  

Another related distinction between modal and amodal representations is based on analogical and digital formats. In an analogical representation, some property of the representational vehicle co-varies continuously with what is represented. It is indeed plausible that many modal concepts can be placed in some "quality space", e.g., a colour space. For instance, the concept RED could be represented by some (convex) region in a three-dimensional "colour space" (e.g., Gärdenfors, 2014). However, such a space can be perfectly digitally encoded. To place RED into an "analogical space" is a higher-level interpretation of some other underlying lower-level representational format. Machery has also argued against the usefulness of characterizing modal symbols as analogical and amodal symbols as digital in nature (e.g., 2007, p.23) based on evidence that some amodal representations are analogue (e.g., representations of numerosity), and that there are visual representations that are not analogical.

Machery then suggests applying another criterion:

This does not mean, however, that we cannot distinguish between perceptual and amodal representations. Following Prinz (2002, p.113), one can propose that perceptual representations are whatever representations psychologists of perception say perception involves. (2007, p.23).

This expert criterion might be perfectly valid, even if it does not allow for a full and detailed further conceptual analysis (e.g., of necessary and sufficient conditions). However, even if no precise conceptual analysis can be provided, it certainly would be surprising if nothing further could be said by an expert to justify the distinction. Moreover, it might be the job of a philosopher to help to make the (possibly implicit) criteria more explicit. A merely deferential criterion of the modal/amodal distinction is therefore not very satisfying and should be only a last resort solution.

We could turn to a neuroscientist instead of a psychologist of perception and apply a neural location criterion that takes into account neurophysiology. Concept

---

34 What I mean is the following: a pictorial representation has spatial extension essentially, and what is represented is spatially extended. Smell might be spatially extended (e.g., a whole room might have a certain smell), but there is nothing spatial in the (modal) concept ROSY FLAVOUR.
representations are, after all, implemented in the brain, so the strategy is to analyse the neural activation patterns and identify their location. A representation is then classified as being modality-specific if during a semantic task the activation of neuron assemblies occurs in areas that are considered by neuroscientists to be sensorimotor processing areas. The fact that the presentation of, say, a cat activates neuron assemblies in, e.g., the primary visual cortex is taken as evidence for modality-specific representations. A lot of empirical support appealed to by modalists and amodalists presupposes this criterion. However, this proposal also turns out to be problematic, as we will see later in more detail. Most of the discussion of this chapter will assume a neurophysiological criterion, as I am concerned with an account of concepts empirically informed by the neurosciences.

Another, related, approach to characterizing amodal versus modality-specific representations is based on the sort of input that a representation receives. Authors like Prinz (2002), Dove (2009) and Dehaene (2011) suggest that amodal representations respond to different types of sensory modalities, not just one. Those authors appeal, for example, to number concepts. For instance, we can classify three things independently whether they are three objects, sounds, or actions. However, this account could be accused of conflating amodal and multimodal representations (see also Haimovici, 2018, p.3, for the same point) and would, therefore, not clearly distinguish modalism and amodalism. I will say more about representational abstraction and abstract concepts in a moment. Finally, Barsalou has proposed an "independent systems criterion", which could be seen as a specific proposal for a neurophysiological criterion: "...cognition is computation on amodal symbols in a modular system, separate of the brain's modal systems for perception, action, and introspection" (Barsalou, 2008, p.617). "Independence" could be functional or anatomic. However, as the later discussion will show, such a strict dichotomic separation is implausible.

With this quick and condensed review, which does not pretend to be an exhaustive evaluation, I want to make the point that we are not short of proposals to tell modal and amodal formats apart. But all of the proposals have issues and there is no consensus as to what the appropriate one is.
4.3. The problem of evidence for modal and amodal representations

The second fundamental problem for the amodal/modal distinction concerns the empirical support for either position. In this section, I argue that the empirical evidence used in the modal/amodal debate is not conclusive (see also the review of Dove, 2016, pp.1110-1111). What we can conclude safely from the evidence, however, is that extreme modal or amodal positions are not tenable, and, indeed, both modalists and amodalists increasingly move to hybrid accounts. The question then remains as to whether some available hybrid account provides a suitable model for conceptual format. I can't review here, exhaustively, the vast body of empirical results, so instead I will focus on the big picture and some representative examples to make my point. For more detailed reviews I refer to the literature (e.g., Barsalou 2016; Dove, 2016; Kemmerer, 2019; Machery, 2016; Meteyard et al., 2012).

To start with, let me differentiate further between the various positions in the modal/amodal debate. Meteyard et al. (2012) usefully introduced a taxonomy of the views located on a continuum from "strongly embodied" to "completely unembodied". Completely unembodied (fully symbolic) views (e.g., Mahon, 2015; Mahon & Caramazza, 2008) hold that concepts are amodal representations and modal information does not play any relevant role in conceptual representation, i.e., semantic content is independent of sensorimotor areas. Strongly embodied (full simulation) views reduce conceptual processing to the level of sensorimotor (modal) representations (e.g., Gallese & Lakoff, 2005; Glenberg & Gallese, 2012). A consensus seems to emerge that extreme views have little empirical support, and a compromise is needed (e.g., Borghi et al., 2017; Chatterjee, 2010; Dove, 2016; Meteyard et al., 2012; Reilly, Peelle, Garcia & Crutch, 2016). To see this, let us briefly review three examples of empirical strategies that have been deployed to reveal the nature of conceptual format. I suggest that the evidence does not adjudicate the debate. However, we can conclude that: a) sensorimotor representations play a pervasive role in conceptual processing (though the question of whether they are a constitutive part of the conceptual representation remains open), and b) some form of abstracted representations is needed (though the question remains as to whether those abstracted representations are amodal or count as modal).
Among the most popular empirical strategies employed is the identification of activation patterns in sensorimotor areas during conceptual processing using neuroimaging techniques like fMRI. Many studies (e.g., Hauk, Johnsrude & Pulvermüller, 2004; Chao & Martin, 2000; Simmons, Martins & Barsalou, 2005) have demonstrated the relevance of sensorimotor activity when concepts are processed. However, while this happens in many instances, there are exceptions. As an example, it turns out that on some occasions processing of an action verb does not activate action areas in the brain (e.g., Barsalou 2016; Dove, 2016; Kemmerer, 2015). Also, Pecher (2018) recently showed that motor representations are not activated automatically; hence their activation is not always necessary for conceptual processing. This suggests that sensorimotor areas are often, but not always involved when concepts are tokened. While this most likely excludes the extreme grounded (modal) view, we still cannot distinguish whether the co-activated representations are part of the concept, or consequence of "spreading activation" (e.g., Mahon, 2015, p.420). Leshinskaya & Caramazza (2016) suggest that tight coupling or coactivation of conceptual and sensorimotor representations is evidence for the interaction of conceptual and sensorimotor representations, but not for concepts being modal. A fundamental difficulty in deciding the debate by this route resides in the complexity of establishing in a principled way how fast or far spreading can be so that the firing neurons still count as a constitutive part of the same representation. A related strategy, also based on neuroimaging, is to establish whether different modality-specific cues related to a concept activate a common representational core in regions that can be considered not to be modality-specific (see Barsalou, 2016; Fairhall & Caramazza, 2013; van Doren et al., 2010). However, evidence for shared cores seems consistent with both weak modalism and amodalism. Weak modalists can account for this phenomenon by claiming that the core is multimodal and abstracted (i.e., it still contains - compressed - modal information).

Scientists have also turned to a strategy based on detecting a causal role of the two types of representation via neurophysiological lesion studies (e.g., of patients with semantic dementia) or Transcranial Magnetic Stimulation (TMS) experiments. The idea is to explore whether modal or amodal representations are necessary for semantic comprehension. If, for instance, the motor-area is permanently or
temporarily impaired, but the understanding of action words remains intact, then it seems that sensorimotor areas are not necessary for concept representation (and strong modalism must be false). For instance, Repetto et al. (2013) showed that the stimulation of the hand portion of the primary motor cortex leads to slower reaction times for hand-action verbs, indicating that sensorimotor areas play a causal role in verb comprehension. Similarly, Gerfo et al. (2008) showed that repetitive TMS (rTMS) stimulation of the left motor cortex delays the processing of action verbs and names. However, Vannuscorps et al. (2016) document the case of a patient with increasing atrophy of sensorimotor regions (leading to an increasing action production disorder), but persistent intact performance with action-concepts. This shows that motor-representations are not necessary for all conceptual tasks. Pobric et al. (2010) showed - with a reverse strategy - that rTMS on the temporal poles leads to reduced efficiency in semantic tasks but does not have an impact on perceptual tasks. The authors conclude that this is evidence that the poles play a role as amodal processing sites. However, all this evidence is not a problem for weak modalists. They only need to admit that low-level sensorimotor representations do not need to be activated in all cases, as full simulation modalists would claim. The weak modalist only needs an account that includes abstracted modal (or multimodal) representations.

As a final example of an empirical strategy, take the appeal to behavioural evidence. Recently, Fisher & Shaki (2018) have studied the performance signature for number concept processing. The results support the claim that the processing of paradigm examples of abstract (and hence purportedly amodal) concepts shows clear characteristics of perceptual processes. The authors have identified a range of effects that are typical for perceptual discrimination and that are preserved when numbers are processed in symbolic form: for instance, distance effects (e.g., 3 and 9 are easier to distinguish than 3 and 4), size effects (e.g., 3 and 4 easier to distinguish than 8 and 9) and spatial-numerical associations (numbers seem to be located on a spatial number line) revealed by motor-behaviour. This seems to be evidence for modalism. But amodalists can recognize the importance of modal representations in higher cognition and argue that conceptual processing sometimes uses perceptual heuristics, while number concepts remain amodal representations (see, e.g., the "Offloading" account in Section 4.5)
The last example involved abstract concepts (numbers). So far, we have not explicitly distinguished between concrete and abstract concepts. That distinction, however, plays a central role in the debate. While it seems quite plausible that concrete concepts could somehow be represented modally, amodalists have been concerned that modalism is incompatible with abstract concepts on both empirical and theoretical grounds. Maybe abstract concepts are then the Achilles heel of modalism that tips the balance towards amodalism. However, I will argue now that abstract concepts are not more of a challenge for modalism than they are for amodalism. Therefore, the impasse remains intact.

4.4. Do abstract concepts support amodalism?

The existence (and pervasiveness) of abstract concepts has been one of the principal arguments against modalism (see, e.g., Dove, 2016; or Löhr, 2018, for discussions). Prominent examples in the literature are, for instance, number representations (e.g., Dehaene, 2011; Fischer & Shaki, 2018; Machery, 2007:34), and concepts like DEMOCRACY and TRUTH (e.g., Dove 2009, 2016; Löhr, 2018). Dove (2016) has summarized some of the main challenges purportedly posed by abstract concepts to modalism: a) generalization, b) flexibility and c) disembodiment. Let us unpack those briefly (4.1) and then see how a modalist can respond (4.2).

4.4.1 Dove’s challenges from abstract concepts for the amodalist

Dove thinks that the "generalization" involved in abstract concepts is a challenge for modalism. Generalization has a horizontal dimension, which consists of the extension of a concept with new exemplars, and a vertical one, which corresponds to an organization in terms of super- and sub-ordinated concepts. According to Dove, the claim that concepts are structured in hierarchies of abstraction is supported by evidence such as cross-modal deficits or hierarchical degradation of conceptual knowledge as well as evidence of the existence of areas that are not modality-specific (2016, p.1112), i.e., show an "abstracted" behaviour. With regards to the "flexibility" involved in abstract concepts, for Dove it seems to be a challenge for modalism that "some individual concepts can be used in either a more or a less

---

35 Interestingly, Dove does not elaborate on how amodalism can account for those aspects
grounded fashion, depending on the circumstances." (2016, p.1113). For instance, an fMRI experiment by Saygin, McCullough, Alac, and Emmorey (2010) showed that when the brain processes "The wild horse crossed the barren field", motion-sensitive visual areas were more active compared to other sentences containing the verb "to cross", like "The hiking trail crossed the barren field". The third challenge rests on the claim, according to Dove, that concepts like ODD or TRUTH seem "divorced from experiential factors" (2016, p.1114) and, therefore, it is difficult to see how abstract concepts can "even in principle" be grounded in sensorimotor representations. Finally, he cites a vast amount of evidence for an abstract/concrete asymmetry (i.e., some areas are preferentially activated for abstract concepts in representing and processing concepts) (2016, pp.1114-1115) as support for amodal representations.

Modalists have embraced different strategies to face the challenges posed by abstract concepts. One suggestion that is gaining momentum is that abstract concepts are grounded not only in the modalities of the five traditional senses, but also in interoceptive states (see, e.g., Barsalou & Wiemer-Hastings, 2005; Connell, Lynott & Banks, 2018; Fingerhut & Prinz, 2018; Vigloccio et al., 2014). This might plausibly work for concepts like FREEDOM or ANGER, but it is unclear how affective grounding could help, for instance, with ODD or TRUTH. Also, not all authors agree that interoceptive states have a central role, and Lenci et al. (2018), for instance, suggest that linguistic representations are needed and play the primary role in abstract concept representation. They deny that the affective load of abstract concepts refutes the position that abstract concepts are exclusively linguistically represented. They claim that affective information could be linguistically derived or a by-product of co-occurrence statistics (but see Vigliocco et al., 2014, who argue against a primary role of linguistic information for conceptual representations). Indeed, some modalists find the idea of combining modal grounding and linguistic representations into a hybrid appealing (e.g., Louwerse, 2018; Pecher & Zeelenberg, 2018) (see also Section 4.5, where I discuss representational pluralism).

However, as I will argue in a moment, the modalist does not necessarily need linguistic in addition to sensorimotor-plus-interoceptive representations for a defence. I have already alluded to elements of a (weak) modalist strategy, namely the appeal to abstracted multimodal representations. Let me expand more on the sort of
representations involved and then respond to Dove's challenges on behalf of a modalist.

4.4.2 A possible modalist response

Modalists, recognizing the need for abstraction, could appeal to a representational structure of concepts based on a modal abstraction and convolution hierarchy (let's abbreviate it by "MACH"). What a modal hierarchy of abstraction amounts to can easily be derived from contemporary neuroscience and AI (specifically deep learning). Modal processing comes with a built-in abstraction process. Take, for instance, the ventral processing stream of visual information consisting of a flow from the retina through to the cortical areas V1 -> V2 -> V4 -> IT. As one advances in the stream, the receptive field size of the representations increases, and the representations get more and more abstract. But they remain—quite indisputably—visual.36 Abstraction per se does not eliminate modality. Single neurons or neuron assemblies represent, say pixels, in early-stage retinal processing. In a later stage, a single neuron or a neuron assembly represents the shape of a certain edge. In each step, the brain abstracts from details available in previous stages. Similarly, the mixing of two or more modalities (convolution) does not lead to a representation that is devoid of modality. Different modalities can be "convolved" or folded into each other (see, e.g., Thagard & Findlay, 2012; Radu et al., 2018; Ramachandram & Taylor, 2017, for deep multimodal learning). "Convergence zones" (e.g., Meyer & Damasio, 2009), "supramodal areas" (e.g., Fairhall & Caramazza, 2013) or "hubs" (e.g., Patterson & Lambon Ralph, 2016 - see also Section 4.5) posited by neuroscientists, could be locations where convolution happens. There, different modalities flow together to create more abstract, multimodal representations (not necessarily amodal ones, as is often claimed). Those representations can be "unpacked" top-down by co-activating appropriate lower-level representations and providing more granularity or detail to the representation (and cognitive phenomenology) in different modalities. For instance, the highest level (most abstracted) representation of the concept THREE might be in the form of three "vague and schematic things" (where "thing" corresponds to some highly abstract concept

36 V1 (primary visual cortex in the occipital lobe) seems to respond to simple local edge structures, V2 more complex curves, V4 even more complex shapes and IT (roughly the inferior temporal cortex) represents complex objects, like faces.
THING, which includes any possible entity not only tangible things). Or, by involving co-activated lower-level representations, it might be in the form of three schematic apples, three specific green apples, three schematic sound events or three specific sounds. Similarly, DEMOCRACY can be seen as a very complex high-level multimodal representation that we might unpack context-dependently in many fashions and mixtures; for instance, in the form of a voting scene, but also as a definition, as an exemplar in the form of a paradigmatically democratic country, or as some subjective feeling of justice and freedom. Whatever has been folded into (by concept formation, or by evolution) the highest-level node of the hierarchical network structure of the concept DEMOCRACY can now be retrieved selectively, and with the level of detail or schematicity needed, depending on the context and task.

MACHs allow a response to Dove’s challenge in the following way. The hierarchical structure can, by definition, account for vertical and horizontal generalization. Representations are organized into abstraction trees. Nodes form a vertical abstraction gradient, and all child-nodes of a parent are related horizontally. Regarding the challenge of flexibility, it is a challenge as much for amodalism as it is for modalism. One needs to come up with a mechanism to account for the high degree of context-sensitivity of concepts, so modalism is not worse off in this regard. A more specific computational proposal is needed to advance here. In Section 4.6, I suggest a mechanism that a modalist could appeal to. The third challenge, disembodiment, rests on the claim, that abstract concepts seem quite remote from direct experiential representations. However, it is not clear why it should be, in principle, impossible to represent abstract concepts that involve categories of events, situations and mental states in terms of abstracted and convolved modal information. Of course, such representations must undergo a very complex abstraction and convolution process using a wide range of modalities (including interoceptive states) and it might be difficult to decompose them into simple experiential components. Finally, the difference in behaviour due to the modal/amodal asymmetry can also be explained naturally given that there is a gradient of abstraction. The ends of the hierarchy might, of course, “behave differently”. At the more abstracted end, representations behave "amodally", while closer to the periphery (the bottom of the MACH) they behave "modally" (perceptually).
Let me highlight that MACHs work for abstract and *concrete* concepts. The challenge of generalization is as much a challenge for concrete concepts, as it is for abstract concepts. In a certain sense, abstract concepts are not qualitatively different from concrete concepts. A concept denotes a category, and any category is abstract by definition. **DOG** is abstract, though you can touch, see, smell, etc., *exemplars* of **DOG**, i.e., dogs. **DOG** is in this sense as abstract as **DEMOCRACY**. The difference resides in certain characteristics of exemplars. Exemplars of **DEMOCRACY** must be very complex states or situations indeed. So abstract concepts are not different in type, but merely require significantly more complex modal abstractions and convolutions, so the modalist can argue. That amodal representations should prima facie be better suited for abstract concepts rests, I suspect, on a confusion. Merely appealing to the "abstract" nature of amodal representations does not explain how they can be representations of abstract concepts. This would be conflating two readings of "abstract", one referring to a property of the vehicle (the representation and its degree of information compression) and one related to the content of what it represents (a certain category whose exemplars share certain characteristics). Amodalists do not have an advantage here then; in fact, quite the contrary. A weak modalist can explain how concepts can mean anything in the following way: if basic level representations get abstracted (compressed) to higher level (still modal, but less detailed) representations, and the meaning is in this sense grounded in basic level representation, then the more abstract representation inherits content from below. The amodalist needs to appeal to arbitrary symbolic relations and explain how those symbols can refer to and can mean anything. There is a range of proposals available, of course (see, e.g., Tillas & Trafford, 2015). My point is merely that amodalism is not a no-brainer default position for abstract concepts and one needs to be careful about being drawn into an intuition based on the above conflation of the notion "abstract".

Let us take stock. Empirical results have not been able so far to adjudicate the modal/amodal debate. The "challenge of abstract concepts" turns out not to be an insurmountable stumbling stone for modalists and is a challenge for amodalism. However, despite this situation, the field has advanced substantially by accumulating quite compelling evidence for the significant involvement of sensorimotor representations in conceptual processing, and also for the involvement of either
amodal or abstracted (multi-)modal representations. Extreme positions on the continuum of Meteyard et al. are therefore unlikely winning proposals. In general, it seems possible to concoct intermediate positions on both sides, to account for most of the evidence. However, the question is not so much whether we can somehow account for the evidence, but rather what account provides the best explanation in terms of other virtues like theoretical simplicity, consistency, coherence, and fruitfulness. So, it is worthwhile having a look at some more specific hybrid proposals to see if one of them provides a way out of the impasse.

4.5. Are hybrid approaches the way out?

An increasing number of hybrids try to accommodate the evidence for the importance of sensorimotor representations and the existence of abstracted representations. In what follows, I briefly review four examples: two proposed by amodalists, and two by modalists. As we will see, hybrids built on the modal/amodal distinction have drawbacks and seem unable to resolve the debate.

Mahon and Caramazza (2008) acknowledge that modality-specific information plays a crucial role in the use of concepts. However, they insist that only the amodal representation is constitutive of the concept:

> On the grounding by interaction view, the specific sensory and motor information that goes along with the instantiation of a concept is not constitutive of that concept (p.68).

However, the "grounding by interaction" account of concepts implies a very anaemic notion of concepts. If I am correct, their implied notion of concept is concerned exclusively with the referent, and hence with questions covered by the intentionality desiderata for a theory of concepts (see Prinz, 2002). Cognitive content and psychological significance are relegated to a secondary, non-conceptual role. Their account is also formulated quite generically, and they provide no specific cognitive

37 Thanks to an anonymous reviewer for helping me see this more clearly.

38 As an example, Keas (2018) names twelve virtues of good theories: "evidential accuracy, causal adequacy, explanatory depth, internal consistency, internal coherence, universal coherence, beauty, simplicity, unification, durability, fruitfulness, and applicability."

39 There are other authors who have gestured at hybrid solutions, e.g., Löhr (2018, pp.20-21) and Binder (2016, p.1096). See also Dove 2016, pp.1115-1117) for an overview.
mechanism of how this interaction is supposed to work. It seems we cannot empirically distinguish it at that general level of formulation from an account in which such modal information is constitutive of the concept and concepts retrieve context-dependently modal information. If an amodal concept representation is often accompanied by a co-activated modal representation and does significant cognitive work, according to what principles is that modal representation not a constitutive part of the concept?

Machery's "off-loading hypothesis" shall serve as a second example of a hybrid account. Machery (2016) acknowledges that we often use perceptual and motor representations to solve cognitive tasks. However, he rejects the conclusion that this implies that (at least some) concepts are modal. He suggests that we offload many cognitive tasks from the amodal conceptual system to sensorimotor representation. Motor and sensory representations are hence not constitutive of conceptual representations but are used heuristically:

In contrast, according to the offloading hypothesis, we often offload the solution of tasks on perceptual and motor systems: While concepts themselves are amodal, we often manipulate perceptual and motor representations to solve tasks. [...] Offloading may happen when the conceptual system does not encode the information needed for solving a given task (e.g., information about perceptual details), while perceptual representations stored in memory do. Offloading also may happen for tasks that can be efficiently solved this way (2016:1094).

This is an interesting proposal, and it seems to imply the existence of some algorithm or mechanism that implements the offloading heuristics. If the amodal system is not able to solve a cognitive task alone, it uses the resources of modality-specific representations. This is a hybrid proposal in the sense that it implies the distinction of two separate representational systems that interact. Again, my concern is whether the offloading hypothesis is specific enough to be empirically testable. What makes a particular activation pattern in the modality-specific regions an "offloading" as opposed to a context-sensitive co-activation of that information? Also, how could it account for some concepts, like specific colour concepts, that seem to come by default with some (maybe vaguely) imagined colour impression?

Let us turn to modal hybrids to see whether they fare better. The Hub and Spokes model (HSM) (e.g., Rogers at al., 2004; Patterson & Lambon Ralph, 2016; Binney,
Parker & Lambon Ralph, 2012; Guo et al., 2013) suggests that both modality-specific (spokes) and amodal information (in the "transmodal hub") are necessary components of a concept representation. The modality-specific aspects of a concept are represented in the corresponding sensorimotor (and linguistic) areas. The hub-component sends and receives information from the modality-specific regions. The hub abstracts away from specific modal features and codes the "semantic similarity structure". The hub-component, therefore, unifies the different modal information sources and provides a coherent and generalizable concept. Both hubs and spokes are necessary and the HSM does not imply that concepts have an abstract form and reside in the hub region (which is proposed to be located in the anterior frontal lobe, the ATL). For the necessity of hubs speaks, according to the authors, evidence from studies of patients suffering semantic dementia (SD): ATL atrophy leads to SD. Cross-category loss of classification and generalization without deterioration of modality-specific areas indicate that the problem must be in the integration of modal information. Evidence for the HSM, however, is compatible with the modal view based on MACHs. The hubs are simply areas that contain modal abstracted and convolved representations. But the evidence for the HSM is not clearly evidence for a dichotomic modal/amodal model. Indeed, some authors have suggested that the role of ATL as "the" hub is overemphasized (see overview of ATL functions by Wong & Gallate, 2012) and the ATL has many other functions and in many other regions representational abstraction happens.

The second example of a hybrid leaning towards a modal view of concepts is the "Symbol Interdependency Hypothesis" (SIH) account (e.g., Louwerse, 2018), which is an account of representational pluralism. It combines modal and linguistic representations as mutually reinforcing. The motivation stems from the following sort of reflection. We might learn concepts without the intervention of sensorimotor input, for instance in school via definitions and verbal explanations. Also, we often bootstrap meanings via the context in which a word appears. Therefore, language plays an important role in concept acquisition. Given the role of linguistic representations, we might say that amodal representations play a role in concept representation and sometimes concepts are represented linguistically, i.e., amodally. This provides a basis for meaning via indirect grounding: the word is grounded indirectly via the surrounding grounded words. This view is, arguably, modally
biased, as grounding is necessary, though the requirement is weakened by allowing indirect grounding. The SIH account then claims that amodal representations encode semantic information via distributional statistics. Words get their meaning from direct grounding and from indirect grounding via the linguistic context. Representations grounded indirectly allow then for at least "quick and dirty representations", while a deeper understanding would require direct grounding. I am very sympathetic towards this approach, but I see various problems as it stands. Firstly, it is not entirely clear what takes the role of amodal representations. Are they linguistic natural language representations? This would mean giving up Fodor's LOTH which does not rely on natural language but mentalese. Giving up mentalese might be an option, of course. However, this assumes that natural language representations are amodal, which can be debated, because they involve sound, gestures and/or visual patterns (see Langacker, e.g., 1987, 2008, who endorses that linguistic representations are modal\textsuperscript{40}). The SIH account claims that the meaning of unknown words is grounded by their "distributional statistics". It is difficult to see how the statistics themselves ground the meaning of the words. It seems to me that we understand an unknown concept appearing in a certain linguistic context not in virtue of the wordforms by which it is surrounded, but in virtue of the content those surrounding wordforms represent. Keeping in our memory information about the statistics of surrounding words might be merely a temporary heuristic, with the ultimate aim being to extract direct grounding indirectly from the surrounding words. The statistics would then play the role of a mere placeholder. It seems more plausible that words and their statistics provide access to meanings but do not constitute them.

In sum, hybrid accounts try to combine the need for abstracted representations with the fact that sensorimotor representations are pervasively present in cognition. However, the amodally biased accounts have an ad-hoc air and are quite unspecific, while the modally biased accounts seem slightly better motivated, but face other problems. So, it is not yet clear that hybrid accounts can resolve the debate. Some authors (e.g., Dove, 2016) have suggested, in the face of the empirical stalemate, that weak modalism is not a position that is distinguishable from

\textsuperscript{40} Langacker suggests that a linguistic representation is a symbolic relationship between two modal representations: a conceptual representation and a phonological representation.
amodalism. However, Dove maintains the dichotomy and claims that it is modalism (embodiment) that collapses into amodalism. I wonder why, if both positions are indistinguishable, he then does not consider the possibility that it is amodalism that collapses into modalism. Dove assumes in his argument that an abstract representation is an amodal representation. But this is an unjustified conflation, as abstract must not necessarily mean "void of modal information". A second possible response to the empirical deadlock could be given along the lines of Machery (2007). Machery refers to "Anderson's problem" (see Anderson, 1978). Anderson already observed the difficulty, in principle, of distinguishing modal and amodal representations: "The correct conclusion from Anderson's argument is that amodal theories and empiricist theories are on par" (Machery, 2007, p.31). Machery then suggests that we need more detailed and specific modal and amodal theories for a given cognitive task that allow us to derive and test "contrastive predictions". However, so far, we have no example of such a cognitive task for which more specific weak modal and amodal theories have been developed and contrastive predictions derived. I agree with Machery that more specificity in the proposals might be required for the debate. However, note that all accounts discussed here are based on some quite unclear modal/amodal dichotomy. When searching for a suitable theory of conceptual representations, ceteris paribus, a more integrated account out of which a distinction between the two representational types arises in a principled way would theoretically be more pleasing. Therefore, I suggest considering for a moment, whether it might not be the very dichotomy, presupposed widely in the debate, which is the source of the troubles. In the next section, I will provide a computationally (and neuronally) more specific account of conceptual representations to show how we could understand the modal/amodal distinction as one of degree. To make the proposal specific enough, I will use a cognitive computational framework, grounded in neuroscience, namely the so-called predictive processing (PP) framework.
4.6. Overcoming the modality/amodality dichotomy: an example

4.6.1. Predictive processing and concepts

There is no space here for a detailed exposition of the PP framework. Given that PP has already been widely covered in the literature and many useful introductions are available, I will only very swiftly summarize the bare-bone essentials of PP, which are necessary to follow my example, and refer to the literature for a wealthier background. I will then describe a recently proposed model for concepts within PP (Michel, 2020a, 2020b).

Predictive processing (PP) (see Clark, 2013, 2016; Hohwy, 2013; Friston, 2010) pictures the brain as a dynamical prediction device that constantly predicts its sensory input and updates its model to minimize prediction error. The brain uses a multi-layer probabilistic prediction model in which approximate Bayesian inference is carried out (e.g., Clark, 2013, pp.188–189; Hohwy, 2013, pp.15–39). The PP model has a hierarchical structure and represents prior knowledge on many levels of abstraction (e.g., Clark, 2013, p.25; Lupyan & Clark, 2015). Information flows bottom-up and top-down in this system. In the downward prediction cascade, the predictions of higher-level layers serve as priors for the lower-level predictions and, in this way, constrain the hypothesis space on the lower level. Computations in the PP model are driven by the goal of minimizing the average prediction error in the long run. The PP system also contains a mechanism of precision-weighting of the prediction errors (Clark, 2016, pp.53-83). The brain must predict the reliability of its sensory input to be able to distinguish between noise and useful signals. In this way, it can avoid modification of the model due to noisy signals. For that purpose, the mechanism assigns weights to the error signals and thus determines the influence of the top-down predictions versus bottom-up driven updates of the model.

To show how modality might be seen as a graded notion, I will use as an example a cognitive-computational model for concepts within the PP framework (see Michel, 2020a, 2020b). According to this model, concepts are "prediction units" (or "concept units" as I will call them here). Concept units are the vehicles of predictions in the PP framework. They play a crucial role in efficient predictions because they are the entities in terms of which predictions are made with the appropriate level of detail. For instance, when crossing a street, it is not efficient for the brain to predict the
presence of a car on a pixel-level of detail. Instead, it should be predicted in a more compressed and schematic way. Concept units are interconnected in a hierarchical network structure covering the whole range, from early sensory representations to representations in the cortical brain areas. The information associated with a concept (features) consists in the connection to other concept units. The information retrieved (i.e., other co-activated concept units) when a concept is tokened can be context-sensitively modulated. Very roughly, the PP precision weighting apparatus allows for switching on and off concept features (i.e., connections to other concept units) that are relevant to the context.

4.6.2. Overcoming the dichotomy

With a specific cognitive-computational model for concepts in place, I will now show how to overcome the modal/amodal dichotomy and suggest how to reconceptualize modality as a graded notion within this model. Let me start by linking the picture of concepts just sketched with the idea of increasingly abstract representations in a hierarchical representational structure, as posited by PP. The higher a concept unit is located in the network hierarchy of the PP model, the more abstract or compressed the information corresponding to that single node is. On the lowest level of the hierarchy, we have representations in the sensory-motor periphery. One might not want to call those low-level representations “conceptual”, but nothing hangs on it. The critical point is that we have a multi-level hierarchical structure of interconnected representational units that are increasingly abstract from the bottom to the top. Furthermore, and crucially, the context-dependent instantiation of a concept might span a network of nodes across an area of varying extension in the hierarchical model. Now, with such a view of concepts, the dichotomy modal/amodal does not cut much ice anymore. To see this, I will argue from two perspectives, the amodalist's and the modalist's one, to be charitable to both (remember, we have concluded that empirical evidence does exclude extreme views but does not decide between weaker versions of modalism and amodalism).

a) From the amodalist perspective

Take, for instance, the neural location criterion, implicitly assumed by many amodalists, which holds that a concept is amodal/modal if it is located in a (generally recognized) amodal/modal processing area of the brain. Assume that we could
localize the highest-level concept unit in an area that is agreed to be amodal. However, the concept token also includes other feature nodes, and some of them might or might not be in brain areas that are agreed to be modal. That depends on the concept and the context. So, rather than saying that a concept is modal or amodal, we should say that a concept can have amodal or modal instantiations: if all of the co-activated features are in amodal areas the concept is amodally instantiated; if at least one feature falls in an area that can be characterized as modal, it is a modal instantiation.

One could object and suggest that one should characterize the concept depending only on the location of the highest-level root-node and ignore the co-activated feature nodes to portray the concept as modal or amodal. If the root node is located in an amodal brain area, we are dealing with an amodal concept; otherwise, we are dealing with a modal one. However, that seems quite arbitrary, because why should the co-activated features be ignored? In many cases they are most likely to be co-activated because they are cognitively relevant and useful in the cognitive task. Also, given the hierarchical structure with the built-in graded notion of abstraction (with increasing abstraction from bottom to top), the introduction of a sharp dichotomy does not seem justified. Instead, it seems more adequate to carry the graded notion of abstraction over to a graded notion of modality.

In this model, a concept is not modal or amodal simpliciter. But this view does not imply that we have to give up either notion. For instance, there is a sense in which we could still give amodality a vital role. There is, namely, a sense in which concepts can be tokened in an amodal mode, without being an amodal concept simpliciter. For that purpose, let me introduce the notion of "shallow" and "deep" processing of a concept inspired by Barsalou (e.g., Simmons et al., 2008; Barsalou, Santos, Simmons & Wilson, 2008) and other authors (e.g., Erickson & Mattson, 1981, or Barton & Sanford, 1993), which might be useful here. Their idea with regard to the depth of processing can be applied to the PP model of concepts as feature networks. The basic idea is that, for example, in reading a task, a concept might be

---

41 For reasons of space and scope I cannot discuss here the relationship between language and concepts, and how reading comprehension might work within the PP model. For the purpose of this chapter not much hangs on a specific account, as long as we deal with concepts as context-sensitive network structures that span wider brain areas.
processed—at the one extreme—only superficially. During such "shallow" processing only a small part at the top of the network of a concept is activated (in the limiting case only the root node of the concept unit itself). For instance, when analysing a syllogism, one need not activate the full concept network of the involved words, and one can ignore most connections to other concepts and treat the words as mere placeholders (though, of course, it is difficult to completely suppress the meaning when reading a word). A shallow representation is enough for the purpose at hand. Or, to give another example, in the case that one has a very superficial understanding of some concept (maybe a technical term one is not familiar with), the processing is quite shallow, simply because the concept network is small or even limited only to the linguistic label.42 At the other extreme, when reflecting very consciously on the meaning of a word, the resulting activated representation might be extremely rich, including, e.g., sensory-motor information regarding exemplars associated with that concept.

The PP story of concept contextualism provides resources to account for the processing modes of conceptual webs that vary in terms of depth. We could imagine that, on some occasions, concept tokens are instantiated only by the root node, possibly together with a few other nodes in adjacent hierarchical levels, without reaching deep into low level peripheral sensory or motor areas (though they could, depending on the context, of course). So, we could have settings in which concepts are processed shallowly. But that would be merely a limiting case on a continuum from very deep to very shallow processing. A concept could appear amodal in a shallow processing mode. But in appropriate contexts, the same concept could also be processed in a modal mode, in which concept units in lower-level sensorimotor areas are co-activated.

b) From the modalist perspective

So far, we have assumed a neural localization criterion, which presupposes the existence of (genuinely) amodal areas in the brain. But, as we have seen, a weak modalist might deny the existence of amodal representations in the first place and point to MACHs. Concepts are more or less abstracted and convolved modal

42 See also Carey, 2004, p.66, for a view that implies the possibility of initially thin concept representations. Concepts can be bootstrapped using words.
representations (that are never fully free of modal information, i.e., amodal). But the modalist could be aligned with the PP view, where concepts are instantiated flexibly as networks with nodes across a continuum from low-level sensorimotor nodes to highly abstracted and convolved ones. On the other hand, amodality could now be seen as an (unreachable) limiting case, or asymptote, of maximally shallow processing of nodes (they may, though, vary in their degree of abstraction, depending on the level on which they are located). The more abstract and the shallower the instantiation, the more the concept "looks" amodal. Some modalists have suggested taking on board linguistic representations (Section 4.5). We can't cover here the relationship between concepts and language but let me hint at the following suggestion (which might allow for fleshing out more consistently the hybrid proposals that combine modal and linguistic representations and which I have criticized in Section 4.5). Modalist could allow for (arbitrary) linguistic labels (i.e., other representations not involved in the hierarchical abstraction gradient) attached to the root-nodes of concepts. This move introduces the possibility of an "amodal" instantiation of a concept, and, in this way, the modalist can "close" the modal-amodal continuum at the amodal end. An arbitrary label in itself would no longer carry abstracted modal information and, if instantiated alone, would be merely a meaningless (shallow) placeholder. Maybe some concepts (namely entrenched lexicalized ones) have such labels as their root nodes.

In sum, from both perspectives, that of the amodalist and the modalist, it turns out that the modal/amodal dichotomy does not look very useful anymore and it should be overcome by reconceptualizing it as a distinction of degree. If the tokens of a concept can (context-dependently) cover a whole range of levels in the PP model hierarchy, there is no reason to call the concept modal or amodal simpliciter, and it would be better to characterize the modal/amodal distinction as one of degree. Concepts do not fall into modal and amodal concepts. The amodal/modal continuum is parallel with the continuum of shallow/deep processing and the continuum of increasing abstraction from the bottom to the top.

4.6.3. Some benefits of the model

The picture of concepts as located in an amodal/modal continuum that I put forward here has various advantages. Firstly, it is based on a cognitive-computational model
that is specific enough to carry the hope that we can test it empirically. Furthermore, it can accommodate both the concerns of modalists and amodalists because it accounts both for semantic and cognitive content. Indeed, if we consider only the root node, we can account, for instance, for the intuitions behind Fodor's amodalism (conceptual atomism). Fodor is mainly concerned with reference and semantics, not with psychological and cognitive significance. The root node plays an "atomic" role. Under certain circumstances, we can idealize matters and consider only the root node and let it stand in for the whole concept. Such an idealization, of course, ignores the context-sensitivity of concepts and the cognitively relevant content or phenomena that led to the proposition that concepts have some internal structure and are not merely atomic symbols.

Secondly, the proposed model of concepts is compatible with (or close to) a range of recent accounts of concepts and can be seen as an underpinning computational model for them. Let me very briefly point to some examples. For instance, the model is compatible with the "improved" LOT\textsuperscript{H} account by Schneider (2011). Schneider claims that Fodor's LOT\textsuperscript{H} is underdeveloped with regard to the notion of a mental symbol. She proposes that a mental symbol's identity is determined by its total computational role. In the view of concepts presented here, the total computational role of a concept is encoded in the way in which its root node is embedded in the structure of the entire hierarchical network, specifically how it is connected with other nodes and how context-sensitive co-activation patterns with feature nodes arise. Furthermore, my account is close to the perceptual symbol accounts of Prinz and Barsalou but spells out more details and provides an additional twist. For example, in Prinz's account, "concepts are proxytypes, where proxytypes are perceptually derived representations that can be recruited by working memory to represent a category" (2002, p.149). But it seems unclear what conveys stability (or identity) to a concept if each tokening of a concept can be different. The root node of the concept (concept unit) in the model proposed here plays such a stable referential role. The flexibility and context-sensitivity of concepts demanded by Prinz is preserved by the feature selection mechanism based on precision weighting in the PP framework.

Tillas & Treford (2015:7) propose an account for concept individualization close to Prinz's account, but which differs in that the individuation takes place "by virtue of a representational core," which is an "abstracted representation that shares enough
similarities with all members of a given category". The root node of the concept in the model I have suggested could be considered to be such a representational core. Tillas & Treford are mainly concerned with the question of how we can "share" concepts given the vastly different individual concept acquisition histories and the significant context-dependence of concepts. They think that the common core plays a key role here because it "secures reference, which in turn provides the ground for communication". However, the questions of how reference works, and how we can "share" concepts (or a language) require much more discussion and cannot be discussed here.

While the suggested PP model tries to address the concerns of both modalist and amodalist, apparently, it is an account with much sympathy for modalism. The PP model could accommodate amodal representations built into the hierarchical structure. However, it seems more natural and parsimonious to say that the amodal appearance of conceptual representations arises as an asymptotic case (namely for shallow processing) out of a predominantly modal view. The overall PP prediction model of an individual is the result of a constant adjustment with top-down and bottom-up influence for global, long-term and average prediction error minimization. The purpose of cognition is to contribute to successful interaction with the world. This implies that all representations tend to be influenced by the sensory bottom-up flow. If a concept does not help in the error minimization mission, it will be over-written sooner or later (or the individual will lose survival fitness), so ultimately it owes its existence to sensory influences. Even if genuinely amodal concepts existed and were inborn (as Fodor famously held), evolutionary pressure would have ensured that only those amodal representations remain in the evolutionary endowment that contribute to dealing with the sensory inflow and world interaction optimally. All concepts tend to be "grounded" in this broad sense in sensory input.

4.7. Conclusion

In this chapter, I have suggested that we should overcome the dichotomic distinction between modal and amodal representational formats, because of two significant problems it faces: firstly, there is no shared understanding of what modal and amodal formats are; and secondly, both views can accommodate the available empirical evidence. Hybrid accounts, as they currently stand, do not seem to provide
a fully satisfying solution either. I have tried to show how we could reconceptualize
the amodal/modal distinction as a graded one, using a specific cognitive-
computational model of concepts (within the predictive processing framework) as an
example. In this model, a concept is a distributed multi-level network of concept
units. A specific tokening of a concept can include, context-dependently, nodes from
all across the hierarchy, from peripheral sensorimotor areas to the highest cortical
levels. Typical amodality is an idealization instantiated by a shallow mode of
concept processing (lowest grain of prediction in PP terms). In this case, concept
instantiation is limited to the root nodes, and no other lower-level feature-nodes are
co-activated. Typical modality, in turn, arises when we process the concept in a deep
mode, also involving lower levels of sensorimotor representations (highest grain of
prediction). In sum, in this view, there are no separate modal and amodal systems or
representational structures in the brain; modality and amodality correspond to
limiting cases of the (context-sensitive) processing depth in a distributed, hierarchical
concept network.
Chapter 5. A hybrid account of concepts within the predictive processing paradigm

Abstract

We seem to learn and use concepts in a variety of heterogenous "formats", including exemplars, prototypes, and theories. Different strategies have been proposed to account for this diversity. Hybridists consider instances in different formats to be instances of a single concept. Pluralists think that each instance in a different format is a different concept. Eliminativists deny that the different instances in different formats pertain to a scientifically fruitful kind and recommend eliminating the notion of a "concept" entirely. In recent years, hybridism has received the most attention and support. However, we are still lacking a cognitive-computational model for concept representation and processing that would underpin hybridism. The aim of this chapter is to advance the understanding of concepts by grounding hybridism in a neuroscientific model within the predictive processing framework. In the suggested view, the different formats are not distinct parts of a concept but arise from different ways of processing a functionally unified representational structure.

Keywords: concept, concept eliminativism, concept pluralism, concept hybridism, predictive processing, coactivation package account of concepts

5.1. Introduction

We seem to learn and process concepts in different and heterogenous “formats”, like exemplars (e.g., Medin & Schaffer, 1978; Nosofsky, 1986), prototypes (e.g., Hampton, 2006; Posner & Keele, 1968; Rosch, 1978) and theories (e.g., Gopnik & Wellman, 2012; Keil, 1989; Murphy & Medin, 1985). Exemplar theory holds that concepts are represented as a set of exemplars stored under a category label.

____________________

43 I take concepts to be certain bodies of information (see Machery, 2009) that are used in many higher cognitive tasks, i.e., abilities like categorization, inductive and deductive reasoning, planning or analogy making. The focus here is on the psychological notion of concepts (see Machery, 2009 and Löhr, 2020), which is concerned with their cognitive-computational significance.

44 I use the term "format" as a placeholder for whatever prototypes, exemplars and theories turn out to be (representational structures, types of knowledge, ways of processing, etc.). Thanks to an anonymous reviewer for suggesting this way of using the term. Also note that "format" is sometimes used in connection with concepts to distinguish amodal and modality-specific representations. This is not the way I use the term here.
Prototypes are abstracted summary representations, for instance, in the form of a list of features with typicality ratings. And theory-theory describes concepts as embedded in theory-like structures or as little theories themselves. Other formats are sometimes hypothesized: for instance, definitions (a set of necessary and sufficient characteristics), scripts (procedural knowledge) or ideals (a description of an ideal member of a category). However, exemplars, prototypes and theories are the formats that are generally accepted; for this reason, here I will focus on those three.

Those formats were posited to account for a large range of empirical, mostly behavioral, data related to conceptual development and conceptual tasks (some of which I will discuss later). But none of the aforementioned accounts turns out to be able to accommodate the wealth of empirical data (e.g., Kruschke, 2005, p. 188, 190; Machery, 2009). Therefore, format variety is now generally recognized as an unavoidable conclusion (e.g., Bloch-Mullins, 2018; Hampton, 2015; Voorspoels et al., 2011) and has been discussed in depth by Machery (2009).

This heterogeneity of formats sparked many early hybrid proposals, most of them combining two formats (e.g., Osherson & Smith, 1981; Nosofsky et al., 1994; Erickson & Kruschke, 1998; Anderson & Betz, 2001). Given the limited scope and other defects of those initial hybrids, Machery (2009) concluded that each format corresponds to a different fundamental type, and we should dispose of the notion of a concept because the formats have nothing scientifically interesting in common.

Notwithstanding this, many researchers find eliminativism implausible and have continued to propose hybrid solutions in defence of the notion of a concept (e.g., Bloch-Mullins, 2018; Keil, 2010; Margolis & Laurence, 1999, 2010; Rice, 2016; Vicente & Martínez Manrique, 2016), searched for unity behind the diversity of concept formats (e.g., Danks, 2014) or endorsed conceptual pluralism (e.g., Weiskopf, 2009, Piccinini & Scott, 2006).

Arguably, hybridism is the approach that has received most attention and support in recent years. Therefore, here I will leave pluralism and eliminativism aside and focus only on hybrid accounts. My overall goal is not to defend hybrid approaches. Rather I want to provide a novel way to spell out a hybrid account in the spirit of Vicente & Martínez Manrique's "coactivation package" account (2016). Vicente & Martínez Manrique (V&MM) have forcefully argued that hybrids that do not consider
"functional integration" of the formats are hopelessly flawed. While I endorse this view, I nevertheless argue that their approach deserves further development and improvements.

I do not develop a full theory of concepts here. Rather, I focus on the aspect of how a concept needs to be structured as a representational device so that it can serve the roles that the different formats (exemplars, prototypes, and theories) are supposed to play in conceptual cognition. A full theory of concepts would need to address a host of additional desiderata, for instance, how concepts compose to more complex concepts, how they can be shared among members of a language community, etc. (see, e.g., Prinz 2002).

The rest of the chapter is structured as follows. In Section 5.2, I discuss hybrid accounts and examine in some more detail Vicente & Martínez Manrique's "coactivation package" hybrid proposal. I identify two aspects that need further development. In Section 5.3, I introduce a model of concepts that is emerging from neuroscience. In Section 5.4, I introduce predictive processing (PP), a cognitive computational framework, and show how the concept model from Section 5.3 can be embedded in it. In Section 5.5, I suggest how the different formats of concepts might arise and how this approach improves the "coactivation package" account.

5.2. Hybrid accounts of concepts

I focus on Vicente & Martínez Manrique (2016) (V&MM) which is one of the most recent hybrids45. Their account, which I call a "functional hybrid", is a reaction to previously dominating "mereological hybrids". To better appreciate the strengths and weaknesses of V&MM's account, and motivate needed improvements, let me set the stage by briefly discussing mereological hybrids.

5.2.1. Mereological hybrids

Mereological hybrids treat instances of concepts in different formats as numerically distinct entities that are combined to create a hybrid entity. For most such hybrids, their proponents do not emphasize and provide principles for a deeper functional

45 Another account that could be considered a "functional hybrid", in the sense defined here, is Bloch-Mullins (2018), which I will briefly discuss in Section 5.3.
integration of the parts. This is not to say that mereological hybrids do not provide some integrating principle, of course, but the characterization of how and why the components are integrated is rather minimal and "thin." That, however, makes them vulnerable to various anti-hybrid arguments put forward by eliminativists and pluralists (see, e.g., Vicente & Martínez Manrique, 2016, for a discussion). In a nutshell, mereological hybrids are at pains to explain what keeps the components together, beyond some minimal description, and hence what justifies calling the cluster of formats a concept. Furthermore, it is unclear what explanatory advantage hybridism would have over pluralism and eliminativism. Secondly, mereological hybrids cannot say much about what formats are possible, how they hang together and interact, and how they are acquired. They do not seek to reveal an underlying principle from which different formats might naturally arise. Hence, they have an ad-hoc air and lack deeper unity.

As an example, in Margolis & Laurence's (2010) account the different formats are "bound to the same mental symbol". However, no constraints are provided for what formats can be bound to a symbol. Also, nothing is said about how exactly the formats are represented and processed, in particular how different formats are selected on some use occasion. Rice's "pluralist hybrid" (2016) is a further instance of a mereological hybrid. In his proposal, we store information in different formats in long term memory. Information chunks in different formats are retrieved and combined dynamically to create a concept, which is then processed, depending on the task, context, and category. Each combination of different formats corresponds to a different concept. This proposal has the advantage that it does justice to the highly dynamic and flexible processes in concept retrieval. But Rice does not provide constraints for what kind of formats are possible. He also does not explain how those formats are represented and how the selection and assembly mechanisms work.

5.2.2. V&MM's hybrid account

I now discuss how V&MM respond to the problems that afflict the mereological hybrid accounts. I argue that while their response focuses on, and advances in terms of a solution to the first problem, they still face issues, including the second problem of mereological hybrids just discussed.
V&MM suggest that *functional integration* is what holds the different formats of a concept together. Contrary to the above-mentioned mereological hybrids, V&MM put the issue of the *functional integration* into the spotlight. For this reason, I suggest calling their approach a "functional hybrid." Their proposal is then that the unity of a hybrid rests on the "functional stable coactivation" of the formats:

> In a nutshell, the idea is that different structures can be regarded as constituting a common representation when they are activated concurrently, in a way that is functionally significant for the task at hand, and in patterns that remain substantially stable along different tasks related to the same category. (Vicente & Martínez Manrique, 2016, p.61)

A concept is, roughly, a "coactivation package" that makes information of different formats available. Different formats are different parts of the concept that are context-sensitively selected:

> Depending on the task at hand, and on background factors, one part or another of this complex structure receives more activation and plays the leading functional role. Taken separately, prototypes, theories, and so on may be not concepts, but they are *components of concepts*. (Vicente & Martínez Manrique, 2016, p.72, emphasis added)

Note that the authors still speak of formats as "components of concepts". But they use "component" in a rather loose sense, not necessarily implying that formats are strictly "separate modules" (p.73).

I agree with the idea that formats should be integrated in such a way that for a given use of a concept the different formats should simultaneously play some functional role. Only some form of functional interdependence guarantees integration. And without integration it is difficult to see why we need hybrids rather than formats as standalone entities, as pluralists and eliminativists claim. Functional integration makes the hybrid resistant to the above-mentioned anti-hybrid arguments, moreover, it undermines eliminativism, because a functionally integrated unit certainly is a scientifically interesting kind that gives rise to generalizations.

However, I see two issues with V&MM's account.
First, what exactly is "functional significance"? V&MM have not spelled out in detail what this notion amounts to. They only provide a minimal characterization:

The idea behind the functionality condition is that only representational components that make a positive contribution to select the appropriate tokening of the concept count as part of such a concept. (p.69, emphasis added)

According to V&MM, the concept components are "functional" in so far as they make a "positive contribution" to the selection of the "appropriate tokening of the concept". I assume here that V&MM mean that "appropriate tokening" involves two elements. Firstly, the "correct" concept should be selected (e.g., DOG instead of HORSE) and, secondly, it should be tokened in an appropriate format (each concept can be tokened in different ways by selecting different "representational components", which I understand correspond to different formats). The interesting question then is: what does this contribution consist of exactly? An answer to this question crucially requires an account of how the context-sensitive selection of formats works, which is not provided by V&MM.

A second issue with the coactivation package account is that it provides no constraints for possible formats. Should we include, for instance, ideals, scripts, and definitions in the coactivation package? The account is simply silent on this question. Formats are given and then merely added to the coactivation package as a range of possible formats. While V&MM strongly emphasize functional integration, without further details about what exactly this consists in and without further constraints on admissible formats, their account risks remaining a programmatic desideratum about functional integration.

I suggest that we can further develop and improve V&MM’s account by adding a level of description from below, i.e., by being more specific about aspects of neural-level implementation. Rather than starting with a set of independently given formats, we should start from a general neurocognitive architecture that is motivated independently of the question of format variety. From this we can then derive the formats.

As such a general neurocognitive framework, I will use predictive processing (PP). But before describing it in Section 5.4, I will first provide a sketch of a current neuroscientific picture of how concepts might be represented in the brain.
5.3. A neuroscientific model of concepts

The hybrid account I propose builds on a model of the neural realization of conceptual representations that, so I suggest, crystallizes out of an increasing body of current empirical and theoretical neuroscience. This model can be articulated in the form of three core claims.

C1. Conceptual representations are realized as extended networks of nodes: A conceptual representation is neurally realized as the activation of a set of neuron assemblies (nodes) in the form of a distributed network that can cover different brain areas, from higher cortical areas down to lower-level sensorimotor ones.

C2. Concepts are hierarchically organized networks: Different subassemblies (nodes) of the network structure of a concept represent information with different degrees of abstraction/schematicity. The network forms a hierarchy of nodes with an abstraction gradient. Very roughly, higher layers of nodes are sensitive to lower-level node patterns, or in other words, they compress lower-level information. The lowest level in the hierarchy corresponds to the sensory periphery, where representations are maximally modality specific. As we go higher in the hierarchy, information represented by the nodes gets not only increasingly abstracted/compressed, but also convolved, i.e., different modalities (visual, acoustic, proprioceptive, affective, etc.) get mixed (see also Eliasmith, 2013).

C3. Context-sensitive and flexible conceptual processing: On different occasions different parts of the network of a concept are activated in a task- and context-sensitive manner. The tokening of the same concept on different occasions can reach into lower levels of the hierarchy to different degrees.

C1 and C3 closely follow the view of the neural realization of concepts suggested by Kiefer & Pulvermüller (2012). They characterize concepts as "flexible, distributed representations comprised of modality-specific conceptual features". Furthermore, with regard to C2, it is well established that the brain is hierarchically organized; neural layers and areas correspond to different levels of abstraction/compression (e.g., Raut et al., 2020: Hilgetag & Goulas, 2020). This suggests that the extended network structure reaching from higher cortical levels down to sensorimotor areas plausibly has an abstraction/compression gradient.
Kuhnke, Kiefer & Hartwigsen (2021) have put forward a model and empirical evidence that characterizes the hierarchical structure in more detail by mapping the different hierarchy levels on specific brain regions. Lower-level monomodal representations are compressed in layers in so-called unimodal convergence zones. Those feed into layers in multimodal convergence zones. The highest level is an amodal\textsuperscript{46} layer that compresses multimodal input. We have here a double gradient in the hierarchy. On the one hand, the higher the level, the more abstract and compressed the information is. Secondly, in multimodal convergence zones we have a mixing (or convolution) of different modalities. That is, neuron assemblies are sensitive to patterns that involve various modalities. The different layers can be identified with different brain areas (e.g., being the "amodal" layer the ATL). Kuhnke et al. (2021) also show that the connectivity between the layers is strongly task-dependent (claim C3).

C1, C2 and C3 are closely interrelated and empirical evidence for them is increasing. Modality-specific (action, visual, gustatory, olfactory, sound, but also interoceptive) representations often activate complex extended neural networks including modality-specific lower-level brain areas (e.g., Hoenig et al., 2008; see also the overview by Harpaintner et al. 2018). What is debated however, is whether a concept includes sensorimotor areas each time it is tokened, and whether abstract concepts like DEMOCRACY or FREEDOM also include sensorimotor information.

It is safe to say that lower-level sensorimotor areas are not necessarily activated on each occasion even for concrete concepts (Barsalou, 2016; Kemmerer, 2015; Pecher, 2018). Van Dam, van Dijk, Bekkering and Rueschemeyer (2012) argue for the flexibility and context-dependency of the activation of lower-level modality-specific areas in the case of lexical concepts. Yee & Thompson-Schill (2016) conclude that concepts are highly fluid and activations depend on the context, including the individual short and long-term experience.

With regard to abstract concepts, studies show that their activation can also include lower-level sensorimotor areas (e.g., Harpaintner et al., 2020), including

\textsuperscript{46} The authors call the highest level in the hierarchy "amodal". However, it seems also appropriate to call it "multimodal", given that in that layer we abstract across a maximally broad range of modalities, so it is just one more step in the abstraction/convolution hierarchy, not a qualitatively different step (see also Michel, 2020b).
interoceptive and areas processing emotions. Harpaintner et al. (2018) highlight the "importance of linguistic, social, introspective and affective experiential information for the representation of abstract concepts." Such modality specific features can be context and task-dependently activated (e.g., Harpaintner 2020). Furthermore, various researchers suggest that abstract concepts are grounded in emotions (e.g., Vigliocco et al., 2014; Lenci et al., 2018), supporting the idea that their neural realizations also potentially extend into sensorimotor and affective47 areas. All of this is evidence that all concepts might have the same fundamental structure. Also, it is evidence for the claim that concepts are sensorimotor grounded in the sense that they are hierarchical networks of nodes that bottom out at the sensorimotor periphery.

It is important to stress that the neuroscientific model of concepts I have articulated here mainly covers the structure of the realization of concepts (C1 and C2), but little research is available about the specific dynamics of the context sensitive activation patterns postulated by C3. Specifically, an account of how the different formats of concepts arise is lacking. In other words, from the available neuroscientific work we cannot yet derive a full neuro-mechanistic account of dynamic concept processing and the format heterogeneity. This is where the predictive processing framework comes in.

My strategy going forward is to embed the flexible, layered network model of concepts in the predictive processing (PP) framework which I will introduce in the next section. I argue that PP can take on board the three core principles of the model and, more importantly, it can bring the wealth of individual findings under a single comprehensive neuro-mechanistic framework. What PP can then bring uniquely to the table is a model of how concepts are processed. This will be central for my proposal that different formats arise from different ways of processing the network structure that realizes a concept.

47 Sensory areas are meant to include both exteroceptive and interoceptive modalities.
5.4. Concepts within the predictive processing framework

In this section I briefly introduce the predictive processing (PP) framework and suggest how the model of the neural realization of concepts just described could be embedded in it.48

5.4.1. The predictive processing paradigm

Predictive processing (or coding) (see Clark, 2013, 2016; Hohwy, 2013; Friston, 2010; Sprevak, 2021a-d) provides a neuroscientific framework or paradigm for how the brain works from a cognitive-computational perspective. PP is an ambitious framework as it aims at providing a general and unified view on cognitive agency, i.e., an account of perception, action, and cognition. It should be stressed that PP is far from being a mature and worked out theory (Sprevak, 2021a, Walsh et al., 2020). However, it is a very popular framework in cognitive science. In recent years, its scope of applications has been extended and is now ranging from low-level sensorimotor phenomena to several psychological phenomena and even consciousness (Hohwy, 2020).

As a paradigm, PP provides guidance and constrains for the development of more specific theories of cognitive phenomena; PP can be seen as a research program based on some programmatic commitments that are generally but not unanimously accepted by the PP community. In the following part I try to synthesize what I consider to be the core commitments that are most relevant for the purpose of this chapter. Most if not all commitments taken in isolation are neither original nor unique to PP (see Sprevak, 2021a) and it is rather the combination and integration of the commitments that characterizes PP.

a) Prediction error minimization of sensory input

In very general terms, PP pictures the brain as an anticipation and expectation organ that constantly fine-tunes a mental model to continually predict its sensory input.

For instance, perception is not passive bottom-up feature aggregation and pattern recognition, as traditionally conceived (e.g., Marr, 1982; Hubel & Wiesel, 1959).

48 Let me stress that I don't aim here at defending the PP framework, therefore I will not put forward arguments or evidence for it. For that I refer to the mentioned literature.
Rather, the brain constantly generates hypotheses of its sensorimotor states (including all extero- and interoceptive modalities) and corrects the model in the case of errors, so next time it does a better prediction job. In a way, the brain constantly hallucinates in a manner that happens (normally) to match reality.

b) The mental model: generative, hierarchical, and probabilistic

Predictions are being generated by a mental model that is generative, hierarchical, and probabilistic. The attribute *generative* captures the already mentioned idea that the model serves to generate hypotheses constantly and proactively about sensorimotor states.

The model is *hierarchical* because the predictions are being done through representations on many different levels of abstraction/compression (e.g., Clark, 2013). In other words, representations, and hence knowledge, are structured in a hierarchy with an abstraction gradient. Higher levels contain representations that are responsive to larger "receptive fields", i.e., they capture more abstract and coarse-grained patterns represented in lower levels. For instance, while on a very low-level pixels in the retina are represented (which change heavily), higher levels contain representations corresponding to concepts like *APPLE*, which abstract over many instances of specific apples (and hence are more stable). In the downward flow of information, the predictions of higher-level layers play the role of priors for the lower-level predictions and, in this way, constrain the predictions on lower levels. Predictions are being carried out all the time and on all levels of the model at the same time.

The model is *probabilistic* because it represents probability distributions about (sub-personal) "hypotheses" about the causes of sensory input. Furthermore, prediction error minimization approximates Bayesian inference as its primary computational mechanism (e.g., Clark, 2013, pp.188–189; Hohwy, 2013, pp.15–39).

c) Precision weighting mechanism

The PP system contains a so-called "precision-weighting mechanism" of prediction errors (Clark, 2016, pp.53–83). Such a mechanism is necessary as the brain must

---

49 We will later see that it would be more accurate to say here that higher levels contain the *root nodes* of the representational structure corresponding to concepts.
predict the reliability of its sensory input (or more generally the inputs of lower levels in the hierarchy) to distinguish noise and useful signals. In this way, useless modifications of the model due to noisy signals can be avoided. Weights are assigned to the error signals, which allows the system to control the influence of top-down predictions versus bottom-up driven updates of the model. This modulatory mechanism is implemented as part of the overall PP prediction model as (second order) "knowledge" about the reliability and relevance of features in each context (see Michel, 2020a).

d) Neural architecture

PP also makes some general claims about neural implementation. The smallest unit in the model is a combination of an "error unit" and a "representation unit" which I will call a "prediction unit" or simply a "node". Prediction units or nodes are realized as small neural assemblies or "canonical circuits" (see Kanai et al., 2015, also Bastos et al., 2012; Keller & Mrsic-Flogel, 2018; Weilnhammer et al., 2018). The error unit is connected to prediction units on higher levels and the representation unit is connected downwards. Furthermore, there are modulatory inputs into the error units that allow the above-mentioned precision weighting mechanism to tune the error signal.

This brief sketch of the PP paradigm which emphasizes the elements that will play a role in the rest of the chapter, should suffice. In the next section I show how the neural model of concepts from Section 5.3 can be embedded in the PP framework.

5.4.2. PP and concepts

My proposal for how concepts manifest themselves in different formats relies on Michel (2020a, 2020b) who suggests that concepts are implemented in PP by the prediction units just described. Specifically, a given concept is instantiated by a prediction unit, taken as the root node of an extended tree of other prediction units.

The idea then is that the activation of a concept's root node makes available a body of information, namely the subnetwork depending on that root-node. This subnetwork

50 My brief exposition of PP is far from complete, and I have omitted many features, e.g., active inference, efficient coding, etc. Virtually every paper related to predictive processing contains introductions to the framework. I can recommend, e.g., Wiese (2017b), Williams (2020) and Sprevak (2021a, 2021b), for a more detailed overview.
can be seen to correspond to Vicente & Martínez Manrique's "coactivation package". When a concept unit is activated, it makes available a subnetwork that can cover various brain regions, potentially including higher cortical down to primary sensory or motor areas. Critically, which other sub-nodes apart from the root-node itself, are selected is regulated by a context-sensitive modulation mechanism (see Michel 2020a). The basic idea is that higher order knowledge about the reliability and relevance of the different nodes is also encoded in the world model. This higher order knowledge then regulates how the prediction error signals are modulated (i.e., more or less suppressed). Such a mechanism is equivalent to a mechanism that can switch on and off certain parts or nodes of the network depending on the context.

There are concept root-nodes that correspond to patterns on all levels of complexity and spatial and temporal scales. There are, hence, concept root-nodes that range from simple sensory-based expectations, like RED, passing through intermediate-level ones like FACE, to abstract concepts like DEMOCRACY, up to complex situation representations that we grasp in some gestalt-fashion. Such concept root-nodes do not necessarily correspond to lexicalized concepts but also include a host of sub-conscious ineffable ("sub-symbolic") representations that are used as prediction vehicles.

This view of concepts within the PP framework can be put in correspondence with the neural account of concepts as dynamic networks from Section 5.3 in the following way:

C1: The extended network of a given concept corresponds to the sub-network in the PP model that consists of the concept root node and all of its child nodes. (Note that each child node is itself a concept root node).

C2: The sub-network corresponding to a concept is organized hierarchically and has an abstraction gradient in the PP model, exactly like in the neuroscientific model.

Regarding C3, we said that neuroscientific evidence suggests that the concept networks are flexibly and context-dependently activated. According to the PP model the depth with which a concept's tree is activated is flexible, namely task and context-sensitive, driven by the error signal weighting mechanism. Lower-level features can be suppressed by the error weighting mechanism when they are estimated to be unreliable or irrelevant. Activation of a concept can be "shallow"
(e.g., a "schematic apple" in which no specific colour is co-activated), in which case only higher-level nodes are activated. Or activations can be "deep", which involves, e.g., a more vivid (modality-specific) mental representation due to co-activating nodes that are located lower in the hierarchy (a mental picture of an apple, with a specific colour, form, size, etc.).

The existence and flexibility of concepts can be motivated within the PP framework in a principled way (see Michel 2020a). Concepts are necessary vehicles for prediction making; it is in virtue of prediction units that predictions are made. An efficient prediction economy requires making predictions with an adequate level of detail. When you want to cross a street successfully, your brain's predictions cannot and need not happen on the situation's pixel-level of precision. Rather the predictions need to be more schematic and have a coarser grain of description.

There are two ways to regulate prediction detail. The first is by using prediction units at higher levels in the hierarchy. The higher the nodes, the more schematic and compressed (hence less detailed) their content. The second is by co-activating a varying number of other nodes; those represent more detailed and concrete features of that conceptual representation.

In conclusion, by embedding the neuroscientific model of concepts from Section 5.3 in the PP framework, we get a more comprehensive model of concept representation and processing. As we have seen, PP can provide an implementational-level proposal of the network structure (a network of PP prediction units with an abstraction gradient). But what PP can crucially contribute is the processing aspect, which is still underdeveloped in the literature. For instance, PP supplies a self-organizing driving force operative in the node network (prediction error minimization), as well as a mechanism for feature selection (based on the precision weighting mechanism). Furthermore, PP motivates the existence of concepts as prediction vehicles, and the need for the right level of granularity, which in turn motivates the existence of the feature selection mechanism.

5.5. The manifestation of different concept formats

With a cognitive-computational account of the structure of conceptual representations in place, I will now show that the different formats correspond to how
the network of a concept is being context-sensitively processed. The different formats mirror not numerically distinct representational entities, but the processing depth and width of the concept's (and surrounding) network structure. More precisely, exemplar effects correspond to relatively deep vertical downward processing (i.e., towards less abstract nodes), prototype effects to relatively shallower vertical downward processing, and theory effects to additional vertical upwards and horizontal processing (i.e., towards parent and neighbour nodes).

5.5.1. Exemplars and prototypes

In this subsection I argue that a concept can manifest itself in "exemplar mode" and "prototype mode" when the node tree associated with the concept is processed from more to less abstract nodes (vertically downwards processing). Processing only higher-level nodes corresponds to prototypes. Processing in addition lower-level nodes corresponds to exemplars. I will first unpack this proposal by explaining how exactly to understand exemplars and prototypes and how they are realized in the PP model. Then I will provide some examples of how we can account for the exemplar and prototype effects that motivated those formats in the first place.

5.5.1.1. What exactly are exemplars and prototypes?

In the standard story of exemplar theory, which aims to address exemplar effects, my concept DOG consists of the memorized collection of representations of specific dogs. They are modality-wise specific as they correspond to instances of dogs. Categorizing some animal as a dog implies using dog exemplar(s) and calculating similarities. Note that the exemplars might have very different levels of specificity, i.e., levels of modality-specific detail or vividity. Sometimes we remember object-exemplars only vaguely with little detail, and sometimes very concretely with a lot of detail.

In the standard story of prototype theory, which aims to address prototype effects, my concept DOG consists of some representation of a typical dog. The representation is more abstract compared to an exemplar. Categorizing some animal as a dog under prototype theory, implies using the dog prototype and calculating the similarity.
Note that the processing, for instance in categorization tasks, of both exemplars and prototypes rely essentially on similarity calculations, primarily over relatively superficial features.

Some researchers think that exemplars and prototypes are the ends of a continuum rather than two distinct kinds (e.g., Vanpaemel et al., 2005, or Verbeemen et al., 2007). Authors like Barsalou (1990) and Hampton (2003) think that prototypes and exemplars differ only to the extent to which exemplar information is retained or abstracted away. Smith & Medin (1999, p.209) characterize exemplars in terms of a relative lack of abstraction. Exemplars can be maximally specific object-particulars but are not necessarily; they can also be subsets. For instance, PLANET is a subset of HEAVENLY BODY, and hence an exemplar for it.

Following those authors, I assume that there is no fundamental difference between exemplars and prototypes in terms of the deeper, underlying representational structure in the first place. In both cases the general structure consists of a set of pairs of features and values. Those features might have different degrees of specificity/schematicity.

5.5.1.2. Prototypes and exemplars in the PP model

The posited structure of a concept as a hierarchical node tree allows us to account for the exemplar and prototype formats. Concept processing in exemplar mode can be cashed out as the processing of the concept's node tree with attention towards relatively more specific information (without necessarily being maximally modally specific), while processing in prototype mode can be cashed out as more shallow processing, i.e., involving nodes with relatively less specific information. In both cases we have more or less deep "vertical downwards" processing of more superficial features. Those features are included in the node tree that origins in the concept's root node.

In PP terms, processing a concept in exemplar mode is processing towards lower-level (i.e., modally more specific) nodes. The tokening of the concept DOG in exemplar mode reaches from the conceptual root node [DOG] down to at least a subordinate node and potentially (but not necessarily) further to lower-level nodes down to the sensorimotor periphery. To conceive of a specific dog, e.g., Hasso, as a dog, implies the activation of the abstract [DOG] node and the subordinated [HASSO]
node and other subordinate nodes, potentially down to specific shapes, colours, odours, etc. So, a whole node sub-tree from \textsc{dog} might be activated.

To categorize a specific dog exemplar, say Hasso, a hypothesis needs to be generated that matches as well as possible whatever sensory input I receive. If my dog Fido is very similar to Hasso, a salient hypothesis is of course that Fido actually is Hasso. So, the hypothesis that reproduces a memory of Hasso fits well with the bottom-up Fido input, i.e., it produces a small prediction error in relation to other hypotheses.

Categorization might also happen via a prototype of \textsc{dog}. If you cannot see Fido well (because he moves quickly and is far away and could be a cat as well) but hear loud barks, given that the feature of barking is strongly cue valid (i.e., the probability that something that barks is a dog is high), there is no need (and it would not be very economic) to recur to more specific exemplar information. The barking can be immediately explained by the hypothesis \textsc{dog} and Fido categorized as a dog.

It is important to stress that, in the proposed view, what is an exemplar and what is a prototype is \textit{task-dependent}. It might happen that in a task a prototype of some concept is represented with more detail than an exemplar of that concept in another task. Consider the following example:51

1) Suppose that a Bach scholar is played a piece of music and asked whether it is typical of Bach. To answer this question, the scholar may draw upon a very rich mental representation of the typical features of Bach pieces, which encodes very specific information about sensorimotor details such as certain kinds of instrumentation, cadences, melodies, harmonies, ornaments, rhythms and so on.

2) Now suppose that the scholar is asked whether the Brandenburg Concertos are a work by Bach. Plausibly, the scholar could answer this question without drawing on deep, specific, information, close to the sensory periphery.

In task 1), the prototypical representation, say \textsc{bach}_{prototype}, used by the scholar to decide whether the piece he is listening to is typical of Bach might perfectly contain very specific features. The important point is that \textsc{bach}_{prototype} is relatively more

51 I am grateful to an anonymous reviewer for providing various potential counterexamples, including this one.
abstract than the exemplar representation in this task, which is the piece of music, say BACH\textsubscript{exemplar}, that she has to classify. In task 2) we deal with a completely different process, again with two representations, say, BACH-WORKS and BANDENBURG-CONCERTO. The question is whether the latter is an exemplar of the former. Indeed, to answer this, one only needs to know that the Brandenburg Concertos are works by Bach (the former is an instance of the latter category). What is needed is that BACH-WORKS is a relatively more abstract representation than BANDENBURG-CONCERTO, and that is sufficient for the latter to be an exemplar of the former. According to the PP model, this is the case if, for instance, BANDENBURG-CONCERTO is represented as a child node of BACH-CONCERTOS. Here the exemplar BANDENBURG-CONCERTO from task 2) is much less concrete than BACH\textsubscript{prototype} from task 1); but that does not undermine the proposed account. What matters is the relative abstractness of the relevant representations within each task.

Let us turn to the probabilistic element of PP: the nodes making up the PP model represent whatever they represent in terms of probability distributions. Specifically, a node represents a probability distribution over nodes in the next lower level. For instance,\textsuperscript{52} RICHARD II might be represented as an exemplar of MONARCHS-OF-ENGLAND because the probability distribution over monarchs encoded in MONARCHS-OF-ENGLAND has at a given moment a sharp spike at the child node RICHARD II. Being an exemplar does not imply, however, that all lower-level nodes have sharp distributions. For instance, my probability distribution over the hair colour feature of Richard II must be very spread-out indeed. As already mentioned, often exemplars are quite schematic (as in the Bach example 2). In the case of a prototype representation, the probability distribution is more broadly spread. A typical feature or exemplar is then one with the largest likelihood. For instance, MONARCHS-OF-ENGLAND might encode a probability distribution over features such that a typical monarch is one who has the most likely features, i.e., those features with the highest probabilities.

Note that in the PP view, there is no explicit "calculation" of similarity formulas, which is central to categorization in exemplar and prototype theories (see, e.g., Machery, 2009 for examples of formulas). Rather, similarity is implicit in the fundamental

\textsuperscript{52} Thanks to an anonymous reviewer for the example, which helped me to make the point clearer.
mechanism of the PP model, namely, weighted prediction error minimization. In weighted prediction error minimization, the top-down prediction and the bottom-up input at each level are compared, i.e., their "similarity" is determined. This mechanism can model both the more abstract prototype level (by focusing attention on higher level nodes, i.e., dampening lower-level nodes that represent more details) and the exemplar level (i.e., lower-level nodes are more error sensitive).

5.5.1.3. Prototype and exemplar effects

As emphasized already, a theory of concepts aims at accounting for a large body of behavioural effects observed during conceptual tasks.

*Prototypes* have been motivated by "typicality effects" that could not be explained by the previously prevailing definitional theory of concepts, according to which concepts are definitions or necessary and sufficient properties. A typicality effect arises when we judge certain objects to be more typical members of a category than others. For instance, a sparrow—in normal contexts—is judged to be a more typical bird than an ostrich. In the standard story of prototypes theory, the concept of *BIRD* consists of a set of properties and a typicality rating for each property. A sparrow would in normal circumstances be a more typical bird than an ostrich.

*Typicality* can be accounted for in terms of representations based on probability distributions through conditional probabilities as they are posited by PP. For instance, if we know that something is a bird, we expect to a higher degree (in a neutral context) that some instantiation is a sparrow rather than an ostrich. So, a sparrow is a more typical bird than an ostrich. In PP jargon: when you are asked to mention a typical bird, your generative model is more likely to "sample" [*SPARROW*] in the next lower level in the node tree below [*BIRD*] than [*OSTRICH*]. This is expressed as the following relation between two conditional probabilities \( p(OSTRICH \mid BIRD) < p(SPARROW \mid BIRD) \) which are encoded in the PP world model.

The PP model can also provide an account of how *exemplar effects* work. Take, for instance, the *old item advantage effect*: memorized exemplars are more easily categorized than new ones that are equally typical (e.g., Smith & Minda, 1998, 2000). Those effects could be modelled within the PP framework as follows. For sensory input like previously encountered and memorized exemplars, the prediction error is better minimized by using the exemplar rather than a prototype. In the case
of "deep processing" which is characteristic for exemplar processing and where
details matter, the most similar memorized bird exemplar just best "predicts" the
target bird you see in front of you because it causes the least prediction error. The
fact that details matter is cashed out in terms of the higher error sensitivity of lower-
level nodes that represent more specific features. The more specific features,
however, are only considered in the prediction if the brain assigns a high precision
estimate to the prediction errors on the level of those features, i.e., when it considers
details to be relevant and reliable. In the above example, where a person hears a
dog barking in a foggy environment, details will be suppressed due to the lack of
reliability of the sensory input. Therefore, more abstract prototype representations
are used. Barking is a property with high cue validity.

So, according to the PP model, depending on the relevance and reliability of the
details, exemplar or prototype modes of processing arise. Note that those are not
two strictly dichotomic modes, but a graduation along the abstraction gradient exists.
As mentioned, concepts within the PP model serve to modulate the granularity of
predictions. Taking up again the example from Section 5.4.2., it is not efficient when
a street is crossed to predict the exact, maybe pixel-level, details of the event.
Rather the event should be processed on a more aggregated level. For instance, we
do not need to predict the exact shape and colour of the car approaching when we
try to cross the street. It is sufficient to conceptualize the scene in larger grain, e.g.,
that some fast-moving car is approaching. Exemplar and prototype formats are
manifestation of this context dependent granularity modulation (or choice of
abstraction level). Also note that what "format", or more precisely, what level of
abstraction is used in each task might vary across individuals. For instance,
someone who is especially afraid of sports cars when crossing a street might pay
more attention to more detailed features. Maybe someone is especially afraid of a
specific car (maybe because in the past Uncle Tim’s car has almost hit her) and,
therefore, she mobilizes even more detailed exemplar information for prediction
making.

5.5.2. Theories

Now I argue that a concept can manifest itself in "theory mode" when the
surrounding node structure in which the concept is embedded is processed (i.e.,
processing in a vertically upwards and horizontal direction from the concept's root node). I will first unpack this proposal by explaining how exactly to understand the notion of "theory" and how a theory is realized in the PP model. Then I will walk through an example of how we can account for a classical knowledge effect that motivated the theory format in the first place.

5.5.2.1. What is a "theory" in the theory-theory of concepts?

It is important to point out that theory-theory is far from being a monolithic position. Discrepancies (or indeterminacies) exist along various dimensions; let me mention two and make explicit what notion of theory I will assume.

Firstly, there are two ways in which the relation between concepts and theories has been spelled out (see, e.g., Weisskopf, 2011): concepts are constituents of theories or concepts are miniature theories that store relevant theoretical (i.e., causal, functional, taxonomic, etc.) knowledge. In the first case, theories are bodies of beliefs or propositional structures with concepts as constituents. In a strong version of this view (e.g., Carey, 1985) concepts are individuated as the roles they play in those theories. In the second case, concepts are structures that are themselves little theories (e.g., Keil, 1989). However, it is not spelled out in detail what this position exactly amounts to in terms of its representational structure. For instance, when Keil says

> most concepts are partial theories themselves in that they embody explanations of the relations between their constituents, of their origins, and of their relations to other clusters of features. (1989, p.281)

the question arises as to what exactly the embodiment of those items looks like. If those items are articulated as beliefs or propositional structures, how is this then different from the concepts-as-constituents view? Even worse, the view seems then to have the incoherent implication that a concept is both a constituent and a theory of which it is a constituent. So, it is crucial to spell out how the knowledge items are represented. The concept-as-constituents view seems not to have this specific problem because there are two things: some theory and a concept that is a constituent of that theory. In turn, this view does not capture the intuition that a concept indeed seems to be some sort of "information package" including a host of
theoretical information. In any case, we have here an unresolved problematic aspect of theory-theory in general because, as Weisskopf points out (2011), "the empirical evidence taken to support the Theory-Theory does not generally discriminate between them, nor have psychologists always been careful to mark these distinctions."

The advantage of the proposed PP account of concepts is, as I will argue later, that it spells out a specific representational structure that allows to perfectly make sense of the idea that a concept can be seen to be both, a miniature theory and a constituent of a theory.

A second aspect where theory-theories vary is the demand regarding the coherence of the encoded knowledge. Kwong (2006) usefully distinguishes two different notions of theory, a literal and a liberal one. A literal theory is analogous to a scientific theory, and cognitive and conceptual development is equivalent to scientific theory formation and change. Here aspects of causal relationships, coherence, and systematic structure are stressed. An example of a literal understanding of a theory notion is Gopnik & Wellman's (2012) account. According to the authors, a theory is a coherent structure of abstract representations, analogous to scientific theories (2012, p.1086).

On the other hand, in the liberal understanding of theory, as endorsed, for instance, by Murphy & Medin (1985), the knowledge structure is more flexible. When they say that "...we use theory to mean any of a host of mental 'explanations,' rather than a complete, organized, scientific account" (1985, p.426), they allow other, informal types of knowledge structures, i.e., formats, in a theory. Such formats are, for example, empirical generalizations (mere correlations of phenomena) or scripts (procedural knowledge, or a chain of events or acts). Liberal theory theorists put less demand on the coherence of a body of knowledge. A representational knowledge system does not need to exhibit formal consistency and rigor, deductive closure, etc., to count as a theory. Such features might be desirable and are most probably normative; however, they are not plausible as a description of how we cognitively store knowledge.

I will endorse the liberal view of theories relevant for concepts because the strict view seems psychologically implausible (see also Machery, 2009, p.102). The liberal notion of theory is closely related to the notion of "folk theories." A folk theory, or
"intuitive theory" is common-sense knowledge about a specific domain, for instance folk biology or folk psychology (e.g., Gerstenberg & Tenenbaum, 2017). The building of such folk theories is less systematic and conscious than scientific theory building.

5.5.2.2. Theories in the PP model

As we have said before, in the proposed PP model, world knowledge is encoded as a huge network of interconnected prediction units (nodes) on many levels of abstraction/complexity. In the upper levels we have prediction units that represent complex situations, contexts, scenes, relations, patterns, patterns of patterns, etc. The lower levels represent for instance concepts of concrete objects or simple features like colour, etc.

The PP framework quite naturally accommodates theory-like structures, as the generative PP model is standardly interpreted as a multilevel causal model (e.g., Friston, 2010; van Pelt et al., 2016). Nodes that correspond to variables form a probabilistic network. The model is hierarchical, i.e., the nodes at one level, roughly, correspond to latent variables that are the causes from which the variable in the next lower level can be derived. However, limiting the relations between the variables to causal relations makes the model too narrow (see also Sprevak, 2021b). A prediction unit can be more generally interpreted as a prior that constrains the values on lower levels, i.e., nodes and sub-nodes have a more general form of "predictive relation", which can also include part-whole relations or taxonomic relations or object-property relations. The reason is that all of those are "predictive" in the sense that in the same way as causes constrain possible effects, genera constrain possible species, and wholes constrain possible parts.

In theory mode, so I suggest, it is the connectivity of a concept root node with higher level nodes and nodes on similar levels in the total model hierarchy that is being exploited. In other words, the theory mode of concept processing arises from horizontal and vertical upwards processing outside the concept node tree, in addition to vertical downwards processing within the concept node tree below the concept's root node. While exemplar and prototype processing remain within the structure of the subordinate nodes of a concept root node, in theory mode, processing expands upwards to more abstract and laterally into neighbouring concepts units.
One might think that theories are represented in terms of high-level, relatively abstract, human-interpretable, lexicalized concepts. For instance, a certain edge form representation in the brain’s visual processing stream is not a concept in the more traditional and common-sense understanding. Perceptual and conceptual representations are normally seen as qualitatively distinct.

However, authors proposing the existence of "folk theories" (e.g., Gerstenberg & Tenenbaum, 2017) do not assume representations in symbolic and lexicalized form. A folk theory of physics, which allows for guessing whether certain tower constructions are stable, requires complex "sub-symbolic" sensorimotor representations. Similarly, I have emphasized within the proposed PP view the existence of many ineffable, consciously not accessible, and non-lexicalized nodes on many levels of abstraction (see also Lake et al. 2017 for a discussion of sub-personal "theories" that are not lexicalized). Those "sub-symbolic" nodes are continuous with the "symbolic" nodes that correspond to more narrowly understood concepts (e.g., only lexicalized, or lexicalizable53 concepts). All the nodes are "concepts" in virtue of them playing the role of prediction units. They just differ in the degree of abstraction. We could stipulate that only narrowly conceived concepts form theories. But nothing hangs on this rather terminological decision. We can consider theories based on narrow concepts to be "embedded" in the total PP model, which consists of both narrow and inclusively conceived concepts.

5.5.2.3. Accounting for knowledge effects

The classical knowledge effect I want to focus on here as an example is reported by Rips (1989a) in his famous pizza experiment. It provides evidence that sometimes we classify some A to be a B, rather than a C, even if A is more similar to C. Rips asked participants to imagine a circular object of three inches and asked whether it was more similar to a quarter or a pizza. The dominant answer was that it was more similar to a coin (because of its small size). Then the participants were asked whether it is more likely a pizza or a quarter. The dominant answer was that it was more likely a pizza (because quarters have uniform sizes, while pizza sizes

53 A feral child might have the lexicalizable concept of WOLF, though it is not lexicalized. In contrast, all sorts of ineffable edge-patterns and shapes are used, e.g., in lower levels of the visual pathway there are prediction nodes that are not consciously accessible and lexicalizable in any meaningful way.
might vary). Here we do not categorize in terms of similarity but rather based on more extended knowledge, e.g., of the manufacturing process of pizzas and quarters from which we can infer their possible variability in size.

Let us now account for the pizza experiment by the PP model. The concept formats involved—prototypes/exemplars versus theory-like common-sense knowledge—seem to be primed by the task. In the first task the subjects are explicitly being asked to make a similarity judgement while the second task evokes a judgement about the causal chain that brought about each object (pizza versus quarter).

Such causal knowledge is encoded in the PP model as specific experiences but also more abstract generalizations that one might have, which also involve other concepts like PIZZA BAKER, PIZZA OVEN, DOUGH, etc. from experiences with how pizzas are made (see Figure 5.1). Hence the concept PIZZA is being processed by carrying out inferences with concept units outside the information package PIZZA itself. A more abstract node in the PP model might be a concept unit representing a complex schema PIZZA-BAKING_SCHEMA which is a sub-domain of common-sense knowledge about baking represented by BAKING_SCHEMA. PIZZA-BAKING_SCHEMA might have sub-nodes that are part of the knowledge about pizza baking, let us say AGENT-FORMS-DOUGH_SCHEMA and DOUGH-PROCESSED-IN-OVEN_SCHEMA.\(^{54}\) AGENT-FORMS-DOUGH_SCHEMA again contains sub-nodes that contain information about how an agent forms the dough, etc. From that knowledge one can infer that it is easy to make, for instance, a pizza that is smaller than usual, simply by applying the same pizza forming process to a reduced quantity of dough. This reduced quantity is possible as the pizza baker is free to choose the quantity she wishes.

Similarly, QUARTER, might be a node subordinate to a more abstract node corresponding to some frame concept unit, which links QUARTER in such a way as to encode common-sense knowledge about the role and production of coins. From that knowledge one can infer that it is very unlikely that a coin has the size of the target object. The agents intervening in the coin producing process do not normally have the "freedom" to alter the size of a coin ad hoc.

---

\(^{54}\) Here PIZZA-BAKING_SCHEMA could be a concept that encodes a "script", i.e., a sequence of actions.
Taking this way of processing the concept structure, the inference is being made that a pizza can easily have different sizes, while coins do not. Therefore, the target object is more likely to be a pizza.\footnote{Given that the PP approach has commitments on the level of neural implementation, at least in principle, there is an avenue for empirical verification/falsification of the model. Admittedly, the current state of the art in brain imaging techniques does not yet provide a sufficient level of temporal and spatial resolution to map out concepts and neural structures in the required way.}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig5_1.png}
\caption{A schematic toy example of a concept unit network for the concept PIZZA and modes of processing.}
\end{figure}

5.5.2.4. Are concepts theories or constituents of theories?

With this approach of the theory format in hand, we can now briefly revisit the question discussed in Section 5.2.1., namely whether a concept (in its theory format) is a theory or a constituent of a theory. It is easy to see that the dispute now looks merely verbal. A concept can be both. A concept, say APPLE, can appear to be a theory when connected nodes are processed that represent theoretically relevant information (i.e., when it is processed in theory mode). But APPLE can also appear to be a "constituent" of some (other) theory, namely when at least the root-node of APPLE is processed as part of the processing in theory mode of some (other) concept, for instance, FRUIT or NUTRITION.
5.5.3. The functional integration of exemplars/prototypes and theories

One might object that exemplars/prototypes and theories do not seem to have the same status in the concept's information package. There are three properties that prototype and exemplar processing share but that are absent from theory processing. Firstly, prototypes and exemplar processing involve nodes of the sub-network of the concept's root node, while at least some nodes corresponding to theory processing lie outside this sub-network. Secondly, we have also seen that the distinction between exemplars and prototypes is a relative affair, but nothing similar has been said for the theory format. Finally, exemplars and prototypes are closely associated with the notion of similarity, which is (at least not obviously) the case for theoretical knowledge.

Despite those differences, all three formats should be seen as deeply functionally integrated in the form of a prediction device. To better understand why theoretical information is also integrated with exemplar and prototype information of a given concept, note that—from a neuro-anatomical point of view—the main difference is that processing theoretical information involves nodes on a level higher than (or the same level as) the concept's root-node, while prototype/exemplar information involves nodes at a relatively lower-level. In both cases, however, the concept's root node is involved and connected to those nodes, and the general structure and processing principles are the same in the whole hierarchy. The specific connectivity implements a layered structure of conditional probabilistic dependencies among the nodes on different levels. It is this informational dependency dynamics which then integrates the higher and lower-level nodes connected to a given root-node into a functional whole. Let me work this out in further detail.

Remember that a PP model is a generative model with latent variables represented as nodes that "explain" (or "generate", or "sample") features represented by lower-level nodes. While lower-level nodes correspond to concepts that are "explained" by some concept in question, higher-level nodes correspond to concepts that "explain" that lower-level concept. For instance, while APPLE "explains" RED, FRUIT "explains" APPLE in the sense relevant here. In other words, using the terminology of generative models, RED is a sampled (a "generated") feature from the probability distribution over features represented by APPLE. APPLE, in turn, is sampled with a relatively high probability from FRUIT, which is a probability distribution over fruit types.
Plausibly, the body of knowledge associated with some concept includes both information about what it is caused/explained by and what it is a cause/explanation for. In this sense, exemplars/prototypes (with more superficial features) and theoretical features (representing more abstract causal, taxonomic, mereological, etc. relations) form a functionally integrated information package. The difference is only one of explanatory (or "generative") direction.

To bring home my point about the tight functional integration of exemplars/prototypes and theoretical information, it might be useful to refer briefly to Bloch-Mullins’ recent work on concepts (e.g., 2018, 2021). There is no space here for a careful discussion of her account and a detailed comparison, but it is worthwhile pointing to some deeper commonalities, which suggest some substantial common ground.

Bloch-Mullins (e.g., 2018, p.607) observes, quite correctly in my view, that the problem with the different single-format accounts of concepts is not that they are each on their own unable to cover all of the empirical data from concept research. The problem is that they do not even have sufficient explanatory depth with regards to the restricted scope of the phenomena they were designed to cover. For instance, she argues that the similarity judgements involved in exemplar and prototype applications cannot be calculated without theoretical (specifically causal) knowledge about how to pick out the relevant dimensions for comparison (pp.609-614). Theoretical knowledge, in turn, can't be applied in categorization without using similarity judgements to determine the relevant range of values that determine the category of a variable figuring in a causal relation (pp.615-621). Normally, the values of the variables by which those causal relations (used for categorization) are described are not identical, but only sufficiently similar to underwrite classification. A second way in which causal knowledge is relevant in categorization is that the dimensions selected for similarity judgements is relevant in categorization is that the dimensions selected for similarity judgements might also include causal relations (Bloch-Mullins, 2018, pp. 622 and 624; see also Bloch-Mullins 2021, pp.61–62; Hampton, 2006, pp.85–86). I suggest a third way in which similarity intrudes categorization based on causal knowledge: grasping and applying theoretical knowledge is itself recognizing analogies/similarities to abstract (e.g., causal) patterns, i.e., causal knowledge is stored as patterns that demand similarity matching.
I am very sympathetic with Bloch-Mullins' view. In the PP model, the similarity of A and B can be fleshed out as A and B being an instance of (being "sampled from") some concept node. If there is some C that "generates" A and B, then A and B are similar with respect to the features that C encodes. But this idea is transferable to theoretical (i.e., causal, taxonomic, mereological, etc.) features. To see this, let us take one of the examples that motivated the theory format of concepts, namely deep "essences" of living creatures (e.g., Medin & Ortony, 1989; Gelman, 2004). For example, assume that HORSE-A and HORSE-B are representations of horse exemplars in virtue of being sampled by some HORSE-ESSENCE which represents the horse essence that "generates" horses. Our folk-biology might be represented minimally as the knowledge that animals have hidden essences that are responsible for (i.e., cause) the existence of certain animal types. In the PP model, this knowledge is captured by some abstract high-level prediction unit that encodes the very general concept of ANIMAL-ESSENCE as part of some animal folk-theory. There are lower-level child nodes of ANIMAL-ESSENCE that correspond to more specific essences like HORSE-ESSENCE, DOG-ESSENCE, etc. Those in turn sample (or "generate") concrete exemplars of the corresponding species, e.g., Fido (the dog).

The advantage of the PP approach is, as previously pointed out, that similarity calculations are not based on algorithms over an explicit list of features but are the implicit result of holistic prediction error minimization. What is then instantiated as being similar to what depends heavily on the "context" which includes background knowledge, goals, foils under consideration, etc., all of which are represented by other prediction units in the network. PP captures well this highly context dependent dynamics of similarity calculations. Similarity judgements emerge holistically from all of the relevant available information in the PP model.

5.5.4. In which sense does the PP model refine the coactivation hybrid account?

Let us get back to the end of Section 5.2 where I pointed out two possible improvements to the coactivation account: spelling out more concretely what functional integration amounts to and providing constraints for "admissible" formats. Let us revisit each of them in the light of the proposal just developed.

First, there is a more specific notion of functional integration that emerges from the PP model. The whole "coactivation package" of a concept serves as a context-
sensitive prediction device for the category represented by the concept. A coactivation package, we have seen, consists of a root-node and the depending sub-network of lower-level nodes. The root-node is the result of abstraction and convolution of lower-level nodes, therefore in a sense it is closely connected to (i.e., it "contains" information of) all sub-nodes. Those subordinate nodes correspond to exemplar and prototypical information. Furthermore, as this package is integrated into the whole overall model, it has external connections to other lateral and higher-level nodes. Those nodes correspond to more theoretical and abstract knowledge associated with the concept, namely causal, taxonomic, mereological, etc., information that "explains" the concept.

Processing in the PP model is holistic, so all of the nodes are interlocked and have an influence on the overall state of the information package associated with the concept, i.e., on which other nodes are selected, and which are not.

With the PP model, an account of the context sensitive modulation of the subparts of a coactivation package comes for free because it is a core feature of the general PP framework. It can be put to work to select the processing depth and direction that determine the appearance of the concept formats.

Secondly, the PP model provides constraints for possible formats, namely those imposed by the PP architecture. One needs to be able to derive the format from the representational resources provided by PP. We have seen that we can derive the three generally accepted, classical formats: exemplars, prototypes, and theories. An interesting next step—that needs to be carried out elsewhere, however—would be to explore whether other candidate formats like definitions, scripts or ideals could be derived from, or are consistent with, the proposed PP model.

5.6. Conclusion

This chapter has attempted to put forward a cognitive-computational model of hybrid concepts within the predictive processing framework. In the view proposed here, formats are—contrary to most other hybrid accounts—not to be understood as components of a concept. Rather, formats correspond to different directions and depths of processing of the same concept structure.
The model aims to further develop and improve Vicente & Martínez Manrique's hybrid account with regard to two aspects. Firstly, it spells out what "functional integration" of the formats more specifically amounts to. Functional integration is necessary for a genuine hybrid account. Formats are functionally integrated in the PP model because they arise as optimal (i.e., prediction error minimizing) ways of processing a unified representational structure. Critical for the functional integration is the context-sensitive selection of subparts of the structure (which then appear as different formats). Such a format selection mechanism comes for free in the PP model. Secondly, the proposed model provides constraints for possible formats because it supplies more detail about how concepts are represented and processed in the mind, providing more specific computational, algorithmic and implementational level commitments.
Conclusion of Part 2

In the preceding three chapters I have sketched an account of concepts within the PP framework. Here I will summarize it and briefly evaluate it with regards to Prinz's (2002) set of desiderata for theories of concepts, which I expand by some additional ones.

Summary of the PP account of concepts

I have pictured concepts as neural structures that correspond to hierarchically organized node sub-networks of the overall PP world model spanning various brain areas. A node corresponds to an assembly of neuros. Each node encodes information of different degrees of abstraction, i.e., a higher-level node contains (abstracted and compressed) information from lower levels. The concept is stably identified by a root-node of a sub-network. The sub-network of the concept can then be activated flexibly by switching sub-nodes on or off. This allows more abstract instantiations of the concept (e.g., prototypes) but also more concrete ones (e.g., vividly represented exemplars).

As those sub-networks "bottom out" at the sensorimotor periphery (including interoceptive and affective areas), we get a natural way to cash out how "embodiment" is mirrored in the brain. Ultimately all concepts are "grounded" in this way in sensorimotor information. In this view, even abstract concepts are embodied, i.e., grounded ultimately in sensorimotor representations. As I pointed out in Chapter 4, this does not imply strong empiricism. Certain concepts, especially higher-level ones, might well be inborn. However, insofar as they serve to successfully predict sensorimotor input, they are nevertheless tuned to the sensorimotor periphery and are grounded in this way.

A representation is "abstract" or "compressed" in the sense that it responds to more and more complex receptive fields, where a "receptive field" should be understood to extend to all sensory modalities and mixtures of modalities. Therefore, abstraction is naturally built into this model. This allows for addressing the concern that abstract concepts cannot be perceptual (i.e., grounded in sensorimotor information).
The advantage of this model is that it takes on board concerns both from amodalists and modalists. The root node can function as what individuates and give stability to a concept. On the other hand, the structure below the root nodes gives the concept a modality specific grounding and allows for context sensitive variation of the information made available for processing depending on the context.\(^\text{56}\)

*Evaluation with regards to Prinz's desiderata for theories of concepts*

Prinz proposes seven desiderata to evaluate whether a theory of concepts does the job it is supposed to do. How does my proposed account of concepts fair with regards to those desiderata?

Prinz's list includes desiderata for both the psychological and philosophical understanding of concepts. As already mentioned, in this dissertation I mainly focus on the psychological aspects of concepts. Regarding philosophical desiderata, therefore, I will comment only briefly on them.

1. **Scope.** The PP account is straightforwardly able to account for the full range of concepts from primitive perceptual ones to complex and abstract ones like DEMOCRACY or PRIME NUMBER. Note that abstract concepts do not suppose a special problem because they are understood as arising from a neurologically plausible abstraction and multimodal convolution or "mixing" process. Moreover, one might say that the notion of concept under the PP framework has an excessive scope, as I also take sub-symbolic as well as very high-level biases to be "concepts" because of their unified role as prediction vehicles. However, I have already insisted that one can restrict the term "concept" to those prediction units that are traditionally considered to be concepts (e.g., because they are lexicalized and consciously accessible concepts). But I also argued that there is no non-arbitrary criterion to clearly differentiate them.

---

\(^{56}\) When this thesis was largely finished, Guy Dove's book "Abstract concepts and the embodied mind: Rethinking grounded cognition" (Oxford University Press, 2022) was published. He makes a case for a very similar model of concepts to that proposed here. I have a review of the book published in *Philosophical Psychology* (Michel, 2023). The main difference between his account and mine is that while I tend to endorse modal monism, Dove is neutral as to a possible role of amodal representations. Also, his account is not couched in terms of the PP framework, but still on a level of description high enough to be compatible with the PP approach.
2. *Intentional content.* I have not addressed intentional content, i.e., how a concept can refer, which is of huge interest for a philosophical notion of concept. Others have done some relevant and interesting work here already, which I will not discuss as it falls outside the scope of this thesis (e.g., see Williams, 2018, for a proposal applicable to PP). Notice that PP does not emphasize the "veridicality" of representations; rather it emphasizes the pragmatic value of them. This is compatible with veridicality coming into the picture in some weaker form. It is plausible to assume that if the world model were not to mirror the structure of the world in an appropriate way it would be of little survival value for the agent.

3. *Cognitive content.* My account is geared toward being an account of cognitive content. I have provided a description of the way a concept is structured and processed. Concepts can be understood as flexible information packages structured as a hierarchical expectation sub-network of features.

4. *Acquisition.* Concept acquisition is just the process of updating and improving the generative model. Acquisition is hence the learning of the model.

5. *Categorization.* A concept under the PP account is inherently a categorization device. The prediction units are predictive classifiers that extract invariants from lower-level prediction units. In a sense, the PP account is radically reducing cognition to multi-level categorization. We cannot predict everything on the level of maximum detail, so cognition is fundamentally based on categorization to regulate prediction granularity.

6. *Compositionality.* I will expand on compositionality (or more generally, productivity, systematicity and compositionality—the "PSC property") much more in subsequent chapters. We have seen in Chapter 2.2 that compositionality is at the heart of the challenge from higher cognition for PP. Here, it suffices to say the following. Let us first distinguish between "concept compositionality" and "propositional compositionality". Given the structure of concepts, it is easy to see how concept compositionality is possible: new concepts can be composed by feature selection and combination. E.g., in the case of pet fish, some features of fish and some of pets (plus possibly other features determined by the context) are coactivated forming the combined concept of pet fish. Now, on the level of propositional compositionality we need to see how different concepts corresponding to word meanings compose the
meaning of a complex proposition or sentence. I will address this later in Chapters 6 and 7. To anticipate the idea very roughly: as my notion of concept includes compressed representations of complex situations, propositional compositionality works exactly like concept compositionality. However, this is a different form of compositionality than the common-sense concatenative compositionality from LOTH or generative grammar.

7. Publicity. The question here is how concepts can be shared and how we can understand each other if each person has her "own" concepts in her head. I am focused on individual minds, and a deeper discussion of how minds coordinate with each other is outside of the scope of this dissertation. I find this concern to be not as pressing as many others (including Prinz) find it. Why should we not be able to successfully communicate without "identical" concepts or models? Given that the external world causally shapes our model of that world, and we share the world and have similar bodies, we tend to infer world models that are sufficiently similar in relevant aspects, which allows us to communicate successfully. Also, our world model plausibly contains a sub-model of how others use certain concepts. So, there are a lot of considerations that suggest that there is not a deep problem if we abandon the naive picture where we literally need to "share the same concepts". This does not mean, of course, that this is not a very relevant and interesting question that is worth answering in much more detail.

I would like to add two additional desiderata that seem especially relevant for the psychological and neurobiological perspective on concepts that I pursue here.

Firstly, the account of concepts should be biologically plausible. My account is designed precisely to perform well here. It is grounded in a neuroscientific framework of the brain that is receiving increasing theoretical and empirical support.

Secondly, an account of concepts should be able to account for the cognitive phenomenology. Note that PP has already provided various accounts of perceptual phenomena (e.g., binocular rivalry, and rubber hand illusion), phenomena in the realm of abnormal psychology (e.g., autism, and schizophrenia) and consciousness. Therefore, PP seems a promising framework and provides resources to tackle issues related to phenomenology. To dig deeper into this issue is, however, not
within the scope of this dissertation and I refer to the further work section at the end for some additional comments.

Overall, I conclude that there is no desideratum that needs to raise serious concerns for the PP account of concepts. On the contrary, given that it is grounded in neuroscience, it promises to contribute to neglected desiderata like biological plausibility. Also, a lot of previous work within the PP framework can be leveraged to tackle interesting questions related to cognitive phenomenology.
Part 3 - Language and Logic

Introduction to Part 3

In Part 2, I have dealt with concepts and how they are represented and processed in the mind. In this part I will deal with language and how concepts and language hang together.

Natural language is often considered the flagship of all human competencies. Depending on one's preferred view, language is important for interpersonal communication and coordination, but also individual cognition. Some consider that language is the necessary vehicle of conceptual thought. The preferred view of some PP theorists (e.g., Clark, 2006; Lupyan & Clark, 2015) is that language is a cognition enhancer, but not a capability that is necessary for conceptual thought. That is also my assumption.

Chomsky has earned the credit for having initiated and led a "cognitive turn" in linguistics with his Generative Grammar (GxG). He considers language to be a specialized module in the brain that is a product of evolution. What we know when we know a language is a lexicon and a set of syntactic rules for admissible combinations of words. This grammatical rulebook must be implemented in the brain as some deeply hidden and abstract "universal grammar". Imagine that this universal grammar has some tuning wheels. Each language corresponds to a specific configuration of how they are tuned. So, what children need to learn is only the right configuration of the tuning wheels, not all of the grammar. This explains why children can learn a language with very sparse input. The many ways in which parameters can be tuned explains the enormous (surface) diversity of languages.

I will follow a completely different language paradigm, namely, Langacker's Cognitive Grammar (CxG), which is a "use-based" paradigm that denies deep structures but derives grammatical regularities ("constructions") from observed language use. Cognitive Grammar (or "Construction Grammar" more generally) is arguably the main rival of GxG. CxG is motivated by exceptional phenomena that are difficult to accommodate in GxG, like, for example, idioms. I argue that CxG, if given a PP twist, has the potential to honour the motivation of Chomsky's GxG, which is to explain language learning with sparse input. In the PP/CxG picture, children can learn with
sparse input because the brain is a Bayesian inference machine with cognitive biases. Somehow simplified, inductive biases allow for supplementing the sparse input to achieve learning.

Let me now turn to the question of how concepts and language are linked. In the Chomskian view, lexicon and grammar are separate entities. Word meanings are concepts, and grammars are formal rules for combinations of words. But this link is much deeper in CxG. In CxG, grammar itself is also conceptual, not merely formal-syntactic. Grammatical rules are just more entries in the "lexicon". This broadly understood lexicon is called the "construct-i-con", because instead of words it contains a host of "constructions". Words are constructions, and grammatical regularities or idiomatic patterns, are all constructions as well. This is a very radical view on language, but it is a thriving branch of Cognitive Linguistics and arguably the most important rival for Chomskian GG currently.

Generative Grammar has been the overwhelmingly dominant paradigm for language. It is helped by the fact that the lexicon-plus-rules conception of language is very much a common-sense conception: when we learn a second language consciously, we do so by learning the vocabulary and grammar rules. By observing how language apparently works, Fodor has concluded that our thoughts must also be language-like. This is the Language of Thought Hypothesis (LOTH). The idea is to explain that if the thoughts we can express in language are systematic, productive, and compositional, then our thoughts must also be that way. The LOT is not a natural language, but also works with an inventory of concepts and rules for their combination. This means, according to LOT, the properties of conceptual thought are derived from the Chomskian understanding of language.

As already mentioned in Section 2.2, under the LOTH paradigm, indeed, PP might face an intuitively difficult challenge in accounting for language and conceptual thought more generally, because PP is not couched in terms of symbols and rules. How, then, can rules and lexicon (both for language and conceptual thought) be realized in PP and how can we get the observed productivity, systematicity and compositionality, that characterize language and conceptual thought?

In this part, I tackle this problem by questioning the very language paradigm of GXG. I suggest that it is from that paradigm that we derive our understanding of how
conceptual thought works. Provocatively, I claim that we have wrong intuitions about the compositionality (or the PSC property in general) of thought because we think wrongly about how language works.

The idea I pursue here then is to find a plausible language paradigm that is a more natural partner for PP. Fortunately, CxG promises to be such a partner, as I will argue. Once we have such a new paradigm, we can endorse a new LOTH, a LOTH*, with a different language paradigm. If this new language paradigm stands in a relation to PP as GxG stands to LOTH, then we seem to have deactivated a concern for PP. The intuitive challenge that the structure of PP does not easily map the structure of LOTH is therefore resolved, because we now have LOTH* to which PP does map. I also argue that we can even extend the CxG approach beyond natural languages to formal logic.

In this part, I proceed in two steps. First, in Chapter 6, I will argue that PP is for CxG what LOTH is for GxG. Once we abandon GxG and endorse CxG, we start to meet the challenge from language and compositional conceptual thought for PP. In the brief Chapter 7, I extend the CxG approach from natural language to formal logic.
Chapter 6. Scaling up predictive processing to language

Abstract
Predictive processing (PP) is an increasingly influential neurocognitive-computational framework. PP research has so far focused predominantly on lower level perceptual, motor, and various psychological phenomena. But PP seems to face a "scale-up challenge": How can it be extended to conceptual thought, language, and other higher cognitive competencies? Compositionality, arguably a central feature of conceptual thought, cannot easily be accounted for in PP because it is not couched in terms of classical symbol processing. I argue, using the example of language, that there is no strong reason to think that PP cannot be scaled up to higher cognition. I suggest that the tacitly assumed common-sense conception of language as Generative Grammar ("folk linguistics") and its notion of composition leads to the scale-up concerns. Fodor's Language of Thought Hypothesis (LOTH) plays the role of a cognitive computational paradigm for folk linguistics. Therefore, we do not take LOTH as facing problems with higher cognition, at least with regard to compositionality. But PP can plausibly play the role of a cognitive-computational paradigm for an alternative conception of language, namely Construction Grammar. If Construction Grammar is a plausible alternative to folk linguistics, then PP is not in a worse position than LOTH.

Keywords: compositionality; Construction Grammar; Generative Grammar; higher cognition; language; Language of Thought Hypothesis; predictive processing

6.1. Introduction
Predictive processing (PP) is an increasingly influential cognitive-computational framework for understanding the mind (e.g., Clark, 2013, 2016; Hohwy, 2013, 2020). PP is ambitious, as it deals with cognitive agency in general, including perception, cognition, and action. The basic idea behind this paradigm is often expressed by the slogan that "the brain is a prediction machine". PP implies a revisionary picture of cognitive agency because what we believe, perceive, etc. are "hypotheses" generated by the brain that are driven to match incoming sensory evidence. The
brain is constantly improving a hierarchically structured model based on a mechanism of prediction error minimization, which approximates Bayesian inference.

I assume for the current purposes that PP is an emerging paradigm, i.e., a set of concepts and principles that guide more specific theory and model building, not a fully-fledged theory or model. As it is still emerging, no consensus exists as to its precise constituting concepts and principles. For this reason, later I need to make explicit what I take those core commitments to be.

A lot of PP oriented research has focused on perceptual and motor, as well as certain specific psychological phenomena, but PP is not yet well understood where higher cognition is concerned. Indeed, it can be seen from a recent review of the philosophical oriented literature in PP by Hohwy (2020) that PP treatments of higher cognition are still marginal. Higher cognition encompasses conceptual thought generally, and, among others, specific competencies like classification, categorization, analogy making, deduction, planning, mathematical discovery and reasoning, theory building, counterfactual reasoning, and language, as well as abilities related to social and cultural interaction, communication, and collaboration between humans.

PP theorists have pointed to the capacity of the models to which they are committed (see 2.3.) to learn and represent complex, structured world knowledge, including representations on many levels of abstraction (e.g., Clark, 2016, pp.171–176). However, the details about those representational elements, and about the compositionality of conceptual thought and language need further development. There is, of course, some incipient work, as well as a lot of related work that is close to (and very relevant for) the PP paradigm.

As to the first type of work, for instance, Friston & Frith (2015) and Vasil et al. (2020) propose PP accounts of communication where agents are seen as coupled generative models. Language has been addressed to some extent within a PP perspective (e.g., Lupyan, 2012; Lupyan & Clark, 2015), but the treatments are limited to pointing to the value of language as a device that enhances cognition in the prediction economy, especially through linguistic labels that serve as "artificial

---

57 Throughout the literature, PP is characterized in many ways (e.g., Williams, 2020: theory, framework; Tate, 2019: research paradigm, framework).
contexts" (Lupyan & Clark, 2015) and facilitate perceptual processing. However, those proposals do not spell out how exactly language and concepts are represented and processed in the mind, nor do they discuss compositionality in detail.

With regard to work on higher cognition from perspectives close to PP, Bayesian approaches have become extremely influential (see, e.g., Jones & Love, 2011, and Colombo & Hartmann, 2017 for an overview and discussion). In particular, hierarchical Bayesian approaches (e.g., Griffiths et al., 2007; Kemp & Tenenbaum, 2009; Lake et al., 2015; Tenenbaum et al., 2011) are very relevant and have been taken on board by PP. However, strictly speaking, PP is Bayesian only derivatively and approximately, and adds further commitments (on which I will elaborate in a moment). Bayesian approaches are often computational level accounts in the sense of Marr's (1982) three levels of description account and have the form of (acausal) mathematical equations, not neuro-mechanistic models (Colombo & Hartmann, 2017, p.455; see also Tenenbaum et al., 2011, p.1284 and Jones & Love, 2011, p.170). But the PP paradigm, as we will see, cuts across all of Marr's levels, i.e., it also includes algorithmic and implementation-level commitments (Sprevak, 2021a).

The lack of coverage of higher cognition that explicitly deals with compositional conceptual thought and language within the PP paradigm might be a symptom of what has been called the "scale-up problem" (e.g., Silva & Ferreira, 2021):

Furthermore, a more radical approach to cognition faces the so-called "scale-up objection", namely, the challenge of proving itself relevant for the investigation of traditional problems related to higher level cognition involving concepts such as contentful information, representational states, symbolic thought, logic, mathematics etc.

This problem seems to generally afflict all cognitive accounts that deviate from traditional symbolic computationalism, a cognitive-computational paradigm famously

---

58 Accounts of language that share some commitments with PP are Pickering & Garrod (2013) and Pickering & Gambi (2018); however, they do not focus on compositionality.

59 Clark, for instance, considers PP to be a "process theory" for Hierarchical Bayesian Models (2016, p.175).

60 In a way, PP takes seriously the concerns raised by Jones & Love (2011) with regard to Bayesian cognitive modelling, namely that it should combine with other branches of the cognitive sciences (e.g., neuroscience) and integrate mechanistic models.
articulated in form of Fodor's "Language of Thought Hypothesis" (LOTH) (e.g., 1975, 2008), in which thought is syntax-sensitive processing of discrete symbols.

The scale-up problem also seems pressing specifically for PP. In his influential target paper from 2013, Clark pointed out that it is still unclear how to extend the PP account to higher cognition (2013, p.201):

> Questions also remain concerning the proper scope of the basic predictive processing account itself. Can that account really illuminate reason, imagination, and action-selection in all its diversity? What do the local approximations to Bayesian reasoning look like as we depart further and further from the safe shores of basic perception and motor control? What new forms of representation are then required, and how do they behave in the context of the hierarchical predictive coding regime?

Williams (2020) recently restated the general scale-up concern of the PP community in the following way:

> As even its most enthusiastic proponents acknowledge, one of the most important challenges for predictive processing is whether the mechanisms it posits can be extended to capture and explain thought (Clark, 2016, p.299; Hohwy, 2013, p.3; see also Roskies & Wood, 2017).

Williams has then put forward various arguments to the effect that PP cannot account for conceptual thought (2019, 2020). I cannot discuss and respond in detail to his objections here; Williams' nuanced argumentation deserves much more space. However, I do want to highlight that most of his objections are grounded in considerations of compositionality as a core feature of conceptual thought. Williams argues, among others, that PP does not have the expressive power needed for conceptual thought, because it is not "richly compositional" (i.e., as expressive as first order logic). But note that the underlying notion of compositionality is based on classical amodal symbol systems and formal logic. It is precisely the aim of the present chapter to consider an alternative to this notion of compositionality.

I assume then the following motivation for the scale-up concern (for which Williams is also an example). It is generally assumed that higher cognition requires conceptual thought, which is productive, systematic, and compositional (I will expand on those notions in a moment). Mental processing is then carried out by manipulating discrete mental symbols in an algorithmic fashion. A classical computational picture of the
mind fits prima facie the bill better than PP because PP is couched in terms other than discrete symbols, algorithms, or rules (see also Piccinini, 2020, pp.125–126).

The core argument in this chapter starts from the idea, for which I claim no originality, that a certain conception of language—common-sense Generative Grammar (GxG)—informs the intuitions about composition, which are then underpinned by LOTH as a cognitive-computational paradigm. By "cognitive-computational paradigm" I am referring to a set of concepts and principles that guide and constrain specific theories and models about the structure and format of mental representations and their processing mechanism. I argue then that if this common-sense conception of language is replaced by Construction Grammar (CxG), PP can play the role of its underpinning cognitive-computational paradigm. In other words, LOTH is to GxG what PP is to CxG.61

The remainder of this chapter is organized as follows. Firstly, I lay out what I take PP to be committed to and compare it briefly with the LOTH paradigm (Section 6.2). I then recapitulate the relationship of LOTH with natural language (Section 6.3). In Section 6.4, I explain the strategy for arguing that PP can in principle meet the scale up challenge for language. In Section 6.5, I introduce Construction Grammar, and specifically its notion of productivity, systematicity, and compositionality. In Section 6.6, I suggest that PP can serve as a cognitive computational paradigm for CxG.

6.2. PP and LOTH as cognitive computational paradigms

6.2.1. What sort of a theoretical entity is PP?

As pointed out already, PP has been characterized in a variety of forms, so it is crucial to clarify what sort of thing we refer to with "PP". Such a clarification is also important because PP is sometimes criticised for being underspecified, ill-defined, impossible to verify, etc. (e.g., Litwin & Miłkowski, 2020). Those criticisms, however, presuppose that PP is a theory that can produce very specific falsifiable/verifiable predictions about target phenomena. But if we characterize PP as a paradigm then

61 While pairing Construction Grammar (or Cognitive Linguistics more broadly) with alternative connectionist or neurocomputational approaches is not a new idea (e.g., Feldman, 2008; Pulvermüller, 2010; Pulvermüller et al., 2013), the novel contribution of this chapter is to use predictive processing as a paradigm alternative to LOTH.
such criticisms miss the point. What might deserve those criticisms are, of course, specific theories of specific phenomena that make use of the core concepts and principles of PP.

I take predictive processing to be a cognitive computational paradigm that is only just emerging and still under construction. For the current purposes, I use "paradigm" in a broad sense, understood as a set of concepts and principles that provide an interpretive framework that guides and constrains the development of specific theories and models of some domain of interest. Such principles can be extracted from "exemplars of good science", of course, as Kuhn believed (Bird, 2018). Here a paradigm is not some specific empirically verifiable theory that serves as an example. This characterization of PP implies that the notion of PP is necessarily schematic.

By cognitive computational paradigm I am referring to such a set of concepts and principles that guide and constrain further algorithmic and implementational level accounts of the nature, format and processing of mental representations that constitute cognition. I take Fodor's Language of Thought Hypothesis to be an example of a cognitive-computational paradigm. As I will use it as a foil in what follows, a brief sketch is in order.

6.2.2. Fodor's LOTH as a cognitive-computational paradigm

Fodor's well-known and extremely influential "Language of Thought Hypothesis" (LOTH) (Fodor, 1975, 2008) is an example of a cognitive-computational paradigm (as opposed to a theory) in regard to what is being discussed here. LOTH in its deterministic version is generally considered to be a dead horse as a cognitive paradigm (e.g., Williams, 2020; Piccinini, 2020, p.312); however, when compositionality is being discussed, it still serves as an influential benchmark, which

62 Connectionism was broadly considered to be the rival paradigm to LOTH. There is extensive literature on the LOTH versus connectionism debate, especially with respect to questions around compositionality, which I cannot discuss here (see, e.g., Kiefer, 2019, for a good overview and a defence for "pure connectionism"). The debate is considered by some scholars to have reached a stalemate (see also Rescorla, 2019). Even classical computationalists and connectionists nowadays tend to move towards positions that recognize the importance of neuroanatomical and neurophysiological constraints for a full multi-level picture of cognition (Piccinini, 2020, pp.201–202). Let me highlight that the PP paradigm, as I have characterized it here, should be seen as such a neurocognitive paradigm. Cognitive theories within the PP paradigm should ultimately provide neural mechanisms (see, Piccinini, 2020, for an extensive defence of the role of neuroscience for cognitive theories).
captures a common-sense view, and which often operates in the form of a tacit presupposition. Also, LOTH is still very much alive in probabilistic versions.63

Aydede (1997) describes LOTH as being characterized by "meta-architectural" properties, which define a class of cognitive-computational architectures that fall under it (p. 65). LOTH, according to Fodor & Pylyshyn (1988, pp.12–13), has the following features with regard to representational format and processing principles:

a. representations of a system have a combinatorial syntax and semantics such that structurally complex (molecular) representations are systematically built up out of structurally simple (atomic) constituents, and the semantic content of a molecular representation is a function of the semantic content of its atomic constituents together with its syntactic/formal structure, and

b. the operations on representations are (casually) sensitive to the syntactic/formal structure of representations defined by this combinatorial syntax.

Fodor does not provide criteria for how LOTH could be empirically verified. Rather, he motivates and then puts forward a set of concepts and principles (on a cognitive level of description) with regard to mental representations and their processing that guide and constrain the development of more specific theories and implementational models of mental phenomena, i.e., LOTH rather than a theory, in a strict sense, is a cognitive-computational paradigm.

I take the PP paradigm to be at a similar level of description to LOTH and competing with it. What is needed then is to spell out what constitutes the PP paradigm. I will highlight the fundamental differences by juxtaposing the key commitments of the two paradigms with regard to representational structure and processing principles.

63 Goodman, Tenenbaum & Gerstenberg (2015), Piantadosi & Jacobs (2016), and Ullman & Tenenbaum (2020) have proposed accounts of concepts and conceptual development relying on probabilistic programs (see also https://probmods.org), which combine structured symbolic representations and probabilistic elements. Note that that such “probabilistic programming languages” essentially follow the LOTH paradigm, though they add symbols representing probabilistic entities to the representational ontology.
6.2.3. The PP paradigm and its core commitments

As mentioned already, so far there is no agreed-upon articulation of the PP paradigm. However, from Clark (2013, 2016) and Hohwy (2013, 2020), and in general from the increasing literature that makes use of the PP framework in one form or another, we can extract largely overlapping concepts and principles. Those we could consider tentatively to be the PP paradigm's core commitments.

6.2.3.1. Core commitments of PP

The core tenet that crystalizes from the PP literature is that the mind entertains a probabilistic, hierarchical, generative model that aims at anticipating the inflow of sensory information. The central operating principle is prediction error minimization that approximates Bayesian inference. The system adapts the model such that the prediction error is minimized on average and in the long run.

It is probabilistic because it represents probability distributions over "hypotheses" and inferences are carried out by approximate Bayesian inference. It is generative because it generates top-down predictions/hypotheses (rather than merely, e.g., classifications by bottom-up processing). And it is hierarchical, because the hypotheses are organized in a hierarchical structure, where higher level hypotheses are the "priors" of lower-level hypotheses. The higher levels represent regularities of larger spatial and temporal scales, i.e., more compressed and abstract information.

PP emphasizes that predictions flow top-down and are being matched by the bottom-up flow of "evidence". To be in a certain perceptual state is to issue a prediction of that state that is consistent (has a minimal prediction error) with bottom-up signals that serve as evidence. PP is a neurocognitive paradigm, and its concepts and principles extend to the neural level, though, as a paradigm, those are still schematic. The PP model is neurally implemented by an interconnected hierarchy of pairs of representation and error units (consisting of a group of neurons). The prediction signals from level N are compared to the representation units on level N-1 and an error signal is generated. The error is weighted by some mechanism that estimates the reliability or relevance of the error signal. This is achieved by an

64 Hohwy calls PP the "Prediction Error Minimization" framework.
estimate of the *precision* of the signal. Details do not matter here, but very briefly, the precision of some variable can be determined via the magnitude of the inverse of the variance of the probability distribution of that variable. The estimated precision allows, on the one hand, for tuning down noisy and unreliable error signals; in this way the system prevents the model from being unnecessarily updated. On the other hand, incoming information that is precise and reliable should lead to adjustments of the model if the top-down prediction does not conform to it. This mechanism then serves as a tool to balance the influence of the top-down versus the bottom-up flow of information: either we rely more on prior beliefs, or we are attuned more to the sensory information.

The top-down flow of prediction signals functions as "priors" that might shape predictions on lower levels. Through this complex interplay of bottom-up and top-down information flow, the model is constantly updated on all levels based on prediction errors, which should be minimized on average and in the long run. Prediction error minimization happens all the time on all levels simultaneously. This makes processing in PP holistic because a given prediction unit is directly influenced by other prediction units in adjacent layers, and indirectly by prediction units in other layers (like a domino effect). That allows for context-sensitive processing because the state of a given prediction unit is determined by the state of many other prediction units that represent this context.

As the hierarchy bottoms out at the sensorimotor level, and the focus is on the prediction of sensory input and the interaction with the environment (to get a "grip on the world", as Clark expresses it), the PP paradigm can be considered an "embodied cognition" paradigm. While embodied cognition includes many different approaches (see Newen et al., 2018), the common theme is the central role that the body, i.e., the sensorimotor apparatus, and its interaction with the world plays in cognition. Consequently, for the PP paradigm it is natural to adopt a modality-specific (i.e., sensorimotor) format of its representations, not amodal formats as in LOTH. At

---

65 The exact algorithmic and implementational level description of precision weighting is still debated (see, e.g., Sprevak, 2021b).

66 In LOTH, symbols are processed "locally", i.e., their processing is context independent. Prediction units on certain levels in the hierarchy can be seen as representing hypotheses as "beliefs" (see, e.g., Smith et al., 2022), so PP allows for context sensitive belief updating. But notice that the issue of how to relate folk psychological notions like belief, desire, intention, etc. to PP is the subject of an ongoing debate (see also Dewhurst, 2017).
higher levels in the PP model hierarchy, those modality-specific representations are more abstract and compressed and are combined into multi-modal representations.

Table 6.1 summarizes the proposal for the characterization of the PP core paradigm through a juxtaposition with LOTH.

<table>
<thead>
<tr>
<th>Feature</th>
<th>LOTH paradigm</th>
<th>PP paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format of representations</td>
<td>• amodal</td>
<td>• modality-specific (sensorimotor grounded)</td>
</tr>
<tr>
<td></td>
<td>• abstract</td>
<td>• different degrees of abstraction</td>
</tr>
<tr>
<td></td>
<td>• deterministic (LOT) / probabilistic (&quot;pLOT&quot;)</td>
<td>• probabilistic</td>
</tr>
<tr>
<td>Structure of representations</td>
<td>• sequential / recursive</td>
<td>• hierarchical network with an abstraction/ compression gradient</td>
</tr>
<tr>
<td>Processing principles</td>
<td>• syntax-sensitive processing/algorithmic</td>
<td>• prediction error minimization</td>
</tr>
<tr>
<td></td>
<td>• local</td>
<td>• holistic</td>
</tr>
</tbody>
</table>

**Table 6.1**: Comparison of key features of LOTH and PP as cognitive-computational paradigms.

6.2.3.2. Possible commitments that are not part of the PP paradigm

The above characterization of the PP paradigm leaves open many aspects about the exact implementation of each of the principles. For the specification of an implementational level more detailed assumptions are necessary. For instance, how many neurons compose a "prediction unit"? Are the principles of PP pervasive in the brain or found only in some specific brain regions? Does PP describe the only type of representation and processing mechanism in the brain, or are there others? Often the hierarchical structure is constrained such that a level is only connected to the next lower and higher levels. But how central is this assumption? Could there be adaptations where connections "skip" over levels?

A further debate is related to motivating the prediction error minimization principle. Friston (e.g., 2010) relates prediction error minimization to free energy minimization. This supposedly solves the "problem of life": how can an organism evade entropic
disintegration? But is this a crucial assumption—and is this link a coherent assumption at all (see Williams, 2022)? While Hohwy seems to endorse it, Clark seems not to, at least not strongly.

Furthermore, there are a variety of proposals for an associated mathematical apparatus (e.g., Spratling, 2017; Clark, 2013) and for a specific neural architecture (e.g., Bastos et al., 2012; Kanai et al., 2015; Keller & Mrsic-Flogel, 2018; Siman-Tov et al., 2019; Weilnhammer et al., 2018). Also, many other more specific questions need to be answered to get at a falsifiable theory or model: how exactly is the prediction error minimized in the brain, by stochastic gradient descent, or other mechanisms? What is an appropriate mathematical description of the node network? What is the mechanism with which nodes are added (or deleted)? Is precision weighting implemented by neurotransmitter dynamics? Which ones? And so forth. The number of open questions is daunting, which shows that PP at this stage should really be seen as a paradigm for a research programme (see also Sprevak, 2021a).

With respect to commitments to mathematical models, let me briefly refer to Williams’ objections from the introduction. According to Williams (2020), PP theorists are committed to so-called Probabilistic Graphical Models (PGMs). His argument is then that those models lack the necessary expressive power for conceptual thought (very roughly: they can only represent facts, not objects and relations). Details do not matter for the current purposes; the point I want to make is that the PP paradigm as I have pictured it is a mechanistic neurocognitive paradigm. Therefore, it does not need to commit to any unifying mathematical model at all.

6.3. LOTH and natural language

It is worthwhile briefly revising the (abductive) core argument that supports LOTH. The purpose is to highlight a crucial point for my argument, namely how our conception of language determines our conception of compositionality.

67 But see Smith et al. (2020) for a recent proposal within the PP framework. The authors propose how latent variables (which he calls “concepts”) in the generative model can be added or deleted.

68 Williams might indeed be right that many PP theorist commit to simple PGMs. But note that PGMs could be extended to more expressive versions, e.g., Relational PGMs (Getoor et al., 2001)

69 Williams considers exactly this strategy (avoiding the commitment to the PGM model) on behalf of PP but thinks that PP then loses explanatory power. However, even if this were true, it would be only for a specific theory, not a paradigm.
6.3.1. The argument for LOTH from natural language

Simplifying very much, one important motivation for LOTH stems from the observable properties of natural language. Natural languages appear to be productive, systematic, and compositional (short: PSC) in a very explicit manner: parts (words) are assembled following certain rules into sequences (sentences). It seems that we can generate from finite means, i.e., an inventory of words and grammatical rules, an infinite number of sentences; or at least we can imagine how we could go on and on infinitely in principle (productivity). It also seems that if we can produce and comprehend sentences like 'Peter kisses Mary' then we can produce and comprehend systematically related sentences like 'Mary kisses Peter' (systematicity). Finally, the meaning of a sentence seems to be determined by the meanings of the words it contains and the way that they are syntactically combined (compositionality).

As language expresses thoughts, the best explanation for language having the PSC property is that thought has it as well (cf. Fodor and Pylyshyn, 1988, pp.37–41).

6.3.2. Folk generative grammar (GxG) as a presupposition of LOTH

Note that "language" in LOTH must be based on some specific conception of language. Then, according to LOTH, the nature of thought has the structure of language under this conception. Fodor’s conception of language is plausibly "folk linguistics", a common-sense Chomskian-style Generative Grammar (henceforth GxG). GxG characterizes the body of knowledge one possesses when one has the competence to speak a language. According to GxG, we hold in our memory a lexicon and (recursive) rules for combining words into sentences. This folk notion of linguistics follows directly from observing the surface form of natural language as consisting of sequences of written or spoken words (or gestures).

One plausible explanation of the origin of GxG folk linguistics is that its supporting intuitions are grounded in action (see also Dutilh Novaes, 2012). Language works

70 There are other arguments for LOTH (see Rescorla, 2019). However, Fodor and Pylyshyn have stressed this one in the context of the debate with connectionism. I therefore take it to be the strongest argument.

71 Note that LOTH might be a "best explanation", but only with respect to PSC. As Fodor himself has pointed out (e.g., 1975, pp.197–205; 2008, Chapter 4), LOTH has shortcomings regarding other desiderata (which should not concern us here), so it is not the best explanation all things considered.
with syntactic rules and words, much like an assembly line where parts are put together to form compounds. That is, we intuitively model language as discrete entities that are composed of or assembled into larger entities. This leads to a concatenative view of compositionality and makes language causally perspicuous to us. There is a sequence of physical entities, written or spoken words, for instance, and they have literally been "put together" following some rules, recipe, blueprint, or algorithm.

The important point here is that LOTH lives up to the PSC desideratum, whose force is grounded in a folk linguistic conception of language. While folk linguistics is quite perspicuous and intuitively very appealing, there are alternatives, as we will see.

6.4. A strategy to address the scale-up challenge for PP

The scale-up concerns have not been articulated in detail in the literature, with some exceptions like Williams (2019, 2020). But any cognitive model deviating from LOTH-based classical computational models seems to evoke a concern about compositionality. Such intuitions were also behind the well-known connectionism–symbolic computation debate. Carried over to PP, it simply is not a classical computational model that relies on the rule-based processing of discrete abstract symbols. In turn, LOTH can straightforwardly account for PSC. Hence PP needs some story for PSC, even if it consists of qualifying it or explaining it away.

A definitive way for PP to meet the scale-up problem for language would be to put forward a specific cognitive-computational model for the language faculty under its umbrella. Such a model/theory should be empirically supported in the strong sense that Litwin & Młkowski (2020) are demanding (i.e., the empirical evidence should be decisive evidence for the proposed model and against contenders, and not only "compatible" with the model). Also, it should ideally be able to make novel predictions. However, my ambition in this chapter necessarily needs to be more modest. I will therefore focus on sketching how PP might plausibly be a cognitive-computational paradigm for some suitable existing language paradigm. If PP can play the role of the cognitive computational paradigm for some plausible conception of language, then PP has started to meet the scale-up challenge.
The critical point to note is that LOTH is a plausible cognitive computational paradigm only for language understood in a certain way. This "certain way" I have characterized as folk linguistics (GxG).

Now, interestingly, the efforts to defend different cognitive-computational paradigms, like connectionism, have often focused on showing how to replicate the common-sense PSC property of language. That is, for example connectionists have often felt pressed to show how to replicate language-like thought, where they tacitly accept that language is to be understood in folk linguistic terms. In other words, many defenders and opponents of LOTH are in the grip of a specific language paradigm, folk linguistics.

The argument for LOTH from Section 6.3 can be teased apart into two independent claims: firstly, the normally tacit assumption that language has certain properties, those captured by folk linguistics, and secondly, the claim that the best explanation for the properties of language, whatever they are, is that thought is language-like.

Criticisms of LOTH have typically focused on undermining the second claim\textsuperscript{72}. My strategy in what follows is different: I grant the second claim but question the first one. I suggest adopting a view on language that is different from folk linguistics. In other words, I suggest a revision of what it means to say that thought is "language-like" (and consequently we also get a different notion of compositionality).

Let me outline then the structure of the argument that PP does not face a fundamental scale-up problem for language based on the strategy just developed:

I) Intuitions about a scale-up problem for PP arise because of a mismatch with the common-sense notion of composition related to folk linguistics, which follows the Generative Grammar paradigm (GxG).

II) LOTH serves as the cognitive-computational paradigm for GxG.

Construction Grammar (CxG) is a plausible language paradigm for which PP can serve as an underpinning cognitive-computational paradigm.

---

\textsuperscript{72} E.g., see Salje, 2019. Some have argued that "mental maps" can give rise to the PSC property of language (Braddon-Mitchell & Jackson, 2007). It has also been argued that connectionist structures can exhibit the PSC property in an implicit way (see Aydede, 1997), e.g., using Smolensky's (1990) tensor product representations.
III) PP can then address the challenge from productivity, systematicity and compositionality (PSC) by deference to the conception of composition of CxG.

We have already established step I) in the previous sections. Let me turn to step II) (Sections 6.5 and 6.6). From I) and II) then follows III).

6.5. Construction Grammar and its notion of compositionality

In this section, I will briefly provide a "theoretical minimum" of Construction Grammar (CxG) for those not familiar with it. CxG is arguably the main rival of the mainstream linguistic theory, namely Generative Grammar (GxG), and differs profoundly from it. After a short general introduction, I will focus on the PSC property, which is our main concern here.

6.5.1. What is Construction Grammar?

Construction Grammar differs from Generative Grammar in important dimensions by which we can characterize a linguistic theory: (a) the way how language is acquired, (b) what sort of knowledge linguistic knowledge is, and (c) how it is represented in the mind.

(a) According to CxG, linguistic knowledge is acquired by extracting patterns on all levels of the linguistic hierarchy (e.g., phonetic, lexical, syntactic levels) from experienced language use. What matters are learned surface structures, not inborn and hidden deep structures as in GxG. Knowledge of a language is having a large inventory of such learned patterns, which are called "constructions".

(b) Crucially, according to CxG, linguistic knowledge is not structured into autonomous modules for syntax and lexicon, where syntax is purely formal, and the lexicon contains meaningful words and expressions. Rather, CxG posits only a sort of generalized lexicon, the "construct-i-con". The construct-i-con contains not only words, but also all of the learned grammatical (phonetic, morphological, and syntactic) patterns. Grammatical patterns are considered to be not purely formal like

---

73 To be more precise, Construction Grammar (CxG) is a family of linguistic theories (see, e.g., the overview in Croft & Cruse, 2004, Chapter 10). The different versions have in common a set of basic commitments that I denote the "CxG paradigm". I will spell out those commitments with particular reference to Langacker (e.g., 1987, 2008) as well as Goldberg (e.g., 1995, 2019), as those are very elaborate and influential versions of CxG.
in GxG, but also meaningful. This is a most radical deviation from a common-sense view of language. The difference between words like "cat" and grammatical patterns, like, for instance, the basic sentence form subject–predicate [S P] is the level of schematicity. Both 'cat' and [S P] have a meaning. However, the meaning of the latter is, of course, much more schematic/abstract (namely something like "someone did something"), but it is a meaning after all.

(c) CxG follows an embodied cognitive paradigm. In other words, the format in which linguistic forms and meanings are mentally represented is not by amodal LOT-like symbols, but representations are *modality-specific conceptualizations*. In other words, the representations are based on and abstracted from experienced sensorimotor information. Importantly, constructions need to be understood as "form–meaning" pairs. For instance, a word has a form (phonology, morphology, etc.) and a meaning (the concept denoted by that word). In the case of [S P] the form is represented, e.g., as an "experience" of the sequence of the slots with first an agent and then an action. This unified view has an economic ontology: we only need modality-specific and no amodal representations. This representational ontology is important common ground with PP, as we will see.

Let me now turn to making more explicit how all those characteristics lead to a view about PSC that is different from common-sense GxG.

6.5.2. Productivity, systematicity and compositionality in Construction Grammar

As suggested previously, the appeal of the common-sense compositionality of language understood as GxG—on which LOTH rests—most likely stems from the perspicuity of the assembly of discrete entities following certain instructions. In other words, PSC mirrors the properties arising from literally assembling atomic units into molecular wholes. Those properties are then also ascribed to the language faculty, which is metaphorically understood in this manner. Langacker expresses doubts about this conception:

---

74 But see Michel (2020b) for a view how some modality-specific representations might appear to be amodal ones.
our conception of composition is greatly influenced by certain metaphors whose appropriateness for natural language cannot be accepted uncritically. (Langacker 1987, p.452)

In the concatenative conception of composition of GxG, the syntactic form is not supposed to contribute semantic information. The semantics of the whole is exhaustively determined by the semantics of the parts and the purely formal syntax.

CxG, as opposed to GxG, is motivated by the observation that some linguistic phenomena are best explained by positing that certain semantic properties are ascribed to syntactic structures instead of the lexicon (see, e.g., Goldberg (1995) for the argument structure of verbs). This step dilutes the distinction between grammar and lexicon. Grammatical constructions are meaningful and linguistic entities are located on a gradient from very schematic (e.g. [S P]) to very specific (e.g., 'doorknob'). CxG then paints a picture where all entities are constructions, i.e., use-based form–meaning pairs. Some constructions have schematic slots that can be filled with other constructions, which in turn might have slots that can be filled in. [S P] can be made specific by filling, e.g., the 'S' slot with a more concrete instantiation, like 'animated_object_noun' until the tree bottoms out at a specific word, like 'cat'. When tokening a linguistic structure, like a sentence, we get a tree-like structure—with an abstraction gradient—that bottoms out, at the level of concrete words.

CxG is characterized by weak compositionality:

> By recognizing the existence of contentful constructions we can save the compositionality in a weakened form: the meaning of an expression is the result of integrating the meaning of the lexical items into the meanings of constructions. (Goldberg, 1995, p.16)

> The composite structure is an entity in its own right, usually with emergent properties not inherited or strictly predictable from the components and the correspondences between them. (Langacker, 2008, p.164)

Strong (or full) compositionality, in turn, allows for predictively deriving the meaning of a composite expression from its parts and the way they are arranged. A linguistic structure is a construction if its meaning cannot be predicted from its parts or from other constructions. In the CxG picture, compositionality is graded. For instance, 'jar
'lid' is close to fully compositional. 'Laptop' cannot be understood outside the context of a metonymy (the place where the computer is typically placed stands in for the computer itself). And 'understand'—composed of 'under' and 'stand'—is not compositional at all (we cannot predict the meaning from its parts).

The notion of compositionality in CxG takes into account three levels of semantic contribution: the components, the construction, and, importantly, the context and rich background information:

Virtually all linguistic expressions, when first constructed, are interpreted with reference to a richly specified situational context, and much of this context is retained as they coalesce to form established units; (Langacker, 1987, p.455)

CxG further implies a notion of partial productivity. In constructions you cannot fill in slots freely. It is often not predictable which inserts are allowed. For instance, consider:

1. Mary goes to school
2. Mary goes to work.
3. *Mary goes to company.
4. *Mary goes to hospital.

To the speaker it is not transparent why 'Mary goes to …' can be combined with some but not other expressions.

Even for common-sense PSC systematicity is only partial. One can say both "Peter kisses Mary" and "Mary kisses Peter". But you can't say both "Peter reads the book" and "The book reads Peter". Some authors endorse unrestricted PSC and allow for the latter type of odd sentences, but also more radically, category mistakes like "Green dreams sleep furiously", to be meaningful and truth-value bearing—they are simply false (e.g., Magidor, 2009). But not all agree, and some prefer to rely on selectional restrictions. But the question is then how to model those restrictions. We should consider a slot in a construction not literally as an empty space, but as an abstract concept (a category) that instantiates that slot. All "allowed" instances of that slot concept can then serve as "fillers".

In sum, in CxG, common-sense compositionality is replaced by a PSC property that relies on the structure of a nested hierarchical tree network. The CxG structure and
processing mode is less intuitive and perspicuous than the building-block-plus-assembly-rule account. One reason is the built-in abstraction gradient. Notice that those abstraction gradients can be straightforwardly modelled by hierarchical connectionist architectures, where information is “compressed” in successive hidden layers, very much like the operations of the visual pathway, where neurons in higher levels are sensitive to larger and larger receptive fields.

6.6. PP as a neurocognitive-computational paradigm for CxG

With a working understanding of both PP and Cognitive Grammar and a qualified PSC property in place, we can now complete step II) of the argument from Section 6.4 to the effect that PP can be seen as a cognitive computational paradigm for CxG. The aim of this section is, hence, to establish the analogy between PP and CxG with regard to representational structure and basic processing principles. This analogy underwrites the claim that PP can play the role of a cognitive computational paradigm for CxG.

PP and CxG have been developed in different research communities relying on different interests, concepts, methodologies, terminologies, and perspectives. By establishing correspondences between the two paradigms, one might run the risk of forcing one into the Procrustean bed of the other by interpreting the terms and concepts too liberally. I bite the bullet here. My ambition is not to argue that there is a formally rigorous structural similarity. Nor can I develop here in detail how CxG can be implemented within a PP architecture, which would be a much larger project. My ambition here is only to argue that there is a striking and suggestive analogy.

The following comparison will focus on the core commitments of both PP and CxG, i.e., treat them as a cognitive-computational and language paradigm, respectively, in the sense defined in Section 6.1.

Table 6.1 lists the core features of LOTH and PP. The core features of LOTH match—by design—the features of GxG, which is the reason why LOTH can serve as its cognitive-computational paradigm. I proceed to arguing that a similar analogy can be fleshed out in terms of at least six features of the structure of representations and processing principles of CxG and PP that are diametrically opposed to LOTH and GxG: 1) All linguistic representations are sensorimotor grounded. 2) The
structure of linguistic representation is bipolar. 3) Representations are organized into a hierarchy with an abstraction gradient. 4) They are context sensitive. 5) Processing is both top-down and bottom-up. And 6) The knowledge of a language cannot be fully formalized.

6.6.1. Sensorimotor grounding of conceptualizations

As explained in Section 6.5.1. (c), Langacker rejects the amodal view of mental representations, which has been the signature of LOTH. PP and CxG allow us to make sense of having a fully modality-specific representational system (as vindicated by neo-empiricism, e.g., Barsalou, 1999, 2009; Prinz, 2002). The "sensorimotor grounding" of representations (be it concrete concepts, abstract concepts, or grammatical structures) can be fleshed out as follows (see Michel 2020a, 2020b). A concept in the PP view, is a certain prediction unit (at some level) conceived as a root-node plus the sub-network that depends on that root-node. The sub-network spans many lower levels in the hierarchy. The lower-level nodes represent more and more concrete features of the concept, while the root node is a most abstract, "gist"-like representation that has abstracted away from many concrete features (but retains its modal nature). The structure bottoms out at the lowest level of the hierarchy, i.e., the first layer of the sensorimotor periphery. Note that concepts can be tokened "shallowly" (cf. Simmons et al., 2008; Barsalou et al., 2008), such that they do not always reach the lowest sensorimotor level. For instance, the concept CAT is represented by a prediction unit that serves as the root node plus a tree emanating from that root node with lower-level prediction units representing many "features" of the cat (information about shape, sounds, furriness, etc.). CAT can be instantiated either gist-like (only the root node is activated), or with multi-modal imagery that is concrete to different possible degrees (the lower-level prediction units are activated, the more concrete and vivid the representation). The crucial point is that in CxG grammatical structures are "also concepts" because of their meaningfulness and are represented as prediction units.

---

Both CxG and PP are also characterized by the probabilistic nature of their representations. However, this is not a fundamental difference compared to LOTH, given that, as already mentioned, probabilistic LOTH versions exist.
In this sense, conceptual representations, including grammatical structures, are modality-specific representations involving sensorimotor information both in PP and CxG. The view that conceptualizations are modality-specific, extended network structures is also increasingly being endorsed in neuroscience (e.g., Hoenig et al., 2008; Kiefer & Pulvermüller, 2012; Pulvermüller, 2001).

6.6.2. The “bi-polar” structure of linguistic representations

As already mentioned, constructions are form–meaning pairs. I suggest that constructions correspond to pairs of associated prediction units in PP. One prediction unit represents the form, the other the meaning. For instance, the word construction [CAT / 'cat'] consists of a representation of the concept CAT in the form of a prediction unit and a representation of the written word 'cat' in the form of another prediction unit. The cognitive content of the concept CAT is information about cats, and the content of the word 'cat' is information about the word form 'cat', which might include its composition of letters, phonetic information, and statistical information about its statistical co-occurrence with other words, among others.

Here we get a picture in PP of two parallel hierarchical networks of prediction units, one for the form, and the other for the meaning parts of the constructions. The form hierarchy represents what we consider the "formal" linguistic knowledge, the meaning network world knowledge or knowledge that is conceptual in a traditional sense. The two hierarchical networks are laterally connected, combining the corresponding parts at all of the different levels (see also Michel, 2019, and Rappe, 2022). Some meaning representations might not have links to form representations (non-lexicalized concepts), and some form representations have no links to meaning representations (e.g., meaningless Jabberwocky words, nonsense sentences, or pseudo-letters).

In sum, in LOTH/GxG, to know a language is to have representations of rules (or generative principles) and a lexicon. In CxG, the knowledge of a language consists in the totality of constructions (or construct-i-con). The construct-i-con corresponds to a subpart of the total PP model, namely those prediction units that are involved in some construction, i.e., constitute form–meaning pairs.
6.6.3. Organization of linguistic representations in a hierarchy with an abstraction gradient

Here is a toy example of how the PP hierarchy works in principle. The prediction units at level N can be seen as abstractions over patterns of prediction units at level N-1. For instance, if level N represents a word form, then N-1 represents letters, level N-2 might represent certain edge forms (of which letters are composed) and level N-3 represents a pixel pattern (that forms edge forms).

This model can implement the construct-i-con. Take again the [[S P] / SOMETHING/ONE DOES/IS SOMETHING] construction. This construction can be made more concrete by replacing the S and P "slots" with more concrete expressions, e.g., [ S(animate object) P(action verb) / SOMEONE DOES SOMETHING]. Still, this remains schematic as we can still make the construction more concrete, e.g., ['Peter swims' / PETER SWIMS]. This is a construct that is a maximally specific sentence that could be a possible utterance. Each slot is an abstraction over possible replacements of the slots "one level more specific". By building a tree of all possible replacements for all levels, we get a hierarchical structure with an abstraction gradient. This tree structure of more and more concrete slot replacements in CxG maps onto the hierarchical structure with an abstraction gradient of the prediction unit network in PP.

6.6.4. Context-sensitive processing

In CxG, conceptual representations are flexible "construals", i.e., they have a variably fine-grained structure, depending on the context of their use. Langacker says:

One dimension of construal is the level of precision and detail at which a situation is characterized. […] Alternate terms are granularity and resolution. A highly specific expression describes a situation in fine-grained detail, with high resolution. With expressions of lesser specificity, we are limited to coarse-grained descriptions whose low resolution reveals only gross features and global organization (2008, p.55).

A specific conceptualization consisting of the activation of some hierarchical substructure of the total network draws—in an open-ended fashion—from a set of
available "domains" (the concept's "domain matrix"). The "domain matrix" can be seen as a pool of conceptual features that can be selected on a specific use occasion. Exactly which features are selected depends on various contextual factors (previous discourse, physical/social/cultural context, background knowledge, etc.). In sum, in CxG, a concept is a network of other concepts and the information retrieved, i.e., what other concepts are co-activated, on a given use-occasion, is context-dependent.

PP provides a computational underpinning for context-sensitive modulation of concept features. This is achieved by the precision weighting mechanism that can switch features on and off depending on their estimated reliability and relevance (Michel, 2020a). In PP we can motivate the context-sensitive modulation of concept features as a means of adjusting the representational granularity. It would not be efficient to always predict a situation with the maximum level of detail. So, both PP and CxG assign an important role to the cognitive capacity to regulate the representational granularity. While CxG merely posits such a selection, PP provides a computational sketch of how such a mechanism could be implemented.

6.6.5. The importance of top-down in addition to bottom-up processing

One of the main tenets of the PP paradigm is the bidirectional, top-down and bottom-up flow of information in the multilayer prediction cascade. What PP especially emphasizes is the importance and pervasiveness of top-down influences or predictions which is a feature neglected by more traditional cognitive approaches. In a striking parallel manner, Goldberg stresses the "simultaneous bottom-up and top-down processing" of constructions (e.g., 1995, pp.24–25). Interestingly, she then supplies an analogy from a perceptual domain, namely vision, citing Wheeler's (1970) work, which shows that letters are recognized faster in the context of a word, i.e., the recognition (top down) of a word aids the recognition of a letter, and vice versa. This is precisely the type of example from which the PP paradigm has received significant support (e.g., Rao & Ballard, 1999).

---

76 CxG relies here on an "encyclopaedic" understanding of meaning (e.g., Langacker, 2008, p.38; also, Kecskes, 2013, p.81ff), as opposed to a "dictionary" view. The "dictionary" view of meaning is roughly the classical "definitional" account of concepts (a set of necessary and sufficient conditions), where a concept is characterized by a (limited and fixed) set of features. The encyclopaedic view holds that the meaning of a concept is potentially open-ended.
Goldberg also discusses the **predictive role** of constructions (e.g., 2006, pp.103–126). She says:

> [...] generalizing beyond a particular verb to a more abstract pattern is useful in predicting overall sentence meaning. (2006, p.105)

Take as an example the polysemous verb "get" (2006, p.106), which is a weak predictor of sentence meaning. Consider:

a. Pat got the ball over the fence.

b. Pat got Bob a cake.

"Get" in connection with a Verb-Object1-Object2\textsubscript{path} structure means a caused motion, while in connection with a Verb-Object1-Object2 pattern it signifies the transfer of something. So, there is value in representing generalizations in the form of such phrase structure constructions. Interpretations of sentences can then be supported by top-down predictions of which of the two cases we are dealing with. For instance, if we get an incomplete input like "Pat got the ball ———" we can infer that we have a caused motion construction and can predict top down that the missing word needs to be an object expressing a path.

It is fair to say that the predictive approach is not developed in much detail in CxG. But my point here is that PP would plausibly be a good cognitive-computational ally with respect to this fundamental processing principle which CxG appeals to.

### 6.6.6. CxG and PP and their formalization

One important consequence of the characteristics of CxG I have laid out is that we cannot formalize the grammar in terms of generative principles or explicit rules.\(^{77}\) The non-formalizability in the form of some explicit and precise formal language is endorsed, e.g., by Goldberg and Langacker:\(^{78}\)

> Since language [...] is neither self-contained nor well-defined, a complete formal description (a "generative grammar" in the classical sense) is held to be impossible

---

\(^{77}\) Of course, some formalizations might be descriptively adequate approximations for a certain range of phenomena. So, I am not claiming that formal approaches are not useful.

\(^{78}\) However, one version of CxG, "Unification Construction Grammar," does build on a formalization where constructions are represented by fixed sets of features. However, this approach has important disadvantages (see Goldberg, 2006, pp.215–217 for a discussion).
Language does not resemble a collection of computer programs. Rather, it inheres in the dynamic processing of real neural networks, [...] (Langacker, 2008, p.10)

I have avoided using all but the most minimal formalization in my own work because I believe the necessary use of features that formalism requires misleads researchers into believing that there might be a finite list of features or that many or most of the features are valid in crosslinguistic work. (Goldberg, 2013, p.29)

The underlying reason for not endorsing fully-fledged formalisms is that CxG emphasizes the meaning of grammatical structures, but "meaning is not easily captured by a fixed set of features" (Goldberg, 2006, p.216).

Also, a PP model cannot be fully formalized via rules and an inventory of discrete concepts with a fixed set of interpretable features. Many prediction units are not lexicalized or do not correspond to interpretable concepts because many of them are located on levels in the hierarchy lower and higher than traditionally understood concepts. Furthermore, the flexibility and context sensitivity of the whole model is also an obstacle to a full formalization. This is, again, in opposition to LOTH/GxG, which is modelled according to a formal calculus, i.e., is paradigmatically formalizable. It is the existence of rules and the explicit manipulation of discrete symbols that makes LOTH in principle tractable. In PP however, the processing is holistic with a crucial role of top-down influences and driven by a self-organizing physical mechanism.\textsuperscript{79} \textsuperscript{80}

Some efforts have been undertaken to computationally model CxG (e.g., Bergen & Chang, 2003; Holmqvist, 1993; van Trijp et al., 2012). However, those do not abandon the classical computational LOTH-type paradigm in their implementational proposals. It might be more promising to endorse the PP paradigm and pursue modern machine learning methods combined with PP-specific architectures (e.g., Maida & Hosseini, 2020; Lotter et al., 2016) for a cognitive-computational implementation of CxG.

\textsuperscript{79} Notice that Friston’s influential Free Energy Principle (e.g., 2010) builds on a formal mathematical apparatus. However, such equations seem not a suitable level of description for language and grammar that captures the PSC property.

\textsuperscript{80} Constructions could maybe be compared to species that emerge in a process that cannot be fully predicted because many contingent environmental and other factors influence the outcome.
Let us take stock. All of the six features discussed in this section represent common ground between PP and CxG, while at the same time they are diametrically opposed to those of the LOTH/GxG paradigm. Therefore, it might be promising that PP and CxG join forces. PP as a cognitive-computational paradigm provides basic concepts, principles, and mechanisms that can constrain and guide the development of more specific implementational level theories and models for CxG.

6.7. Conclusion

Fodor's Language of Thought account (LOTH) is generally recognized as a benchmark where accounting for the productivity, systematicity and compositionality of language and conceptual thought is concerned. As predictive processing (PP) is not couched in terms of symbolic syntax-sensitive computation like LOTH, it seems to face a scale-up challenge regarding higher cognition.

I have argued that predictive processing is not in a worse position than LOTH with respect to the scale-up challenge from higher cognition if one is willing to accept a different language paradigm associated with a different notion of composition. In the same way as LOTH plays the role of a cognitive-computational paradigm for common-sense Generative Grammar, I suggest that PP can play that role for Construction Grammar. PP mirrors relevant properties of the representational structure and processing of Construction Grammar in a way that is similar to how LOTH mirrors those properties of Generative Grammar. PP can then inherit the notions of compositionality, productivity, and systematicity from CxG. The proposal is, interestingly, still a form of LOTH because it accepts that thought is language-like. The novel approach, however, is that it adopts a different language paradigm.

Acknowledgements

I would like to thank Mark Sprevak, Sofiia Rappe, and two anonymous reviewers for their useful comments and suggestions on earlier drafts.
Chapter 7. Logic in the predictive mind

Abstract

Formal logics are often considered formal, "de-semantized" languages that have a purified grammar, as opposed to grammatically more messy natural languages. Moreover, logic is often located in a Platonic realm of abstract truths (see also Chapter 8). In the defence of an embodied PP view of thought, it therefore seems especially pressing to explain how logical thinking arises and functions cognitively. In Chapter 6 I have tried to build a bridge between PP and a linguistic research program, Cognitive Grammar, by showing that they share key assumptions about the underlying structure of conceptual and linguistic representations in the mind. In this way we have tried to respond to a challenge for an embodied, modal view of cognition, by showing how a paradigm of higher cognition that does not follow the classical LOTH can account for the PSC property. In this chapter, I suggest that by applying the principles of Cognitive Grammar at the level of discourse (and not only words and sentences), one could also provide an embodied predictive processing story for paradigm formal-symbolic competencies related to logic, like deductive inference.

Keywords: cognitive linguistics; Construction Grammar; deductive inference; embodied logic; predictive processing

7.1. Introduction

The focus of Cognitive Grammar, so far, has been predominantly on the sentence and sub-sentence levels, as those are most fundamental for linguistics. Cognitive Grammar has also, so far, focused mainly on natural language and has not yet dealt with formal languages, like logics. A natural next step is, therefore, to extend the representational structure of CG with its abstraction gradient and consider suprasentential patterns, i.e., text and discourse-like structures. In this section I will develop the suggestion that logical inference patterns are very general and abstract linguistic representations, i.e., form-meaning pairs, on a discourse level. In this way, so I argue, we can extend the CG approach to, for instance, deductive inference, another example of a higher cognitive capability that has been tightly associated with
amodal representations. This approach leads us to a cognitive account of deductive inference that does not rely on such amodal mental representations.

7.2. Discourse construction grammar and the dialogical nature of logic

7.2.1. Discourse construction grammar

There are already proposals within Construction Grammar (CxG)\textsuperscript{81} to establish a "grammar of dialogue". As a reminder, the central idea of CxG is that linguistic knowledge is represented as an inventory of "constructions" (the "construct-icon"), which are patterns extracted from language use. They occur on many levels of abstraction and complexity (e.g., morphemes, words, phrases, idioms) and form flexible hierarchical networks. Constructions correspond to linguistic representations as form-meaning pairs. Some attempts to develop a "discourse construction grammar" have aimed at explaining sentence-level phenomena that are influenced "top-down" from the discourse or genre levels. For instance, Östman (2005) has proposed such an extension of CxG from the level of sentences (and below) to discourse (see also Ruiz de Mendoza Ibáñez & Gómez Gonzalez, 2014; Du Bois 2014). Here is an example of such top-down influences from discourse to sentence level. The fact that a text is a newspaper text sanctions sentences like "Mother drowned baby" acceptable. In other contexts, such a sentence would be considered ungrammatical. Another example is the text-type of a recipe. The text-type of a recipe implies mutual expectations of a reader between the content and its format. For instance, a recipe has a characteristic format with an initial list of ingredients and then a list of instructions. Another example is fairy tales, which have the characteristic opening and closure "once upon a time" / "happily ever after".

To establish that inference patterns are constructions, we need to show how they are form–meaning pairs. As to the form, this is quite straightforward, and follows the standard "formal" understanding of inference patterns. Modus ponens has the characteristic form: "If A then B. A. Hence B", where A and B are slots for sentences or propositions. Therefore, the rest of the section deals with fleshing out the meaning

\textsuperscript{81} I do not distinguish here between Cognitive Grammar and Construction Grammar. Langacker says that "Cognitive Grammar is most closely akin to Construction Grammar" (Langacker 2017, p.263)", and indeed, both rely on the same representational structure, the main difference being that Cognitive Grammar has focused more on fleshing out the conceptual meaning of grammatical structures.
of inference patterns. The meaning, e.g., of modus ponens is represented, so I argue, as a concept unit in the hierarchical PP model. MODUS PONENS is as much a concept as CAT or DEMOCRACY. The key ideas is that the meaning of MODUS PONENS arises out of experienced dialogical situations which we can easily take on board through the extension of CxG to the discourse level.

7.2.2. The dialogical nature of logic

The strategy of getting at the meaning of inference patterns and establishing those patterns as constructions involves considering logic grounded in specific dialogical situations. In this way the form-meaning pairs of inference patterns arise as experiential generalizations of observed linguistic behaviour. The meaning then is determined by the dialogical situation. As already pointed out, logic is often treated as corresponding to the realm of abstract mathematical objects, existing independently from mind and agent. Therefore, let me first establish how logic might be grounded in linguistic dialogical behaviour. For that purpose, I follow Dutilh Novaes (2012) in her view that logic has a dialogical nature and origin, i.e., it is ultimately a discourse phenomenon.82

We generally do not reason by strictly following the canon of classical logic and often violate the principles of rationality (e.g., Tversky & Kahneman, 1974; Dutilh Novaes, 2012). Rather, for reasoning we rely significantly on background information and prior beliefs, and we are influenced by many biases. This motivates the view that logical inference is a competency that is learned as a cultural artifact:

"[...] the exact form of the deductive approach to reasoning and arguing must (at least in the majority of the cases) be learned upon intensive training [...] (Dutilh Novaes, 2012, p.156).

Dutilh Novaes stresses that, historically, logic arises from debates and dialogues. A deduction is an argument, i.e., a type of discourse, not an inner mental process. More specifically, it is an argument put forward by a debater to compel other debaters to accept the conclusion of the argument if they accept the premises:

82 See also James Trafford (2017, p.vii): "[...] far from seeing logic as floating free from the trappings of this world, I argue that it should be fully embedded within themes of social structures, agents’ bodies, and the power relationships between us.”
"...the multi-agent dialogical situation would be the true locus primus of deduction. Deduction in mono-agent situations—both in linguistic situations (e.g., an agent conducting a mathematical demonstration) and in non-linguistic situations (e.g., an agent performing a mental process of reasoning)—is in fact a derivative notion.

(Dutilh Novaes, 2012, p.155).

It is quite plausible, therefore, that syllogistic patterns and logical rules are also schematic representations distilled from situated language use, i.e., schemas on a discourse level. In the case of a deductive rule those capture entire patterns of sentences expressing the premises and the conclusion in the context of some discourse situation in which a real or imagined opponent must be convinced. Such schemas correspond to steps that compel the opponent to accept the conclusion granting the premises. My suggestion then, following Dutilh Novaes, is that logical rules, like syllogisms and modus ponens, are conceptual representations capturing successful moves in such a dialogical situation. They are represented by concept units in the hierarchical generative model in the brain. Given their abstractness they are located high in the conceptual hierarchy, i.e., they are very schematic concepts.

7.3. The embodied nature and perspicuity of logic

Regarding the meaning of logical inference rules, one further issue needs to be addressed. Logical inference rules are argument constructions, which correspond to a successful strategy in convincing a critical adversary. That such rules lead to inescapably convincing results is because such arguments represent sufficiently small deductive steps that are maximally perspicuous. Now, what is the basis of this perspicuity? What is it that guarantees that following an inference rule, the adversary is inexorably compelled to accept the conclusion if she grants the premises? The meaning seems to require a source of universal irresistibility. I argue that, following the dialogical approach to logic, this element is grounded in the appeal to perception and experience:

Greek mathematics reflects the importance of persuasion. It reflects the role of orality, in the use of formulae, in the structure of proofs, and in its reference to an immediately present visual object. [my emphasis] (Netz, 1999, p.297f, cited in Dutilh Novaes & French, 2018).
The reference to the visual grounding of proofs has a connection to Lakoff & Nuñez (2000), who suggest that symbolic logic is grounded in "image schemas". Image schemas are sensorimotor based representations capturing generalizations from experience. Lakoff & Nuñez suggest that the meaning of logical rules is grounded in sensorimotor representations. I suggest that the representations they appeal to do not provide the meaning (the meaning is as abstract as suggested in the previous section83), but rather the perspicuity. Let me explain.

Lakoff & Nuñez suggest that logic is based on image schemas that conform to "spatial logics", i.e., it is grasped in terms of objects, containers, and containment relations. For instance, the law of modus ponens is grounded in the following way. Given two container schemas A and B, and an object X, if X is in A and A is in B, then X is in B. Lakoff & Nuñez flesh out the relation between this visual-spatial object-container pattern and the logical modus ponens pattern as a "source domain-target domain mapping", where the target is an abstract concept, and the source is a concrete and experiential concept. Source-target domain mappings could be seen as categorization relations in a hierarchical model of concept units (as described in Chapter 9). The object-container pattern concept (short: OC)—the source concept—is a privileged, prototypical instantiation of a more schematic concept of which MODUS PONENS (short: MP) is also an instance. Let us call us this more abstract parent category MODUS PONENS* (short MP*). OC and MP are child nodes of the schematic, ineffable superordinate concept MP*, which sanctions an isomorphism between OC and MP and therefore allows for using the more concrete OC instead of MP in reasoning and communication. OC has the advantage that it can be made public and shared by joint attention thanks to its visual-spatial nature. Switching to visual-spatial (source) concepts is therefore a good strategy for the purpose of compelling oneself and the opponent to accept the conclusion given the premises. The spatial relation pattern of objects and containers is the ground for immediate perspicuity and for the acceptance of the modus ponens. Often, Venn-diagrams are used to explain, understand, or verify, for instance, a modus ponens reasoning; they

83 In Chapter 9, I discuss Lakoff's "Conceptual Metaphor Theory" (CMT) to which I allude here. A core claim of CMT is that we need mappings with concrete perceptual or motor schemas to understand abstract concepts. I do not deny the importance of those mappings but propose that they are not necessary for understanding or grasping abstract concepts. Rather the role of such mappings is to enhance cognition and communication.
are two-dimensional visual-spatial instances of OC. OC is the vehicle for the acceptance of the modus ponens move. In sum, the meaning of MP—as fleshed out in the previous section—is isomorphic with OC, and the appeal to OC is the source of the perspicuity of MP.

7.4. Syllogism as an example

To illustrate how symbolic deductive inference can be emulated in the PP model proposed here, consider now a simple syllogism (S): All men are mortal (P1). Socrates is a man (P2). Hence Socrates is mortal (C). We need to spell out the deductive inference in terms of an error minimizing prediction process in the hierarchical structure of concept nodes described above. We posit an abstracted schema syllogism in the form of a concept unit that represents the syllogistic inference pattern grounded in the dialogical situation as described above. I suggest that deduction can be understood as error-minimizing co-activations of the syllogism concept node and various subordinated child nodes that instantiate fillers for the schematic slots for the premises and conclusion. Let us distinguish between comprehension (recognizing that a pattern of propositions conforms to a syllogism) and production (inferring the conclusion from the premises under the syllogism scheme) and spell out how each case works.

a) Syllogism comprehension: Grasping the syllogism (S) consists of categorizing the three sentences, the two premises P1 and P2, and the conclusion C, under the schematic, higher-level construction SYLLOGISM. The brain needs to instantiate a concept unit at the level superior to the level of the premises and conclusions as a "prior". In the cognitive process, in PP terms, attention is focused on the three sentence-representations (i.e., we have a high precision estimate in the PP model for them). Now "sense" needs to be made of the occurrence of exactly those three sentences in the given situation. In other words, the sentences are held fixed and an adjustment of the prior, i.e., of the top-down expectation, needs to be made such that the three sentences are "expected" or predicted. In Bayesian terms, the multilayer PP model aims to derive "a cause" of the occurrence of the three sentences, i.e., the activation of a superordinate concept unit. Various "hypotheses" are available one step up in the hierarchy that might predict or explain the occurrence of the three
sentences. In the specific context in which the speaker utters the three sentences, not all the hypotheses have the same probability. The hypothesis on the superordinate level that generates the lowest prediction error in the specific discourse situation is SYLLOGISM. Other possible hypotheses (e.g., THREE-SENTENCES-IN-THE-SAME-LANGUAGE, SENTENCES-ABOUT-HUMANS) are discarded. They do not minimize the prediction error in the given context and are hence not relevant for the context and have low explanatory value.

b) Syllogism production: In the case of the production of a syllogism, i.e., the generation of the conclusion C, from P1 and P2 and SYLLOGISM, the following hypothesized process is being carried out in the PP model. Representations of the two premises P1 and P2 are activated, as well as of SYLLOGISM. Those are held fixed (again, they receive a high precision estimate). However, the prediction error is not minimized as SYLLOGISM "predicts" a pattern that is not actually instantiated yet (the "conclusion slot" is not yet adapted, and C is not yet instantiated). Therefore, the error signals keep adjusting the representation related to the conclusion slot until an error minimizing fit occurs. It is the syllogism conclusion C that achieves the minimal prediction error.

In conclusion, with the help of the hierarchical architecture of increasingly schematic concept representations, we can emulate symbolic deductive inference in the PP model. This requires positing concept units high in the hierarchy that correspond to inference schemas, which are extracted as abstractions from language use in specific (adversarial) dialogical situations.

7.5. Conclusion

I have suggested that we can generalize the Construction Grammar approach to the supra-propositional level and in this way establish that logical inference patterns are constructions on a discourse level derived from experience with specific types of dialogical situations. Those logical constructions correspond to concept units at higher levels of the representational hierarchy of the mental model. Logical deduction can then be implemented as inference in a hierarchical generative model driven by error minimization. While Construction Grammar shows how language and
logic could be accounted for in the predictive processing framework, the latter could be seen as an account of neuro-computational implementation of the former.
Conclusion of Part 3

In Part 2 I have sketched an account of compositionality (and the PSC property in general) for conceptual thought and language, as well as for logical thinking. With this I hope to have addressed the concern that PP does not look sufficiently compositional and is hence not able to deal with higher cognition.

In Chapter 6 I have discussed the challenge from productivity, systematicity and compositionality both for language and for conceptual thought for PP. The central idea is that what motivates the concern that PP might not be able to be scaled up to language and conceptual thought, is our notion of compositionality derived from common-sense generative grammar. Once we leave GxG behind and endorse CxG, we get a new way of understanding productivity, systematicity and compositionality and we get a new LOT, LOT*. LOT* uses not GxG, but CxG as a language paradigm. I have made a suggestive analogy regarding the representational structure and processing principles between PP and CxG. This analogy supports the idea that PP partners well with CxG. Hence PP promises to be for CxG what LOTH is for GxG, i.e., a cognitive computational underpinning. In this way, we have gone some way toward dissolving the LOTH-based concern from higher cognition for PP.

In Chapter 7 I have put forward a PP account for how we think in formal logical terms. The idea is that logical rules are constructions that are derived from observed dialogical behaviour in which an adversary is inevitably compelled to accept a conclusion once she accepts the premises. In other words, logical rules are constructions on the dialogical level. With this move, logic fits squarely into a CxG picture.

The main message from Part 3 is that to appreciate the potential of PP with regards to higher cognition, a paradigm change for language is necessary. Once we embrace such a paradigm change, the challenge from higher cognition based on concerns related to compositionality does seem to lose its force. It was not the purpose to map out in full detail how CxG could be implemented with PP. I have merely provided a high-level and conceptual argument to the effect that by switching the language paradigm, PP looks less problematic. Note that, as already emphasized, PP itself is far from fully specified. A maximally detailed mapping
project would also previously require that the general PP apparatus were spelled out in more detail. Therefore, this part, unavoidably, has a programmatic flavour.
Part 4 - Applications

Introduction to Part 4

In Part 4, I will apply the PP account of concepts and language to various interesting phenomena that have been discussed in the literature in a range of disciplines like philosophy of language, psychology, and linguistics. My discussion is conceptual and schematic, so I am not trying to define formal models or detailed neural circuits. I intend to show that the conceptual apparatus of PP allows us to shed light from a different perspective on those outstanding problems.

I have chosen those phenomena because of their general interest in a wide range of disciplines (linguistics, psychology, philosophy, and neuro- and cognitive science) and the very general implications they have for the understanding of thought. Those phenomena are not yet fully understood and, as of now, various competing accounts for each are on the market.

My motivation for Part 4 is the following. Introducing an account of higher cognition (Part 2 and Part 3) based on a new cognitive paradigm, should not be only a redescription of existing accounts with a new vocabulary. It should shed new light on relevant standing problems. The more challenging the problems the better they can serve as test cases.

Note, that I will not need a fully worked out multi-level-of-description model of higher cognition. For instance, how exactly errors are minimized is not relevant for the level of analysis that is relevant for the current purposes. The commitments inherent in the paradigm (as opposed to a detailed theory) are sufficient. However, I do aspire to contribute to the debates by providing new perspectives (but not necessarily a definite solution). Even if PP is ultimately wrong, or my proposed accounts of semantic paradox, metaphor and copredication within PP turn out to be wrong, I will still be satisfied if the PP account can say something novel and interesting about those problems.

Semantic paradox (the Liar Paradox)

The first problem—which I tackle in Chapter 8 with the toolbox of the PP paradigm—is the family of semantic paradoxes. The Liar Paradox is arguably the most famous
and most discussed among those paradoxes. The huge number of solutions proposed is impressive in itself; it shows how we can come up with dozens of interesting approaches to the same problem. The Liar Paradox is considered a *deep* paradox (as opposed to more shallow "cocktail paradoxes"). We still have no agreement about how we should look at it and it seems to point to some fundamental inconsistency with our concept of truth or the classical common-sense logic we use in thought. Most solutions assume that the Liar is a formal puzzle, so they are very technical and formal solutions that often suggest that we need a modified logic, not classical, common-sense logic. I do not want to just add yet another such formal approach. My take on the paradox is completely different. I do not view the Liar as a problem in formal logical systems, but as an issue where cognitive processing fails. I assume the following: "No thought, no paradox". If there is a paradox, something must be wrong with our thinking. The world is not paradoxical (it is simply as it is). Therefore, I will use the PP account of higher cognition to have a fresh look at the Liar Paradox from a different angle. This requires, beforehand, saying more about linguistic representations (and expanding somewhat the account of language sketched in Chapter 6), and how linguistic and non-linguistic representations are linked. Note that I do not aim to resolve the Liar Paradox; rather I want to explore it through the lens of PP as a problem with cognitive processing.

*Metaphor comprehension*

Metaphor comprehensions is another phenomenon awaiting a better understanding that I tackle in Chapter 9. It has been a tremendously popular topic in philosophy, linguistics, and psychology in the last decades. Metaphor is a deep phenomenon as well, firstly, because it is considered a signature phenomenon of human creativity. For instance, metaphors have helped in our scientific understanding of the world, but also produced pearls of human creation in literary contexts. But it is deep for another reason. Metaphor is considered by many, e.g., by the influential Cognitive Linguistic movement, as fundamental in conceptual thought, not just a poetic decoration or some helpful stylistic device. According to this view, going back to Lakoff, Johnson, and others, we need metaphorical mappings to understand abstract concepts. Not everyone agrees and there is an ongoing debate on how metaphor comprehension works. There are two main camps: those who see metaphors as analogical
mappings, the "Implicit Comparison View," and those who see them in terms of a categorization process, the "Category Inclusion View". Hopefully, by grounding metaphor research in a fruitful cognitive paradigm, we can progress in the understanding of how metaphors are represented and processed in the mind and if they indeed play a fundamental role especially where abstract thought is concerned.

Copredication

Finally, I will examine (with Guido Löhr) copredication. This is a linguistic phenomenon that is increasingly discussed in the linguistic, psychological, and philosophical literature. Copredication sentences use a single nominal for two predicates that require incompatible senses. In "The book is heavy and informative", 'book' is first used in the sense of a physical object and second in the sense of informational content; but those two senses are strictly incompatible as one is concrete and the other abstract. Copredication provides a puzzle for truth semantics. How can the same word refer at the same time to two incompatible entities (a physical object and abstract content)? We propose to focus on an account of how acceptability intuitions for copredication sentences arise in the first place. By leveraging PP, we hope to improve on a previously proposed psychological account for copredication.
Chapter 8. The Liar Paradox in the predictive mind

Abstract

Most discussions frame the Liar Paradox as a formal logical-linguistic puzzle. Attempts to resolve the paradox have focused very little so far on aspects of cognitive psychology and processing, because semantic and cognitive-psychological issues are generally assumed to be disjunct. I provide a motivation and carry out a cognitive-computational treatment of the Liar Paradox based on a cognitive-computational model of language and conceptual knowledge within the predictive processing (PP) framework. I suggest that the paradox arises as a failure of synchronization between two ways of generating the Liar situation in two different (idealized) PP sub-models, one corresponding to language processing and the other to the processing of meaning and world-knowledge. In this way, I put forward the claim that the Liar sentence is meaningless but has an air of meaningfulness. I address the possible objection that the proposal violates the Principle of Unrestricted Compositionality, which purportedly regulates the conceptual competence of thinkers.

Keywords: Liar Paradox, predictive processing, semantic paradox, unrestricted compositionality, Yablo Paradox

8.1. Introduction

Consider the following version of the Liar Paradox. Mary says, I am now lying. If what Mary says is true, she must speak falsely, and if she speaks falsely, she must be telling the truth. So, we have a contradiction, or we must assume that what Mary says is neither true nor false. The extreme simplicity of its statement and the concern that the paradox might reveal fundamental inconsistencies in our basic intuition about logic, language, and the concept of truth (see Beall, Glanzberg & Ripley 2019, Section 3) have led to an enormous number of attempts to resolve it.

Many of the proposals to resolve the paradox try to argue that the Liar sentence is meaningless. If the Liar sentence is meaningless, then any reasoning with it is
pointless. Therefore, the paradox is blocked. An alternative approach to resolving the paradox is to modify classical logic and semantics. For instance, one might declare that the Liar sentence is truth-valueless, i.e., it falls into a gap between true and false (paracomplete approaches, e.g., van Fraassen, 1970; Kripke, 1975) or has a third truth-value; or—quite counter-intuitively—that it is both true and false (paraconsistent approaches, e.g., Priest, 1984, 2006). Barwise and Etchemendy (1987) proposed another account based on non-classical logic. They construct a semantics based on non-standard set-theory, namely, *hyperset theory*, in which sets can contain themselves.

The first sort of solution, via the meaninglessness of the Liar sentence, is often unsatisfying because of either a lack of independent motivation for declaring the Liar sentences meaningless or because the solution comes at the price of highly counterintuitive consequences. For instance, Tarski suggests that the truth predicate is stratified and can only be applied from a meta-language level to an object language level. However, in natural language truth does not seem to work that way; the condition that no language can contain its own truth predicate seems too restrictive. There is, for example, no problem with an honest person saying "I am always telling the truth." Also, Etchemendy and Barwise’s solution leads to strange implications. Their account is based on hyperset theory, which presupposes a notion of self-containing sets (hypersets are no longer conceptualised as collections of things, but as directed graphs allowing for cyclical relations). This has the weird implication that we cannot make statements about the entire world (1987, p. 174).

On the other hand, to easily give up classical logic—as the second main type of solution suggests—does not seem to be a good idea either, given its success as both common-sense logic and the logic of science and mathematics (see also Williamson, 1996; Leitgeb, 2007, p. 283). It seems astonishing that logic would have to be replaced in the domain in which the Liar is situated: a quite trivial domain of everyday life (with lying and truth-saying people). Solutions with gaps or third truth values also suffer from the ‘revenge problem’. One can reformulate the Liar sentence in a way that makes it reappear in a different form. For instance, if Mary says *I am either lying or saying something that is neither true nor false*, then this sentence is also subject to the same sort of paradoxical reasoning with regard to the question of whether it is true or false or neither true nor false.
The discussions and analyses of the Liar Paradox in the literature along the above lines implicitly assume that it is a formal logical-linguistic problem. Logic, language, and concepts like TRUTH are further treated implicitly, like abstract, mind-independent objects and cognitive-psychology has played no significant role in any influential solution attempt. Maybe Leitgeb’s (2007) requirement that "[...] ultimately every successful philosophical theory of truth has to stand the test of formalization [...]" (p.276) best summarises the formalistic tradition in regard to resolving the paradox. There are some exceptions, though, with regard to the treatment of the Liar Paradox as a formal semantic paradox. Most notably, Martinich (1983) suggested that the paradox is pragmatically based (e.g., p.64) in the sense that the speaker fails to execute a proper assertoric speech act when uttering a Liar sentence. However, Martinich does not explain on independent grounds why the speech act fails (beyond the fact that it produces a paradox) and does not provide any discussion on cognitive-computational processes. Also, there are some recent psychological approaches to the Liar Paradox in the context of experimental psychology. However, those treatments are purely descriptive of how participants actually judge Liar sentences in terms of truth value assignments (e.g., Elqayam, 2006; Ripley, 2016) and they do not pretend to ‘resolve’ the paradox, nor do they discuss cognitive-computational processes. Rips (1989b) is an exception and proposes a “computational model” (p.90) for how we reason with liar-type sentences. According to him, we carry out inference processes that are formalizable by a Gentzen-type natural deduction framework. Kearns (2007) follows Martinich’s speech act approach but formalizes it into a system of “illocutionary logic”. Rips and Kearns, therefore, remain in the formalists’ territory.

I do not deny that formalisations are desirable and useful. However, one could suspect that this requirement has some cognitive origin itself. Our ‘desire’ to formalise could be some unconscious norm or bias, which strongly penetrates our scientific thinking. Furthermore, empirical findings from cognitive neuroscience support the idea that the way we think is shaped by the specific architecture of the brain, the body, and the environment. Thinking is not (only) a process carried out via language-like syntactic operations with amodal symbols (see Fodor’s (1975) “Language of thought”). Rather, thought relies on re-enacting sensorimotor states involving the activation of modality-specific representations in the brain (e.g., Hoenig
et al., 2008; van Dam, van Dongen, Bekkering & Rueschemeyer, 2012; van Dam, van Dijk, Bekkering and Rueschemeyer, 2012). But if this view of cognition is in the right ballpark, then there might be fundamental doubts about the possibility of full formalisations of language, logic, and concepts, because formalisations rely on amodal symbol systems.

Therefore, this chapter reconsiders the Liar Paradox by providing a cognitive-computational treatment. I suggest that this approach allows for independently motivating the meaninglessness of the Liar sentence and for retaining classical logic. It explains the paradoxical feel of the Liar and avoids other counter-intuitive implications. My plan is as follows. In Section 8.2, I will provide a motivation for a cognitive-psychological treatment of the Liar Paradox (as opposed to the prevailing “formalist” treatments). Then I will briefly lay out a cognitive-computational model for language and concepts within the relatively new, but already well-established, predictive processing (PP) paradigm (Section 8.3). In Section 8.4, I will explain how the paradox might arise cognitively within the PP model and how we could motivate the meaninglessness of the Liar sentences. In Section 8.5, I will discuss a critical possible objection involving the Principle of Unrestricted Compositionality of concepts, and in Section 8.6 I will conclude.

8.2. Motivating a cognitive-psychological treatment of the Liar Paradox

Why might it be worthwhile to tackle the Liar Paradox from a cognitive-computational perspective, as opposed to taking it as a formal linguistic-logical puzzle with a solution to be sought inside a formalised system? One answer is that the cognitive access to the Liar situation and resulting paradox—i.e., the grasping and appreciating them—is mediated by language and cognitive processing. We are confronted with a Liar situation in which a particular sentence, the Liar sentence, is uttered. Only under the assumption that the Liar sentence expresses immediately, directly, and transparently a mind-independent proposition (i.e., some abstract entity) could we avoid the intermediate step of some cognitive and language processing. However, such a view is not very plausible. It might be that, ultimately, the paradox is inherent in the way the mind-independent world is structured and we merely ‘discover’ the paradox. But it seems more plausible that the paradox arises from how we represent the world conceptually and linguistically and how we reason. If there is
cognitive mediation, then there is at least a possibility—that we should not easily ignore—that the paradox arises because something goes awry with the cognitive processing. That much, so it seems, must even be admitted by die-hard referentialists (e.g., Fodor & Pylyshyn, 2015). For referentialists, semantics depends only on the (mind-independent) referents of concepts, which are established, for instance, via some causal connection between the referent and the mental representations of those concepts. For referentialists, cognitive content is no more than some “[...] aura of associations, attitudes, feelings, beliefs, quasi-beliefs, recollections, expectations [...]” (Fodor & Pylyshyn, 2015, Chapter 2 and p.146) associated with a referent.

A second way to justify why looking at the Liar Paradox from a cognitive angle might be valuable is as follows. There is already a huge number of solution proposals available. Those solutions all produce some counter-intuitive consequences. Let us assume that there is a solution, and whatever the solution is, we have to pay some price in terms of counter-intuition. This is not implausible; consider, for example, the fact that one of the most successful scientific theories in history in terms of experimental support, quantum physics, is highly counter-intuitive. Now ‘intuition’ is very much a cognitive-psychological notion. Why don’t we turn, therefore, to cognition to try to adjudicate a resolution to the Liar based on some understanding of how the counter-intuition arises? Here we would not reject an existing formalistic solution, but instead would use a cognitive approach as a complementary tool for adjudication.

A third possible answer goes far beyond this minimal concession, which even referentialists could make, and points to the paradigm of embodied cognition. According to this paradigm, conceptual representations, meaning and thought are grounded in the sensorimotor experience with the world and are hence shaped by the specific characteristics of our body and brain. Even iconic examples of formalistic disciplines, like formal logic, mathematics, and grammar, are grounded in cognition and not a world of mind-independent objects. Thought cannot be reduced to the syntactical processing of amodal symbols. Let me provide three examples to illustrate the idea that formal symbol systems are cognitively grounded.
Firstly, Dutilh Novaes (2012) suggests that formal languages, like logics or mathematically expressed theories in physics, are cultural artefacts. Their function is to de-bias thinking, especially in scientific contexts. Since the work of Tversky & Kahneman (1974), it has been well-known that the way humans actually think is strongly biased and violates many principles of rationality. Reasoning with the help of formal languages, according to Dutilh Novaes, is externalised “sensorimotor engagement” (Dutilh Novaes, 2012, p.162) with symbols, applying certain rules mechanically without taking into account the meaning of the symbols. One way to spell out what it means to be ‘formal’ is via the notion of de-semantification. De-semantification allows us to “switch off” cognitive content that is automatically associated with concepts (“semantic activation”) and hence to suppress biases, for instance, in the form of prior beliefs (Dutilh Novaes, 2012, p.206–207). The mental manipulation of symbols is a “pushing around” of those and involves modality-specific (sensorimotor) areas of the brain. Dutilh Novaes further observes that writing in a formal language is not an a posteriori expression or description of thoughts and cognitive phenomena. Instead, the formalism is a vehicle of thought; languages are cognitive tools that enhance cognition. Furthermore, Dutilh Novaes takes an explicit position with regard to the object of formalization. It is not a portion of the world, but theories or concepts that we have. Hence what is formalised are intensional objects (e.g., Dutilh Novaes, 2012, p.224). For example, Peano arithmetic is a formalisation not of a series of numbers but our theory of it. With this characterisation of the object of formalisation, we can avoid the question of whether numbers or logics exist independently of the mind or are merely mental constructs. All we are theorising about are the ideas and notions of numbers that humans have, not numbers themselves. Dutilh Novaes’ position allows us to remain agnostic with regard to difficult ontological issues.

As a second example, take Lakoff & Núñez (2000). They suggest that mathematics and logic are not disciplines constituted by abstract, mind-independent objects and truths. Rather, all mathematical concepts are embodied and cultural artefacts, i.e., “a product of the human mind” (Lakoff & Núñez, 2000, p.9). Central to Lakoff and Núñez’ account is the idea of “image schemas”, which can be understood as a basic modality-specific mental representation. Visual-spatial image schemas correspond, for instance, to concepts expressed by prepositions like in, on, at or above. The
“Container Schema”, (which is a gestalt consisting of some boundary, some inside and outside) is central to mathematics (on which, e.g., set theory is grounded). They are both conceptual and perceptual (Lakoff & Núñez, 2000, p.31). Take a branch of mathematics, arithmetic, as an example. According to Lakoff and Núñez, we conceptualise the objects and principles of arithmetic in the form of “conceptual metaphors”. One such metaphor is ARITHMETIC IS OBJECT COLLECTION. We understand the concepts and principles in the “target domain” ARITHMETIC by transferring them from the “source domain” OBJECT COLLECTION. The source concepts stem from experience, like taking away or adding objects from collections. From correlations between manipulating object collections (which are sensorimotor operations) and arithmetic operations arise neural connections that constitute a “conceptual metaphor at the neural level” (Lakoff & Núñez, 2000, p.55). Image schema like the IN schema have modality-specific “spatial logics” on which amodal, formal symbolic logic is grounded. For instance, the law of modus ponens is grounded in the following way. Given two container schemas A and B, and an object X, if X is in A and A is in B, then X is in B. But it is not only neural visual-spatial areas, but also the motor control system, that are involved in mathematical conceptual thought (Lakoff & Núñez, 2000, p.34–35).

As a last example, take Langacker’s Cognitive Grammar which applies the idea of the embodiment of cognition to grammar. Contrary to the traditional view, Langacker (e.g., 2008) denies a formalistic view of grammar, the idea that it can be represented exhaustively in the form of an amodal symbolic system:

The picture that emerges belies the prevailing view of grammar as an autonomous formal system. Not only is it meaningful, it also reflects our basic experience of moving, perceiving, and acting on the world. (Langacker 2008: 4)

Langacker’s surprising claim is that grammar is meaningful, and that grammar and lexicon differ only in degree, not nature. He motivates this view via examples of “partially schematic units”, which can be classified neither as paradigmatically lexical in nature, nor as pertaining exclusively to grammar. The following linguistic schema: Vs X IN THE Nb (where Vs is a verb meaning ‘striking’ like hit, kick, strike, or poke and Nb is a body-part noun like shin, back, face, eye, or knee) is an example of such a partially schematic unit (Langacker, 2008, p.20). Now, according to Langacker, (e.g.,
2008, p.5) grammar is meaningful because it is “symbolic”, much like lexicon. A “symbol” is a pairing of a “semantic representation” (which is a complex embodied conceptualisation, much along the lines of Lakoff & Núñez (2000)) and a “phonological representation” (including as well gestures and orthographic representations). Grammar and lexicon form a gradation in terms of specificity/schematicity. Grammatical symbolic units are just more schematic than lexical ones, but they are not different in nature.

All three examples show how cognition takes centre-stage when analysing formal systems. If they are in the right ballpark we should consider formal systems like logic, language or mathematics not as mind-independent abstract amodal symbol systems but as grounded in modality-specific representations. Maybe the deadlock in resolving the Liar Paradox via the formalistic approach is due to neglecting the embodied cognitive basis of formal systems. Therefore, it seems at least worthwhile to explore a cognitive approach to the semantic paradoxes and see where it leads us. Such an approach requires a specific cognitive computational model and an account of concepts and language, which I will outline in the next section.

But before that let me address the following possible objection to an embodied cognitive treatment of matters related to logic. Normally, the nature of logic is expressed as a dichotomy: it is either descriptive of or a norm for rational thought. But, for instance, the principles of classical logic are not descriptive of how in fact we reason, as the literature on cognitive biases teaches us (see e.g., Tversky & Kahneman, 1974, for the classical paper). That leaves us with option two. But if logic is a mind-independent norm then how we actually think seems irrelevant to a treatment of the Liar Paradox. I suggest denying the dichotomy and taking logic to be both descriptive and normative. It is descriptive of what we think the norm of rational thinking should be. Following the idea of Dutilh Novaes regarding what the objects of formalisations are, the analysis here will be of an intensional object (our understanding and ideas of the rational norm), rather than the “norms themselves” (if they exist at all in the Platonic heaven). In other words, we can go agnostic about whether the norm exists independently of the mind, without falling into an unpalatable idealism or solipsism.

84 As we will see later in Section 5.2., the idea of such an internal norm can be spelled out naturally under the framework of a hierarchical generative model.
8.3. Predictive processing and a dual account of language and conceptual knowledge

8.3.1. The predictive processing framework

According to predictive processing (PP) (see Clark 2013, 2016; Hohwy 2013, Friston 2010), the brain is an embodied, multi-level prediction machine that continually anticipates its sensory input, relying on a mental prediction model. The PP model structure has the form of a generative probabilistic model in which approximate Bayesian inference is carried out (e.g., Clark 2013, pp.188–189; Hohwy, 2013, pp.15–39). The PP model has a hierarchical structure and represents prior knowledge on many levels of abstraction (e.g., Clark, 2013, p.25; Lupyan & Clark, 2015). In the top-down prediction cascade, the predictions of higher-level layers serve as priors for the lower-level predictions and, in this way, constrain the hypothesis space on the lower level. The main feature specific to the PP story is that the computations in the brain are driven by prediction error-minimisation and, therefore, the error signals play a central role. The predictions are compared to the actual sensory input in so-called **error units**, and the **residual error** of the predictions is calculated. It is then the error signals that flow laterally and upwards in the hierarchical network and may lead to updates of the model at different levels of the hierarchy.\(^{85}\) The PP model is constantly adjusted in order to converge towards a version that minimises the overall average prediction error in the long run. There are two fundamental and interrelated ways to minimise the prediction error: firstly, by updating the internal model to fit the predictions to the sensory flow; and secondly, by generating actions to fit the sensory flow to the predictions (**active inference**), i.e., the brain with its body can change the world to fit its prediction. In this way, PP brings action, perception, and cognition into a unified framework under a unified (embodied) cognition paradigm. The prediction error minimising process is supplemented by a mechanism of precision-weighting of the prediction errors (Clark, 2016; pp.53–83). The brain needs to discriminate noise and useful signals because noise should not force an update of the model. The brain must, therefore, predict the reliability of the sensory input, assign weights to the error signals and thus determine the influence of the top-down predictions versus updates driven by the bottom-up

\(^{85}\) The mechanism of processing only the error signal is also called **predictive coding**.
error-signals of the model. An example from Clark (2013, p.198) might serve to illustrate this point. In a situation of thick fog, visual sensory information about the shape of an object is less reliable than, e.g., tactile or auditory information. In such a context, the precision-weighting mechanism predicts that the bottom-up visual signal has low precision. Bottom-up error signals related to the shape are then tuned down to avoid an update of the brain's prediction model, and the influence of other sensory modalities or top-down predictions increases.

While the PP model has a unified architecture (often represented as a hierarchical probabilistic graphical model), it is compatible with a certain functional modularity, where a module corresponds, roughly, to some domain. For instance, we can speak of a sub-model corresponding to folk-psychology in which concepts and knowledge are encoded relevant to predicting and explaining the behaviour of others based on desires and beliefs. Another example is folk-physics, a sub-model that encodes common-sense knowledge about how physical objects behave. Those sub-models should not be seen as ‘encapsulated’. Rather, they might arise out of hard-wired biases or be the consequence of the error-minimisation based optimisation of the PP model. Such domain-specific knowledge is still part of a single, large and highly interconnected overall PP model. This is important to keep in mind for the next section.

8.3.2. The Language and Situated Simulation (LASS) Model

The treatment of the Liar that I propose is based on the idea that thought proceeds in a synchronised way in two sub-models (as qualified in the previous section), one for (formal) linguistic processing and the other for processing meaning and conceptual knowledge. This duality of sub-models should be taken as an idealization. There might be more than two sub-models involved in meaning and conceptual processing, for instance, a ‘pragmatic’ sub-model. Such a pragmatic sub-model would encode the knowledge that allows the individual to derive the speaker’s intentions and meanings. For simplicity, those other sub-models are grouped into the sub-model for meaning and the sub-model for conceptual processing. These two idealized sub-models rely on language-like linguistic representations and modality-specific

---

86 See, e.g., Gerstenberg & Tenenbaum (2017) for a discussion of intuitive theories.
87 Thanks to an anonymous reviewer for raising this point.
representations, respectively. The paradox arises—I suggest—as a failure in the synchronisation of the two modes of processing. Several authors (e.g., Paivio, 1990; Simmons et al., 2008; Dell & Chang, 2013) have proposed accounts of thought and language processing that make a dual distinction of representational types. For instance, Dell and Chang (2013) suggest a Dual-Path model for language production and understanding, with a dual structure consisting of coordinated processing on a “meaning pathway” and a “sequencing pathway” (e.g., Dell & Chang, 2013, p.4). I will adopt a version of the account of Simmons et al. (2008) (also Barsalou et al., 2008), however, with a specific PP-twist to it. The authors propose that conceptual processing relies on both language-like and modality-specific representations, in a way that depends on the context and cognitive task. According to their LASS (Language and Situated Simulation) model, in conceptual processing both the linguistic and simulation sub-systems become active. However, activations of the two types peak in different orders, depending on whether there is a linguistic or a perceptual cue: “Once a word is recognized, associated linguistic forms are generated as inferences, and as pointers to associated conceptual information” (Simmons et al., 2008, p.107). Once the linguistic forms are associated, the system can pursue different strategies corresponding to different levels of conceptual depth. The linguistic form (or associated statistical information) might be sufficient for the purpose at hand; therefore, we might have a “shallow activation” of meaning. After the activation of the linguistic form, the brain might also activate modality-specific regions within itself, which generate a simulation of perceptual or motor mental states that would be active if it were to interact with the referent of the linguistic form. Those simulations are “situated” or context-specific and are often quick and automatic (less than 200 ms after word onset) and “deeper” conceptual representations, as opposed to the shallow, merely syntactic, ones (Simmons et al., 2008, p.107). Different tasks imply a different mixture of the two representational forms. The linguistic system represents mere form, while the simulation system represents the “meaning”. The authors stress that speaking of “two systems” is not to imply that there are two rigid modules; rather, it is a “simplification so that we can focus on mechanisms of interest” (Barsalou et al., 2008, p.253). This corresponds to my idealization in the form of two sub-models mentioned at the beginning of this section.
8.3.3. The LASS model and predictive processing

The reason why I endorse the LASS model is that it is fully compatible with the predictive processing framework, and the PP approach provides further unity to the dual LASS account. The PP framework with its precision weighting mechanism also supplies a computational underpinning for the LASS model. Let me briefly develop those claims.

LASS and PP are compatible because we can treat the two sub-systems of LASS as sub-models in the overall PP model that we entertain. Language and its linguistic objects are further ‘things in the world’ that are modelled by the brain. The linguistic sub-system has a certain functional identity as a highly interconnected sub-model and encodes knowledge of a specific domain of things in the world, in this case, linguistic objects, like phonemes, words, and sentences. The PP model has an innumerable number of such sub-models, which are understood as closely interconnected parts of the model that track a certain world domain. For instance, when talking about moving objects, we can say that the sub-models corresponding to the two domains folk-physics and language are active and salient. The linguistic sub-model supplies the formal aspects of language, and the folk-physics sub-model provides the conceptual meaning.

There is also increasing evidence that language production and comprehension are predictive on all levels of the hierarchy of linguistic objects: phonemes, words, sentences, and discourse (see, e.g., Allopenna et al., 1998; Barsalou, 2009, p.1286; Kuperberg & Jaeger, 2016; Kutas et al., 2011; Pickering & Garrod, 2013; see also Gagnepain et al., 2012). Therefore, the linguistic sub-system of the LASS model fits with the prediction-centred paradigm of the PP framework. PP can further provide a computational underpinning for the idea reflected in the LASS model that both sub-systems might receive different emphasis, or attention, at different moments. Clark (2016: 64–65) has suggested that the PP precision estimation mechanism can give control to different areas in the brain. For instance, visual information can be given priority over auditory information in the case that the estimated reliability of the latter is low. Similarly, I suggest that the mind can focus attention on the language sub-model or some other domain sub-model with non-linguistic conceptual content. In one case, one focuses more on formal-syntactical aspects during language comprehension (e.g., when drawing logical inferences content is irrelevant). The
different words (as label-like linguistic objects) function as empty placeholders (in Dutilh Novaes’ terms “de-semantisized objects”) and computations run on the language sub-system only. But one can also focus more on conceptual content and lower the precision estimate for the language sub-system. This allows one to make sense, for example, of grammatically incorrect sentences, by prioritising the prediction of conceptual content, and not being too picky about grammaticality.

For convenience, throughout the remainder of the chapter I will call the linguistic sub-model the LSM and the sub-model(s) of the domains relevant for a cognitive task the WSM (world sub-model). The LSM represents all formal, linguistic aspects (including statistic information about word co-occurrences, for instance). The WSM represents the meaning and conceptual (non-linguistic) knowledge. Again, this is a simplification, and I follow what the LASS authors have pointed out: there is no rigid modularity.

8.4. A cognitive approach to the semantic paradox

With a cognitive model for language and conceptual processing in hand, we now turn to a treatment of the Liar Paradox. The central idea is that we can conceive of a Liar situation in two ways, one corresponding to a simulation/prediction in the world sub-model (WSM) and the other corresponding to a simulation/prediction in the linguistic sub-model (LSM). For a sentence to ‘make sense’, there needs to be a synchronous, stepwise prediction in both sub-models. In a different context, Altmann & Mirkovic (2009) speak of a dual, synchronous “unfolding” of sentence and real-world event representation, which fits nicely with the way I suggest we should consider the relation between the LSM and WSM:

[...] language is not processed in isolation of the world it describes; instead, comprehension consists in realising a mapping between the unfolding sentence and the event representation corresponding to the real-world event that is being described. (Altmann & Mirkovic, 2009, p. 602)

In this way, the WSM can be seen to provide conceptual or categorical constraints with regard to how words can be combined grammatically to create meaningful sentences. I suggest that Liar sentences are meaningless because Liar situations, on very careful reflection carried out in the WSM, cannot arise in the way that is
necessary for a paradox, i.e., a synchronously unfolding prediction in the WSM and LSM fails. But still, the Liar Paradox retains a certain psychological pull, as it looks meaningful. This, as I will show, can be explained by an accommodation mechanism fleshed out via the PP-specific precision weighting mechanism and its multi-level processing architecture. But first I need to establish a basic assumption that is necessary for my treatment of the Liar Paradox, the assumption that meaning is speaker meaning.

8.4.1. An assumption: meaning as speaker meaning

I make one substantial assumption, namely, the Austinian-Wittgensteinian view of meaning as speaker-intended meaning. A defence of this view would exceed the scope of this chapter, but fortunately, the view is endorsed by others as well, and I refer to the literature for further defence (e.g., Azzouni, 2013; Goldstein, 1981, 1982; Rayo, 2013). It is by using a word or sentence in a specific context with a particular intention that the speaker endows this word or sentence with meaning. Without the (at least imaginable) presence of a (conscious, intentional) speaker, sentences are merely dead wiggles and noises. The impression that The cat is on the mat means something ‘by itself’ as a sentence without actually (or imaginatively) being uttered by a speaker with assertive intentions is merely an illusion. Azzouni (2013) explains the illusion that linguistic items mean by themselves through the phenomenon of “semantic perception”. In the same way that we cannot avoid ‘seeing’ the functionality of a screwdriver, we cannot avoid ‘perceiving’ the meaning of a word:

    We don’t experience the presence of a speaker’s intentions—even when we are aware of them—as causing or influencing or determining what is said in these cases. Rather, we simply experience the expression as just meaning this or as just meaning that. (Azzouni, 2013, p.130)

If random wind movements arrange some leaves on the ground in the form of the word cat, it has a meaning only because we can imagine a person intentionally arranging those leaves to form the word cat to communicate something. If all meaning is speaker meaning, the test for the meaningfulness of a sentence—with careful thinking—is a speech act simulation where an agent with a certain intention is uttering the sentence. I speculate that Azzouni’s “semantic perception” involves
unavoidably and subconsciously projecting certain intentions onto a parrot uttering a ‘meaningful’ sentence.

In the view defended here, the place where a person generates and processes meaning is the brain’s WSM. WSM-based meaning processing, when coordinated with the LSM, not only includes the comprehension of the literal (or semantic) meaning of words and individual sentences, but also comprises the derivation of the pragmatic meaning of a sentence, and of the understanding of whole discourses. Pragmatic meaning comprehension involves perspective-taking on the listener’s side, i.e., the simulation of speech-acts and speaker intentions, from the perspective of the speaker (e.g., van Berkum et al., 2008). I also assume that the comprehension of the literal or semantic meaning of sentences requires such simulation-inferences. That might be controversial but is compatible with the Davidsonian view (e.g., 1986) which denies the existence of fixed, conventional (literal) word meanings. Such a view is gaining momentum among philosophers, psychologists, linguists and cognitive scientists (e.g., Barsalou, 2009, 2011; Casasanto & Lupyan, 2015; Ludlow, 2014; Rice, 2016) who take concepts to be highly dynamic and context-dependent entities. The view does not deny that a semantic/pragmatic distinction might be useful, but certainly questions a clear, dichotomic divide. Such a divide would seem arbitrary in the multi-level processing architecture of PP, where inferences are carried out on different levels of abstraction and contextual complexity.

The view that semantic and pragmatic processing are not separate (and sequential) cognitive processes is supported by recent empirical findings that show that meaning processing is immediate and holistic (e.g., Hagoort & van Berkum, 2007; Bašnáková et al., 2014). This finding fits with the PP model, because error minimization can be seen as a holistic and simultaneous process that brings the network on all levels into an overall optimal state. Crucial for the rest of the chapter is the assumption that linguistic processing in the PP model can have different ‘levels of depth’ (I follow Barsalou in the use of the notions ‘shallow’ and ‘deep’ here, see Section 8.3.2). For instance, meaning typically described as pragmatically inferred is deeper than that described as literal meaning in the sense that the former integrates a larger amount of available information, namely, additional contextual and situational cues. Even the processing of what is normally described as literal (or semantic) meaning of a sentence can have different levels of depth. If conceptual knowledge is represented
in the WSM in terms of constraints for concept application, then the level of depth of processing is correlated with how exhaustively all those constraints are being taken into account.

With those clarifications about the notion of ‘meaning’ and assumptions about how meaning is processed at hand, we can now move to the discussion of how the Liar Paradox arises.

8.4.2. Two ways of conceiving of the Liar Paradox

Let us call a scene or situation that leads to a Liar Paradox a ‘Liar situation’. The situation in which Mary utters the Liar sentence, *I am now lying* is an example of a Liar situation. Another, non-self-referential, Liar situation was first described by Stephen Yablo.88 Consider the following infinite chain of numbered sentences, each of which states that all of the sentences after it are false. As can be easily verified, no consistent assignment of truth values can be made to the sentences:

0. All following sentences (1, 2, 3, ...) are false.
1. All following sentences (2, 3, 4, ...) are false.
2. All following sentences (3, 4, 5, ...) are false.
   ...

I suggest that the Liar Paradox arises as a failure of the synchronization between two ways of generating (predicting) a Liar situation. Those two ways correspond to predictions in the sub-models, the WSM and the LSM.

8.4.2.1. Conceiving in the LSM (linguistically focused prediction)

The first way of conceiving of the Liar situation corresponds to relatively ‘shallow processing’ (see also Erickson & Mattson, 1981; Barton & Sanford, 1993). The focus of the mind is on the LSM, not on the ‘deep meanings’ and implicit conceptual constraints contained in the WSM. To show the generality of the approach I will use the above examples of self-referential and non-self-referential Liar situations and try to sketch what is going on inferentially when we conceive of the Liar Paradox in this mode.

---

88 Yablo (e.g., 1985) discovered this version of the liar, the so-called *Yablo Paradox*, which shows that the problem with the liar is not self-reference.
Self-referential Liar situation

We have a person, Mary, uttering a sentence (the Liar sentence). Her sentence is grammatical, and it contains only meaningful words. The subject matter of what Mary utters can be easily identified. She is saying something about the truth of the utterances of an agent. The agent happens to be herself. Also, the sentence does not contain any apparent category mistakes because we can clearly say of persons, including Mary, that they say the truth or lie. The sentence appears to be meaningful. The PP model has generated the sentence with some (shallow, as we will see) involvement of the WSM. In this way, the PP model produced ‘comprehension’ of the sentence. The paradox arises then when we take the apparently meaningful sentence to be an assertion and start reasoning with it.

Non-self-referential Liar situation: Yablo’s paradox

All sentences are grammatical and contain only meaningful words. The subject matter of each sentence is easy to identify. Each sentence is expressing something about the truth or falsehood of the sentence that follows. Also, the sentences do not contain any (apparent) category mistakes because one can say about sentences that they are true or false. All sentences appear to be meaningful. The PP model has generated the sequence of sentences with some (shallow, as we will see) involvement of the WSM. In this way, the PP model produced ‘comprehension’ of the sequence of sentences. The paradox arises then when we proceed and try to assign consistent truth value assignments to the apparently meaningful sequence of sentences.

8.4.2.2. Conceiving in the WSM (semantically focused prediction)

Surprisingly, one very simple fact about Liar situations does not play any role in the literature about the Liar Paradox. On careful reflection it becomes obvious that Liar situations cannot actually arise in reality, at least not in the way that is necessary for a paradox to arise. When I say Liar situations cannot arise in reality, I mean that we cannot conceive—on very careful reflection, taking into account all of the conceptual constraints—of situations where the sentences are being uttered with the right intentions. But this means that they are meaningless—at least under the assumption discussed above that meaning is speaker meaning. The ‘on careful reflection’
corresponds to the second way of conceiving of the Liar situation. It is cognitive
processing of the Liar situation with a strong focus on the WSM, i.e., with an
emphasis on ‘simulating deeply’ (not shallowly) an actual Liar situation. Let me
explain.

Self-referential Liar situation

Mary makes an (apparent) assertion about her trustworthiness. Given this context,
there is, however, no conceivable situation in which Mary would say, I am lying. If
Mary were telling the truth she would say, I am telling the truth, and if she were lying,
she would say that same sentence. Of course, the WSM and LSM can generate a
scene (through a coordinated unfolding of the predictions in each sub-model) where
Mary utters the Liar sentence. However, in the WSM prediction Mary’s intention
cannot be the one required to make the Liar sentence a paradoxical sentence,
namely, the intention to say something true or false about her trustworthiness. As a
consequence, the WSM cannot generate the Liar situation synchronously with the
LSM in the way required to make it a paradox. Therefore, the Liar sentence is
meaningless. To find the Liar sentence meaningful requires being slightly sloppy with
the WSM and ignoring the subtle role of the utterer’s intention. How this ‘sloppiness’
might arise in PP terms I will discuss in a moment.

Non-self-referential Liar situation (Yablo’s paradox)

In the Yablo paradox, there are no agents with intentions. However, I have assumed
that to make sense of those sentences, we need to be able to at least imagine a
situation where intentional agents utter them with the appropriate (assertoric)
intentions. We can transcribe Yablo’s sequence in the following way:

Person 0: ”What persons (1, 2, 3, ...) are saying is false.”
Person 1: ”What persons (2, 3, 4, ...) are saying is false.”
Person 2: ”What persons (3, 4, 5, ...) are saying is false.”
...

A similar line of thought as with the self-referential situations shows that the Yablo
situation cannot be generated with deep (as opposed to shallow) involvement of the
WSM either. To make sense of the sequence of sentences, we need to imagine
them to be uttered by agents with appropriate intentions. To say something true or
false intentionally about a sentence requires knowing the sentence. For instance, if
everyone speaks at the same time, they all make statements without knowing what
the others are saying. So, they cannot make honest (or dishonest) assertions about
the truth value of the others’ sentences. It is also not possible that everyone speaks
in sequence, such that they all refer only to known sentences. The reason is that in
the Yablo situation, the last person should start speaking then. However, there is no
last person in this infinite sequence. Therefore, ‘comprehension’ of the Yablo
situation involves strong attention being paid by the mind to the formal aspects (i.e.,
a strong focus on the LSM), including the recognition that the sentences are in
correspondence with a mathematical structure (an infinite omega sequence). The
apparent ‘comprehension’ of the Yablo situation succeeds only by partially ignoring
constraints from the semantic engine (the WSM). No synchronous generative
unfolding with deep involvement of both sub-models is possible. Hence the
sequence of sentences as a whole (and at its highest level of integration) is
meaningless.

In both cases, the self-referential and non-self-referential, the Liar sentences can be
uttered, of course, but they cannot be asserted, where assertion implies that the
utterer intends to say something true or false. The meaninglessness consists in the
fact that words are put together in a way that cannot correspond to a situation
generated synchronously with deeply (and not only shallowly) involving the WSM.
The WSM therefore provides constraints on how words can be put together
meaningfully to describe whole situations. While the Liar sentence (or a sequence)
as a whole is meaningless, it certainly is not completely meaningless. It is
significantly more meaningful than linguistic forms like: Dsad djd dj hhd or Green go
or having. The multi-level processing of the PP model can accommodate the intuition
that meaning comes in degrees.\(^89\) The Liar sentence contains meaningful words; it
contains meaningful phrases; a subject matter can be clearly identified; it can be
identified as a sentence with a subject and predicate, etc. Cognitive processing
succeeds at those lower levels. It is only at the highest level of integration in the
WSM of all those bits and pieces that meaning fails. However, this failure is enough

\(^89\) Thanks to Andy Clark for this suggestion.
to deprive the Liar sentence of the status of being meaningful as a whole and any truth evaluation is pointless.

8.4.2.3. Accommodation effects: explaining the psychological pull of the liar

I will now explain in more detail how, and in what sense, the Liar sentence appears to be meaningful. This is essential to account for the pull of the Liar Paradox. The pull of the Liar Paradox consists in the resistance to accepting that the Liar sentence is meaningless. (Even I still cannot believe that the Liar sentence is meaningless!)

The brain's model exhibits flexibility in interpreting other speakers of the language community. A precise application of grammar and orthography is not necessary to be able to understand others. We can perfectly well understand grammatically incorrect sentences. We can even deal with words that are used incorrectly—malapropisms (Davidson's classic paper “A nice derangement of epitaphs” (1986) makes this last point). Let me call such effects accommodation effects. Accommodation can be explained in the PP framework via the precision-weighting mechanism in the following way. The PP precision weighting mechanism allows us to tune up or down the influence of the error determined in the error unit corresponding to a certain representation. With this mechanism one can modulate representations (predictions) by switching off or dampening the influence of other representations.90 Accommodation, for instance of a malapropism, occurs when the error signal for that word is ignored or dampened. In this case, top-down influence of the prediction of the right word (which is inferred from the context in which the word stands in the sentence and of the speech act itself) dominates and the prediction succeeds, despite a large error between the predicted word and the word actually uttered by the speaker.91 Metaphors can also be explained by an accommodation mechanism cashed out in PP terms. Metaphors are often taken to be category mistakes, i.e., grammatically correct sentences that are made up entirely of meaningful words but are semantically anomalous. Shakespeare's famous Juliet is the sun is a simple and

90 See Michel (2020a) for a proposal of a detailed mechanism for context-sensitive modulation of concept features with PP precision weighting.

91 Another example of an accommodation process, quite close to malapropisms, is Moses sentences (see also Erickson & Mattson, 1981; or Barton & Sanford, 1993): people tend to answer to the question How many animals of each kind did Moses take on the Ark? with Two. They overlook that it should say Noah.
often analysed example (see, e.g., Schroeder, 2004). However, the PP model can explain metaphors via accommodation. Metaphors can be ‘made’ meaningful by flexibly creating an ad-hoc concept to ‘force a fit’. Features of the concept $\text{SUN}$ are suppressed via the precision weighting mechanism and the ad-hoc concept $^*\text{SUN}$ is created, which can represent a more inclusive and abstract concept composed of features like ‘something central’ or ‘something that generates positive feeling’. Not all category mistakes can be accommodated as metaphors. For instance, *The relativity theory listens to an ill breakfast* seems to be nonsense that cannot be accommodated, i.e., ‘forced’ to be meaningful.92

Accommodation can be applied to the Liar sentence as well. That the Liar sentence appears to be meaningful might be explained in the following way. On the one hand, the sentence is fully grammatical; the agent appreciates this by focusing attention on the LSM, in which the sentence can be easily generated. For the sentence to also be meaningful to the agent requires success through a synchronous, parallel unfolding of the prediction in the WSM while the stepwise prediction in the LSM is going on. While the synchronous unfolding cannot succeed on ‘careful, constrained conceiving’ (i.e., deep as opposed to shallow processing) in the WSM, it can succeed if we raise the error thresholds for the prediction, in other words, if we simulate the Liar scene with a coarser grain, ignoring certain features of the Liar situation. This effect might lead to abstracting away the speaker's intentions. In fact, in most cases we can ignore the speaker’s intentions and consider sentences as having meaning on their own. Our bias towards perceiving meanings as properties of sentences (see the idea of semantic perception described by Azzouni, mentioned before) might be an evolutionary adaptation. In most cases it works. However, on some occasions, like the Liar situations, it leads to trouble.

One might object to the fact that we have left unexplained the exact sort of conceptual constraint that makes the sentences in Liar situations meaningless. But this is only a pressing concern for those who think that we must be able to formalise and make explicit the conceptual constraints in the PP models (which I suggested are encoded in the WSM sub-model). Although Liar situations are quite simple

92 Carston (2002) also suggested that some metaphors can be explained by the creation of ad-hoc concepts where content is narrowed or broadened, until one can ‘make sense’ of the metaphorical sentence.
situations, they involve a complex web of common-sense concepts related to language and concepts like truth and falsehood, as well as folk-psychological knowledge of intentions, etc. It is unclear how those could be formalised explicitly in the way that formalists demand, such that we can filter out all possible Liar situations. In Section 8.5.2, I will delve further into this issue.

Let me stress that the treatment of the Liar Paradox requires the specific PP apparatus. Specifically, the mechanism of precision weighting is central to the explanation. The ultimate purpose of the overall PP model is “to get a grip on the world” (Clark, 2016, p.202), the central tenet of PP. The LSM could be seen as a model of the WSM (and hence as a second-order model of the world) that can be made explicit linguistically. It enhances the grip-getting capacity and efficiency of the WSM in many ways, e.g., by allowing the sharing of predictions via public language or the critical evaluation of predictions that are publicly available (see Clark, 1998, for various other ways applicable also to PP). But the LSM can also lead us astray, as the Liar Paradox shows, if we reason with the LSM with only shallow involvement of the WSM, i.e., by not carefully honouring all of the conceptual constraints within it.

A further conclusion is that the multi-level processing of the PP framework offers resources to deal with partial meaningfulness, which we have encountered in the Liar Paradox. An application of those ideas in non-linguistic contexts also seems possible. For instance, take Escher's famous vexing visual paradoxes. Escher pictures present 'impossible' three-dimensional objects (e.g., Escher, 1971; Penrose & Penrose, 1958). A lot of local and partial patterns can be recognised as meaningful without any problem; however, no stable perception of the whole three-dimensional object is possible. Unfortunately, detailed discussion of this phenomenon needs to be carried out elsewhere. But a treatment similar to the Liar Paradox seems possible along those lines.

8.5. An objection: the Principle of Unrestricted Compositionality

I will focus on what I consider the most serious possible objection to the view that Liar sentences are meaningless. This objection is related to what I suggest calling the Principle of Unrestricted Compositionality (PUC). The PUC holds that all grammatical sentences with meaningful words are meaningful and truth evaluable.
(Specifically, category mistakes come out simply as false under the PUC.) As Liar sentences are grammatical and have only meaningful words, then if the PUC is correct, my treatment of the Liar Paradox must be flawed, as it relies on Liar sentences being meaningless.

The PUC is motivated by the *productivity* and *systematicity* of language. The *productivity* ensures that we can generate infinite novel combinations of concepts. If we can entertain the thought a is G, then we can also entertain that b is G, c is G, etc. If we can entertain that a is F then we can entertain that a is G for every concept of G we have. The *systematicity* allows us to understand that Peter kisses Mary when we understand that Mary kisses Peter. The PUC goes back to Evans & McDowell’s (1982) *generality constraint*:

> If a subject can be credited with the thought that a is F, then he must have the conceptual resources for entertaining the thought that a is G, for every property of being G of which he has a conception. (Evans & McDowell, 1982, p.104, as cited in Camp, 2004)

Evans & McDowell stress that for an agent to competently master a particular concept, she needs to be able to combine it arbitrarily with any other concept. But the PUC has not been the mainstream view. Other writers (e.g., Strawson, 1959; Peacocke, 1992) prefer to impose categorical constraints on concept combinations to avoid ‘category mistakes’ like *Colorless green dreams sleep furiously* (Chomsky, 2002, p.15). The strong intuition is that those are not meaningful sentences; one cannot grasp under which conditions those sentences would be true. The inability to understand what is being said in cases of absurd concept combinations should not undermine the conceptual competence of the agent. But Camp (2004) and Magidor (2013) have recently put forward a whole battery of arguments defending the PUC. According to them, even category mistakes like *Caesar is a prime number* are meaningful and can be truth-evaluated. Let me briefly discuss three of their arguments.

8.5.1. *Argument from (material) inferential roles of category mistakes*

Camp (2004, p.212) suggests that category mistakes have substantial *inferential roles* and, therefore, are meaningful and can be truth-evaluated. She argues for this thesis in the following way. From the supposition that Caesar is a prime number you
can draw the material inference that ‘Caesar is not an efficacious emperor’. This inference, according to Camp, is material, not formal, because it requires using the meanings of the terms involved. If the inference were merely formal, we could replace subject and objects with constants and ignore meanings and draw the inference in a purely mechanical way. The conclusion clearly makes sense; therefore, if one can infer meaningful things from category mistakes, then category mistakes cannot be meaningless.

However, this argument begs the question. One can only make the presupposition ‘Caesar is a prime number’ if that sentence is meaningful. But that is what Camp wants to show. Now, maybe she might respond that I beg the question by taking the sentences as meaningless and for that reason reject the idea that it can be taken for a supposition for the sake of argument. So, I first have to show on independent grounds that the sentence is meaningless. However, even if we could presuppose that Caesar were a prime number, what Camp calls a substantial material inference is nothing more than what could be called inferential luck because the conclusion is meaningful for the wrong reason; also, the inference turns out to be formal and not material (i.e., meaning involving) after all. To see this, let us spell out the inference in more detail (this time assuming, for argument’s sake, that Caesar is a prime number):

Assumption: Caesar is a prime number.  
P1: If something is a prime number, then it is abstract.  
P2: Abstract things are not efficacious.  
P3: Prime numbers are not efficacious (by virtue of being abstract)  
-> Caesar is not efficacious (because of his abstractness in virtue of being a prime number).

We have a meaningful conclusion for the wrong reason. We have been able to conclude something that obviously makes sense because non-efficaciousness can be meaningfully predicated on both persons and abstract things. The simple recipe for further examples is as follows: assume that X is a Y (where “X is a Y” expresses a category mistake). Take some predicate P that can be meaningfully predicated on Xs and Ys. Then the inference runs:

Assume X is Y.  
Y is P.
Hence X is P.

As this scheme indicates, we do not need the meaning of X, Y, and P, merely the condition that P can be predicated on both X and Y. So, the inference is rather non-material and formal.

8.5.2. Argument from failure to formalise categorical restrictions

Magidor (2013, p.48) argues that the best attempt at formalising categorical restrictions in natural language (namely, via Montague grammar) has failed. No other formalisation is forthcoming; hence there are no good reasons to think that categorical constraints can be formalised.

However, Magidor has not shown that categorical restrictions cannot be formalised, just that they cannot be formalised in some simple and elegant way as type theory aims to do (type theory builds on a few, two or so, “types”—like “individuals” and “truth values” from which further types are built by combinations) (Magidor 2013, pp.48–56). Magidor still leaves room for very gerrymandered restrictions or having one specific type for almost each concept (2013, pp.5–52). My response is that I would expect from the human conceptual apparatus that its categorical restrictions would turn out to be quite messy. An elegantly formal model, inspired by static axiomatic systems or traditional semantic toy models, also seems inadequate. The human conceptual apparatus is highly flexible and dynamic and conceptual content associated by tokening a concept seems highly context-dependent (e.g., Barsalou, 2009, 2011; Casasanto & Lupyan, 2015; Ludlow, 2014; Rice, 2016). Furthermore, as I have pointed out in the introduction, there is strong evidence for concepts involving modality-specific representations, i.e., non-language-like representations. So, it is not clear how such a neat formalisation could be carried out in a language-like format at all. Probably the representational system that can be formalised is some approximation, but the brain certainly does not implement a formal calculus exactly.

While I do not think that the actual conceptual apparatus can be neatly formalised for descriptive purposes, it is possible that our generative model in the brain contains some very high-level ‘hyper-prior’, i.e., some sub-personal, internal ‘norm’, that makes us unconsciously ‘desire’ (especially in scientific contexts) to have our mental representations and concepts formalised. The fact that we so persistently undertake-
-in scientific contexts--axiomatisations and formalisation might be a symptom of such an unconscious bias.

8.5.3. Argument from scientific progress

Camp (2004, p.230) further argues that scientific progress sometimes consists in formulating hypotheses that look senseless and are only minimally understood, for example, that light is a wave and a particle at the same time.

However, I do not think that this is an argument for unconstrained compositionality. There are at least two responses that allow for avoiding the acceptance of the PUC. Firstly, what Camp’s example shows is that there can be concept change and that accommodation mechanisms are at work (like in the case of metaphors). The pre-theoretic concept of light is not dualistic, while the revised, scientific concept is. Of course, the first time one spells out the idea that light is both a wave and a particle it sounds like nonsense (which it is—according to the old concept of light), and unless the idea is spelled out more specifically, leading to concept change, it remains nonsense. The fact that a meaningless sentence can trigger a concept revision does not imply that it was meaningful before the conceptual change happened. Concept change might be considered what I call an accommodation process over some larger time scale and with more permanent effects on the functional webs by which concepts are implemented. The connections of the (pre-change) concept of light to other concepts might be adjusted in such a way that taking light to be both a particle and wave ‘makes sense’ again. A second response (thanks to Andy Clark, in a personal conversation) is to insist that the dual concept of light is really a sort of disjunctive concept, where we have two partial models, one with light as a particle and the other with light as a wave. Science has not really integrated those partial models into one unified concept.

A last comment about the PUC as a mark of the conceptual competence of speakers is in order. It seems to me that avoiding meaningless concept combinations (and finding them ‘meaningless’) is a sign of conceptual competence, not the incompetence of the speaker. The skill of correctly combining concepts demonstrates that she knows the categorical restrictions. The meaning of concepts is implicitly encoded in the categorical restrictions. Someone systematically combining words in nonsensical cross-categorical ways would probably not be
judged competent in the language. Also, the PUC, rather than a requirement for conceptual competence, is a requirement of grammatical competence, because it is knowledge of the grammatical structure that underlies the possibility of unconstrained combinability. Semantic knowledge consists in the skill of being able to select those combinations that make sense.

I conclude that Camp’s and Magidor’s arguments for the PUC, and hence against the meaningless view of category mistakes, are not decisive. Therefore, the meaningfulness solution to the Liar Paradox, for which I have tried to provide a supporting cognitive account, is not undermined.

8.6. Conclusion

Firstly, I motivated a cognitive approach to understanding and resolving the Liar Paradox in opposition to the popular formalistic approaches. Building on a specific cognitive computational architecture for the mind, based on the predictive processing framework, I argued that the liar sentence is meaningless. It is meaningless because the situation that the sentence evokes cannot be conceived of in the ‘world model’ of the brain (in a way that is necessary for a paradox). Still, the liar looks meaningful, and it is difficult to escape its grip. This pull I explained via an accommodation process, fleshed out with a PP-specific mechanism of precision weighting of prediction errors. If the liar sentence is indeed meaningless (at the highest level of integration), the Principle of Unrestricted Compositionality must be wrong. I tried to show that some recent arguments for that principle are not decisive and do not threaten the meaninglessness solution.

Acknowledgements

I would like to thank Andy Clark, Mark Sprevak, Brian Rabern and two anonymous reviewers for their very helpful comments and suggestions.
Chapter 9. The Category Inclusion View of metaphor within the predictive processing framework

Abstract

In metaphor research two competing approaches have been dominating in the last decades. On the one hand, Category Inclusion Views hold that understanding a metaphor involves the modulation of a source concept into a more general category that includes the initial source and target concepts as instances. On the other hand, Implicit Comparison Views are based on structural analogies or mappings between source and target concepts. Empirical evidence seems to favour the Category Inclusion View only slightly, and it is still unclear which approach is better supported, or whether a hybrid approach is needed. I suggest that it might be useful in the search for more decisive empirical support to embed a theory of metaphor processing into some promising mechanistic neurocognitive framework. I attempt to take a first step in this programme by arguing that the predictive processing framework can underpin some version of the Category Inclusion View better than the Implicit Comparison View. This leads to the conditional claim that if predictive processing, which is increasingly influential in cognitive science, turns out to be an appropriate approach to understanding the mind, then this would support the Category Inclusion View and vice versa.

Keywords: metaphor; predictive processing; category inclusion view of metaphor; implicit comparison view of metaphor; relevance theory; conceptual metaphor theory

9.1. Introduction

Metaphors are linguistic constructions in which an expression (denoting the source concept) requires an interpretation that seems semantically distant from its literal, conventional meaning. In "Juliet is the sun", "sun" denotes the source concept that needs reinterpretation given that Juliet is not (literally) a heavenly body. In "We are at the crossroads of our lives", "crossroads" needs reinterpretation given that lives do

---

93 There are also non-linguistic metaphors. However, this article will focus on linguistic ones. One advantage of the PP based account that I will sketch here is that its extension to non-linguistic metaphors (e.g., visual metaphors) seems feasible, not least because PP builds on modal, i.e., sensorimotor representations.
not (literally) have physical roads that cross. The expression that needs a reinterpretation can be different parts of the sentence. The metaphors that traditionally received most attention are nominal metaphors like "Juliet is the sun" and "My lawyer is a shark". However, there are also verbal metaphors like "This deal stinks", where the verb "stinks" needs reinterpretation, or metaphors where the object needs reinterpretation like "She is feeding a princess" (where "princess" is meant to refer to her cat) (see Rodríguez Ronderos, 2021).

One central issue in metaphor research that is not yet settled is how they are cognitively represented and processed. How does this remarkable reinterpretation process work? Within the psycholinguistic literature two theories are the main contenders (e.g., see overviews in Holyoak & Stamencović, 2018; Rodríguez Ronderos, 2021). On the one hand, the family of Category Inclusion Views (CIV) holds that understanding a metaphor involves modulation of the source concept into a more inclusive category that can include the initial source and target concepts as instances. For instance, when interpreting "Juliet is the sun", "sun" is being modulated into an ad-hoc concept SUN* such that "Juliet is the sun*" can be understood as literally true under that interpretation of "sun". Similarly for a verbal metaphor like "This deal stinks", STINKS is being modulated into STINKS* such that a deal can "literally" stink*. Such an approach has been proposed in different guises by what I call Category Theory (CT) (e.g., Glucksberg & Keysar, 1990) and the Relevance Theory of Metaphor (RTM) (e.g., Carston 2002, 2010a, 2010b; Carston & Powell, 2009; Sperber & Wilson, 2008).

On the other hand, accounts following the Implicit Comparison View (ICV) suggest that metaphor involves the processing of a structural analogy between the source (e.g., sun) and target (e.g., Juliet) concepts. Comprehending a metaphor consists in grasping a set of salient similarities or an analogy, very much like when one says, "Juliet is like the sun". Therefore, metaphor processing consists in representing an analogical mapping between properties of the source and the target domain. Key proposals within this camp are Structure Mapping Theory (e.g., Gentner & Markman, 1997; Gentner & Bowdle, 2008) and Conceptual Metaphor Theory (CMT) (Lakoff, 2008; Lakoff & Johnson, 1980a,b). According to those approaches, metaphors are stable conceptual mappings that pervasively structure the way we think. Blending theory (Fauconnier & Turner, 2003) is considered to be a generalisation of CMT,
because it does not rely so much on fixed mappings but allows the flexible creation on the fly of novel and more complex mappings. Notice that Structural Mapping Theory and CMT/Blending Theory are considered to be compatible (Gentner et al., 2001; Holyoak & Stamenković, 2018; Murphy, 1988; Coulson and Oakley, 2005) so I will focus here on CTM as the most representative theory of the Implicit Comparison View (CIV) camp.

A recent meta-study (Holyoak & Stamenković, 2018) and empirical research (Rodriquez Ronderos, 2021) support the hypothesis that the CIV is empirically slightly more supported. However, this conclusion is not yet decisive, and it is still a live option that both approaches need to be combined into a hybrid to be able to accommodate the empirical data. For instance, Gibbs & Tendahl (2006), Tendahl & Gibbs (2008), as well as Wilson (2011) have tentatively explored whether RT and CMT can be fruitfully combined. Others (e.g., Stöckl, 2010) have suggested pluralistic accounts into which CIV and ICV could be synthesised. However, no full hybrid theory fleshed out on the cognitive-computational or neuro-mechanistic level has been put forward so far.

This chapter attempts to contribute to the debate regarding cognitive processing of metaphor. In the face of the current stalemate, the methodological idea is the following. Take some plausible and promising general neurocognitive computational and mechanistic paradigm that is motivated and developed independently from metaphor research. I will take predictive processing (PP) as an emerging and promising framework for that purpose. Ask whether CIV and ICV can be naturally derived within this framework, i.e., which metaphor theory can be cognitive-computationally underpinned by it. If CIV but not ICV sits better with that cognitive paradigm, then this would provide some support to CIV. This presupposes, of course, that the cognitive framework turns out to be adequate. But merely by embedding a metaphor theory into a general brain theory, new strategies and options for empirical verification should arise, given that the brain theory provides concepts, resources and constraints that allow for formulating testable hypotheses.

The phenomenon of metaphor has not yet been addressed by the PP literature. Therefore, there is also a motivation for the PP camp to deal with metaphor if it is to live up to its ambition to be a framework that also covers higher cognition. In this case, we can also see how PP (or at least an account of concepts and metaphor
within PP) could be empirically confirmed or disconfirmed. This is significant, because PP needs to demonstrate that it can generate hypotheses that are compatible with evidence but also that those hypotheses could turn out to be wrong empirically (see Litwin & Miłkowski, 2020). If we have two competing metaphor theories, and the PP account proposed here speaks for one and not so much for the other, then we have a case where discriminatory confirmation is possible. It is a case where divergent predictions can be made.

In the scope of this chapter, I can carry out only a first small, but necessary part of this programme. I will focus on arguing on theoretical grounds that an increasingly popular and influential cognitive paradigm, predictive processing (PP), can underpin CIV better than ICV. If PP turns out to be a correct framework for modelling the mind, including higher cognition, then this would support CIV as opposed to ICV. PP with its neuro-mechanistic commitments can also be used by the CIV camp to derive empirically testable hypotheses.

To achieve this objective, the rest of the essay is structured as follows. In Section 9.2, I introduce the neurocognitive framework, predictive processing (PP) and provide a view on how concepts are represented. In Section 9.3, I provide a PP account for different types of metaphors. In Section 9.4, I show how PP could underpin a version of the CIV. In Section 9.5, I argue that PP does not support the ICV well. In Section 9.6, I argue that while PP and the specific account of conceptual representation proposed does not support ICV well, it can accommodate those elements of ICV that typically feature in hybrid theories.

9.2. The predictive processing paradigm and conceptual representations

9.2.1. The PP paradigm

PP is an increasingly influential framework in Cognitive Science that pictures the brain as an embodied prediction organ that constantly tries to anticipate its own sensory inputs (Clark, 2013, 2016; Hohwy, 2013, 2020; Friston, 2010). For the predictions, the brain uses and constantly improves a probabilistic, hierarchical,
generative model. The model mirrors relevant features of the causal structure of the world that generates the sensory signals. This allows it to simulate and anticipate sensory inflow (hence "generative"). Representations in the model are probabilistic, i.e., they are in the form of probability distributions. Inference in the model is approximate Bayesian inference, implemented by a prediction error minimisation mechanism. Prediction errors in lower levels are signalled to higher levels, to adjust the model and allow it to make better predictions next time. Predictions flow from higher to lower levels in the hierarchy. The aim of the model adjustments is to achieve an overall model that minimises the prediction errors on average in the long run (e.g., Clark, 2013, pp.188–189; Hohwy, 2013, pp.15–39, 2020). There are two fundamental ways to minimise prediction errors: updating the model to better fit the incoming evidence and acting on the environment to fit the prediction. The first—which will be the relevant one for the current purposes—can be achieved in different ways: by changing prior probabilities over hypotheses, by adding new variables and links, or by changing the level of detail of the prediction (see, e.g., Kwisthout et al., 2017, for more detail).

9.2.2. The PP prediction model

We can visualise the mental PP model as a huge hierarchical network consisting of nodes in the form of "prediction units" (PUs), which represent expectations on different levels of abstraction or a spatiotemporal scale. The PUs at each level are connected to the PUs on the next lower levels and constrain their activation. But each PU is itself constrained by higher-level PUs. PUs encode prior probability distributions over the PUs on the next lower level. Those prior probabilities get constantly updated to posterior probabilities (which then become the new prior probabilities), depending on the PU's "context". A PU's context is those PUs that are laterally (similar level) connected and connected from higher levels in the hierarchy. We can imagine PUs as "expectations" (sub-conscious, but also conscious ones, which we notice are violated when we have, e.g., a feeling of oddness/infelicity in sentence comprehension).

Perception, cognition, and action are then explicated as adjusting the expectation network to minimise the violation of all of the active PUs. The model needs to balance in each moment all sorts of prior beliefs and top-down expectations and
bottom-up evidence to achieve an error minimising (i.e., as coherent as possible) whole. A specific mechanism, the "prediction error weighting apparatus" (e.g., Clark, 2016, pp.53–83), suppresses unreliable upward flowing error signals (e.g., noisy visual input). If you are in a fog, your visual input should be regarded as unreliable, and you should rely more on prior knowledge. Any violation of expectations that is not being suppressed corresponds to a prediction error that needs to be "explained away" by changing higher-level expectations (the higher-level PUs (new) prior probability distribution over its lower-level nodes).

To illustrate the model, and especially how the different layers of abstraction work together, consider the following example. Imagine you are at the Opera. The circumstance that you are in the Opera is represented by some PU, which serves as a very high level (situational) prior and determines a range of expectations at lower representational levels. This PU that represents the "tacit belief" that you are at the Opera itself is the result of perceptual input processed upward through the visual pathway in the brain, which compresses pixels to edge forms, to more complex shapes and even more complex representations of environments and things, namely of, e.g., the foyer of the opera building where you currently are, with chairs, stairways, spectators, etc. The representations on each level serve as predictions/expectations (priors) for the lower levels. So, the edge forms are predictions for pixel patterns, the complex shapes for edge patterns, the objects and foyer interior for the complex shapes, the conceptualisation of a foyer scene for objects and the foyer's interior, and finally the belief that you are at the Opera for the foyer scene. If you have the belief that you are at the Opera, the representational hierarchy of your brain model has found an error minimising equilibrium of expectations on all levels of the hierarchy. There might be other priors at an even higher level, like the belief that you exist, that you are awake, that something exists at all, and so on. Details do not matter here; the key point of the example is that the model has a hierarchical structure all the way from small scale perceptual representation (like retinal "pixels"), up to everyday beliefs and very general (tacit) "hinge beliefs" like that I exist. Now as you move into the auditorium, you would be surprised to see an elephant on the stage. If that were the case, your brain would hum. A huge prediction error needs to be supressed. Activity in the expectation network sets in with the aim of explaining away the unexpected. Normally you would
expect opera singers and orchestra musicians wearing formal suits and you would be surprised to see a nude musician or one in a pink suit.

In sum, the brain's predictive model is a gigantic hierarchical "expectation network" of interconnected PUs. As we go up the hierarchy what PUs represent gets more and more compressed and covers larger and larger receptive fields, i.e., larger spatiotemporal scales. Cognition is driven by the search for a brain state that minimises the violation of all expectations corresponding to PUs on all levels of the hierarchy.

9.2.3. Concepts as expectation sub-networks

In this section I sketch a proposal about how concepts are structured and processed within the PP framework. According to this account, words are associated with subnetworks of the overall PP model. They are subnetworks of prediction units (see Michel, 2020a,b). The root-node of such a subnetwork is what stably corresponds to a concept. However, the cognitively relevant content of a concept can vary flexibly and dynamically. The tokening of a concept is the selective and context-sensitive activation of lower-level prediction units from the concept's subnetwork. Child nodes of the root-node can be seen as "features" of the concept.

Michel (2020a) has proposed a mechanism regarding how context-sensitive modulation of concept features could work. In the extreme case only the root node is activated. This means that the concept is represented in a very compressed "gist"-like manner (see also Eliasmith, 2013; Thagard, 2019). At the other extreme, the activation of the concept could be deep, such that, additionally, nodes all the way down from the root node to the level of perceptual representations are activated. This means the concept is activated more concretely, i.e., with more concrete sensorimotor information.

The picture proposed here is inspired by Langacker's (e.g., 1987, 2008) view on language and concepts ("Cognitive Grammar"), as well as neo-empiricist accounts of concepts (e.g., Barsalou, 1999, 2009, 2011; Prinz, 2002). Therefore, it does not seem promising to try to formalise and make explicit the concept's expectation
network because here we are not dealing with a LOT-like amodal symbol system. The reason is that not all nodes (probably only very few) are "interpretable" representations (concepts traditionally understood, which are lexicalised). The intermediate complex patterns in the visual path for instance (edge shapes of all sorts and orientations) do not correspond to "concepts" in the conventional sense, they are "sub-symbolic". The hypothesis is that only some PUs on some levels of the hierarchy correspond to consciously accessible, lexicalizable concepts. Others operate sub-personally and manifest themselves in different weaker forms, often as intuitions in the case of higher-level prediction units, or they do not manifest themselves at all, especially the lower, "sub-symbolic" levels.

According to this model of concepts within the PP framework, PUs play the role of representational devices in terms of which predictions are made. To entertain a thought is just to activate a web of prediction units such that the prediction errors are minimised on all of the levels. To believe that the door is open is to activate a host of appropriate prediction units in the hierarchy brought into an overall error minimising balance. The belief that the door is open, in the view here, might be accompanied by "imagery" (e.g., that a specific door is open) if many PUs on lower levels, close to the sensorimotor periphery, are activated.

Exactly which other lower-level PUs are selected when a concept is instantiated depends on the context. For instance, when a spectator at a concert is using the word "piano", it might activate the concept unit PIANO together with features in the information package that represent nice sounds and an instrument played artfully with the hands. Alternatively, in the mind of a furniture remover during a relocation job, the same word might activate PIANO and the information that it is a heavy and bulky object.

For efficiency reasons, the brain needs to make predictions at an adequate level of detail. Imagine you want to cross a busy street. It would be a waste of resources to predict the environment at the level of detail of the exact retinal pixel pattern's temporal evolution. Predictions should be made with a coarser grain in terms of

---

95 While a strict full formalisation does not seem promising, it still might be extremely useful to model and describe certain general patterns in a formal manner. The point is that the PP picture claims that the mind does not work as a fully formalizable amodal symbol system. But I do not claim that it might be fruitful to describe some or many aspects of cognition in formal terms.
whole objects (e.g., cars and one's own body) and other more abstract notions (e.g., speed and direction). The representational detail of concepts can be modulated in two different ways. Firstly, a concept unit can be activated with a more or less extensive sub-network. At the other extreme, in a "shallow" mode, the sub-network is not extended too far away from the root node. For instance, a car can be represented with a concrete colour and shape ("deep mode") or merely as a schematic vehicle ("shallow mode"). The second way to modulate detail is to choose a concept unit at an adequate level of the hierarchy. For instance, the root node of "car" might be the very schematic conceptual representation CAR. The concept units SPORTS-CAR, DELIVERY-VAN, etc., might be more concrete instances of CAR represented at the next lower level of the hierarchy. When someone warns you "Be careful, a car!" she might just be thinking of a generic CAR. When a visitor at a sports car exhibition says, "What a nice car!" she might have in mind the more concrete SPORTS-CAR.

Meta-linguistic information, e.g., that a certain sentence is interpretable, works as priors as well. When I hold the fixed belief (a strong prior) that a sentence is interpretable, priors on the lower, namely word level are driven to adjustments until the most probable interpretation is obtained. In this adjustment (inference) process, many other priors are involved, representing relevant prior knowledge that needs to be brought to bear during the interpretation process. Such priors could include contextual information and certain "maxims" like Gricean conversational maxims, which are part of the tacit or not so tacit background beliefs of the agent that influence expectations. If I am at a Dadaist exhibition and read an absurd sentence, a contextual prior might be that the sentence is non-interpretable, and hence I expect it to be non-interpretable and do not even try to make sense of it.

9.3. A PP approach to metaphor processing

Based on this model of concepts or word meanings as sub-networks or hierarchies of priors, and the idea of continual modulation for error minimisation, I will now propose a mechanism for metaphor processing. I will first summarise the main idea and then immediately work through some examples. I will then show (in Section 9.4) that the proposed PP account of metaphor cognitive-computationally underpins a version of the Category Inclusion View (CIV).
The core idea is that metaphor comprehension results from the modulation of the expectation hierarchy associated with the word (or expression) denoting the source concept. "Modulation" means that certain (higher-level) nodes of the hierarchy are context-dependently selected, and others (lower-level ones) are suppressed. The target concept serves as a contextual prior for the modulation of the source concept.

To avoid a large prediction error in the processing of the metaphor, the source concept cannot be used with its conventional meaning, which produces a category mistake in combination with the target concept.96 ("Conventional" meaning is the configuration of PUs from the concept's network that has the highest prior probability given no special context.) Rather, the mind needs to infer a configuration of priors that minimises prediction errors, i.e., one where all of the priors are consistent (i.e., they form a consistent network of expectations). This is then perceived as a felicitous sentence, one that makes sense.

The best possible configuration is one in which those features of the source concept are suppressed that are not expected by the target concept in the context (= have a low conditional probability given the target concept's and context priors' activation). In other words, as features are suppressed, we get a more abstract meaning; the source concept is being modulated into a more abstract meaning.

Another way to minimise prediction error would be to have a higher-level prior that represents the prior belief that the sentence is meaningless. Such a hypothesis is plausible under certain circumstances, for instance, when the sentence has been pronounced by a madman or a parrot. Then the mind would not try to make sense of the sentence by adapting feature nodes of the source concept.

However, in most cases of metaphor processing, the attitude is charitable, that is, there is a higher-level prior that predicts that the sentence is meaningful, and this belief is clamped or fixed (assigned a high reliability by the error weighting mechanism). The sentence must hence be made meaningful by adequately modulating the source concept features, specifically lower-level ones.

96 Notice that I am not claiming that we first try the conventional meaning, notice a huge prediction error, and then search for a better non-conventional interpretation. The PP prediction error minimisation process does not work as a sequence of steps at all but is a holistic self-organised constraint minimisation process.
In the rest of this section, I will spell out the proposed PP metaphor mechanism in more detail, using specific examples of two different types of metaphors, namely: resemblance (I) and indirect (II) metaphors.

9.3.1. Simple nominal metaphors (resemblance metaphors)

What is going on when a reader comprehends "Juliet is the sun," according to the account proposed here? The reading takes place in a broader context. Aspects of this context are also represented in the mind and serve as priors. One such contextual element that is represented as a high-level prior is that the sentence is recognised as grammatical and presupposed to be meaningful. Here, "meaningful" shall refer to a psychological notion that the sentence is perceived as understandable, that it is cognitively acceptable or "felicitous". It is - let us say - a credible writer to whom we ascribe an intention to express a statement that can be accepted as felicitous by the hearer. This contextual element is represented in the mind as some strong high-level prior (hyper-prior) with the content "sentence is meaningful". This hyperprior is "clamped", so any outcome in which the sentence is infelicitous produces a high error signal, which drives further modulation of the prior network at lower levels to find a better "interpretation" that better minimises prediction errors.

Let us say that some configuration of PUs yields an infelicitous interpretation of the sentence, i.e., there are unexpected elements in the sentence for which no modulation has been able to suppress errors. What is necessary now is to adjust the priors at subordinate levels, e.g., those representing word meanings, to minimise the system's prediction error. Words have a probability distribution for their possible senses (i.e., for different ways to select features from their sub-network). Without a special context the conventional sense is the most probable (has the highest prior probability) and taking the conventional sense of the words "Juliet" and "sun" would lead to a large prediction error under the assumption (the hyperprior) that the sentence is felicitous because Juliet is not a heavenly body. When processing the sentence, the word "sun" is unexpected as a continuation of "Juliet is the..."

How can the prediction error be minimised? The brain can minimise the prediction error by trying to adjust the meaning of the word "sun" by suppressing certain features of the subnetwork of sun. In this way, we get the more schematic (or
abstract) sense $\text{SUN}^*$, which is a slimmed down (abstracted) version of the full network of $\text{SUN}$. This more schematic sense might be something like "something with a central role" or "something that produces a positive feeling", depending on the writer's intention and on other factors like the preceding text, the cultural context, etc. The sentence now appears to be felicitous because after hearing/reading "Juliet is ..." we can expect "something with a central role" or "something that produces a positive feeling".

Note, that metaphorical meanings are notoriously difficult to paraphrase. So, the example here is just illustrative. Also, metaphors are often considered to be "open textured", (i.e., they are suggestive of many possible interpretations). In the PP account this open texture can be accommodated quite naturally because the network needs to suppress only features that create a violation of expectations (e.g., "is a heavenly body", "carries out nuclear fusion", etc.). All of the remaining features are still available for further modulation and use, leading to the impression of open texture. Open texture reflects the remaining consistent possibilities to select subsets of nodes that have not been suppressed.

This mechanism implies that the target concept $\text{JULIET}$ can "literally" be categorised as an instance of $\text{SUN}^*$. Now with the priors on all levels (contextual representations, sentence as a whole, words, etc.) "fitting together" the prediction error is minimised. The ordinary $\text{SUN}$ is, like $\text{JULIET}$, an instance of $\text{SUN}^*$ because $\text{SUN}^*$ is sufficiently schematic to include both $\text{JULIET}$ and $\text{SUN}$. Note that it is not the case that there is one specific and concrete sense that has been settled on in the prediction error minimisation state; that would also speak against open texture. Rather, we could now interpret $\text{SUN}^*$ in different, more specific ways.

Conventionalised and novel metaphors are processed in essentially the same fashion. The difference is one of a degree in the modulation effort. In the case of conventionalised metaphors, the source concept is more easily modulated into the meaning that fits with the target concept. It is easier because the probability of the hypothesis corresponding to the conventional modulation is larger (precisely because it has been conventionalised, its prior probability is higher). Novel metaphors might require more cognitive effort; they are more surprising. The more often they are used in a certain context, the higher the prior probabilities of the modulated senses get, conditioned on that context.
One might ask how does the system "know" which features to supress and which not to? The short (and maybe unsatisfying) answer is that the suppressions work via a holistic error minimising process. It is unlikely that this can be formalised by any rule-based algorithm that makes the selection perspicuous. PP is not a classical computational paradigm with amodal symbols and rules. Rather, in the brain, a self-organising holistic constraint-minimisation process is going on continuously, namely in the form of a prediction error minimisation process that happens on all levels in the hierarchy simultaneously.

9.3.2. Indirect metaphors

Let us run through a different type of metaphor to illustrate how the PP mechanism of processing works:

"We are at the crossroads of our lives."

The metaphor is indirect or implicit because the target concept is not mentioned; rather is has been substituted by the source concept. In this example, "crossroads" stands for something like "a situation in which we need to take a crucial decision about various options".

Again, assume that in the given context, the listener/reader strongly assumes that the sentence is meaningful, i.e., it can be felicitously interpreted. When reading "crossroads of our lives", a modulation mechanism kicks in to adjust the feature hierarchy associated with CROSSROADS. Features of CROSSROADS are suppressed to yield a more schematic notion CROSSROADS* with the content, say, "situation in which we need to take a crucial decision about various options". In this more abstract representation, more concrete physical path features are suppressed, and it remains unspecified that the decision is related to one among various paths at a physical crossroads. The conceptual representation of the life situation and crossroads can now both be considered as instances of CROSSROADS*.
9.4. PP as a cognitive computational underpinning for the Category Inclusion View

The PP account of metaphor just described closely follows the general tenet of CIV, which is based on the idea of ad hoc category creation or concept modulation. But note that CIV is not a single specific theory but a family of such theories. There are two main incumbents within the CIV camp: the Relevance Theory of Metaphor (RTM) (e.g., Carston, 2002, 2010a, 2010b; Carston & Powell, 2009; Sperber & Wilson, 2008), and the Interactive Property Attribution Model (e.g., Estes & Glucksberg, 2000; Glucksberg & McGlone, 2001; Glucksberg, 2011; Glucksberg, McGlone, & Manfredi, 1997; McGlone, 1996; McGlone & Manfredi, 2001).

Beyond the common high-level feature that metaphor is based on concept modulation or ad-hoc concept creation, the RTM and CT accounts are quite different in their detail. Therefore, it is necessary to further consider how each account might be related to the PP model proposed here.

9.4.1. Glucksberg et al.’s Interactive Property Attribution Model

According to the Interactive Property Attribution Model, the source concept is modulated into an ad hoc category with a selection of salient and relevant features that are attributed to the target concept. The source and target concepts have asymmetric roles in this attribution process. The target concept serves as a local context for the source concept (Glucksberg & McGlone, 2001, p.53). The source provides properties for attribution to the targets. The source has a double reference: it refers both to the literal sense and the higher-level ad hoc category that has the literal source and target concepts as instances. Note that only the relevant properties of the source are captured by the ad hoc category. The relevance is determined by the "dimensions for attribution". E.g., for "road", aspects such as shape, surface, safety, and speed are relevant, while other aspects are generally not salient, like the colour or cost. The ways in which roads can meaningfully differ establish the relevant dimensions of attribution. Different targets activate different source properties. For instance, in the "the road is a snake" the curvy form of snakes is relevant; in "my lawyer is a snake" the sly behaviour of snakes might be relevant. This requires two sorts of world knowledge to be represented: in regard to the target, what dimensions
are relevant; and in regard to the source, "what kind of things it can epitomize" (Glucksberg & McGlone, 2001, p.55).

Gluckberg's account and the PP account of metaphor described above appear to be a good fit. PP can serve straightforwardly as an implementational cognitive-computational framework for Glucksberg's account. In PP, concepts are represented as expectation hierarchies implemented by a host of prediction units/priors. The target concept features can be seen as priors that influence the processing of the source features (i.e., serve as a context, which is just to say that the source priors are under the influence of the target priors).

Take the example "roads are snakes". ROAD can be seen as the parent node for the next lower-level features like "Road-SHAPE", "Road-SURFACE", etc. Given the general context, the most salient features are the ones that are more strongly expected (i.e., have the highest prior probability). Those then serve as context priors for dynamically selected features of the source concept SNAKE. The relevant "dimension priors" provided by ROAD set a context (serve as priors or expectations) for the selection of features of SNAKE. There is a set of properties of SNAKE that are strongly expected given ROAD (i.e., have a high conditional probability). For instance, WINDING is made salient (i.e., expected) by ROAD and serves in turn as a prior for processing SNAKE. In SNAKE then some properties are suppressed (those not expected by WINDING) and others are activated (those expected by WINDING). This leads to a modulation of SNAKE to SNAKE *. In "The road is a snake", 'road' allows us to expect "snake" when its sense SNAKE has been modulated into SNAKE *. In this configuration then, the prediction errors are minimised, and the resulting representation constitutes the comprehension.

At this stage one might wonder what PP actually adds to this picture of metaphor; is the PP account not merely a redescription of Glucksberg's view with PP vocabulary? My answer is that firstly, PP is more general, as it covers not only direct metaphors, which are the scope of the Interactive Property Attribution Model, but also indirect ones (see 9.3.2). Secondly, PP is an implementational-level account with substantially more commitments on the computational, algorithmic and

97 Stöckl (2010) characterises Glucksberg's account as a linguistic account (as opposed to CTM, which he characterises as a cognitive account. However, I take it to be a cognitive account as well aiming at explaining the flexible semantic processing of word meanings. See also Section 9.6.
implementational level than the Interactive Property Attribution Model,\textsuperscript{98} which is a specific metaphor theory, not a general cognitive theory. Thirdly, the PP metaphor account arises very naturally from PP as a general cognitive framework. It is not an ad-hoc account only for (nominal) metaphors. So, even if the PP account of metaphor is not a new story (at least for nominal metaphors and on the level of description targeted by Glucksberg, i.e., without neuromechanical commitments), its embedding into the PP framework has significant additional benefits; it extends its scope and adds lower levels of description (neuro-mechanistic ones) and motivates it within a broader picture of cognition.

As an example of how the PP account could be used to provide a neat mechanistic account for existing empirical findings, consider McGlone & Manfredi (2001), who provide discriminative support for CIV and against some versions of ICV.\textsuperscript{99} CIV predicts that priming with ground-irrelevant properties of the vehicle concept slows down metaphor comprehension. For instance, if a comprehension task for "The lawyer is a shark" is primed by the mentioning of "The shark is blue", then comprehension should be slower. This is because this sentence primes a literal reading of "shark" in the metaphorical sentence. To move to the second, more abstract, referent of "shark" implies a switching cost. According to some versions of ICV no such priming effect is predicted because the comparison process starts from the literal senses of the source and target concepts, hence a ground-irrelevant prime should not slow down the process. The PP explanation is that the activation of the source is due to the prime: the shark node and some sub-nodes representing a specific colour property are activated. The prediction error generated by the metaphorical sentence is large and the blue feature needs to be suppressed (and some other, relevant feature activated). This adjustment of the priors in the model implies a larger cognitive effort. In the unprimed situation, comprehension can start with an activation pattern of the concept root-node and sub-nodes in which less sub-nodes (or even only the root node) are activated, and hence the effort of subsequent suppression is avoided.

\textsuperscript{98} This is a comparative claim. It does not imply that PP is fully specified on all levels; rather PP is still heavily under construction (see, e.g., Sprevak, 2021a-d).

\textsuperscript{99} The results are compatible with the structure mapping approach, but not with "pure" property comparison models like Ortony's salience imbalance model (see McGlone & Manfredi, 2001, p.1215).
9.4.2. **Relevance Theory of Metaphor**

A second theory of metaphor within the Category Inclusion View is the Relevance Theory of Metaphor (RTM). I argue that while there is a lot of common ground, PP and RTM are in tension. The reason is RTM's commitment to the LOTH paradigm and a truth conditional orientation towards meaning, as opposed to PP's commitment to modality-specific representations and a notion of meaning that is oriented towards the cognitive-psychological significance.

Relevance theory is a theory of communication that emphasizes the importance of inferential processes in language processing. Communication consists in "providing evidence for an intended inference about the communicator's informative intentions." The mind constantly tries to infer the speaker's intentions. Interpretation happens based on decoded linguistic meaning that is fed into a process to recognize the speaker's intentions under the "principle of relevance", i.e., the presumption that the information communicated is relevant. The relevance principle has its roots in the Gricean principles of communication (e.g., Carston 2010a, p.162), According to RT, all lexical meanings involve pragmatic modulation of the literally encoded meaning driven by this principle. Interpretations of metaphors also follow this general principle of expectation of optimal relevance. Metaphor implies that some features of the source concept are broadened (i.e., abstracted) and others are narrowed (i.e., more specific features added) until some balance is achieved in terms of a good enough trade-off between processing effort and cognitive effect leading to the formation of an "ad-hoc concept":

> [... modulation or adjustment of the meaning encoded by a linguistic constituent involves an interaction among the lexically encoded concept, the other concepts encoded by the utterance and contextual information, constrained by the hearer's expectation of relevance [...]. The outcome of this process is what is known as an ad hoc concept ('ad hoc' in that it has to be inferentially derived on, and for, the particular occasion of use). (Carston, 2010a, p.158)

"Cognitive effect" is understood here as, roughly, appropriate inferences. While a word has only one linguistically encoded meaning, it is polysemous in terms of what it contributes to the truth conditions (its "explicature") in a specific context of application. In metaphor processing, the pragmatically derived meaning corresponds to an abstracted source category and the linguistically encoded meaning to the initial
source concept. RT emphasises a deflationist approach to metaphor. No special mechanism is needed for metaphor, as the inferential process is general, so metaphor is just one manifestation on a continuum of inferential modulation effects.

Despite RT being a cognitive theory specifically of communication and pragmatics and PP a general theory of cognition, there is some significant common ground. Both emphasise the inferential character and the flexibility of the cognitive content of lexical items. However, there is a discrepancy in the fundamental cognitive-computational paradigm that might be an obstacle to a closer alliance between RT and PP. RT relies on the Fodorian LOTH paradigm (see, e.g., Carston 2018, p.210). Conceptual and linguistic representations have an amodal format and thoughts have a propositional form, based on components and their syntactic combination. Contexts are sets of propositional assumptions, and inferences look like logical deductions in a formal symbols system. PP, however, is a paradigm emphasizing the importance of modality specific representations.

Notice that the PP approach is, like the RT approach, deflationist. The difference is only that the inferences over amodal representations in RT are replaced in the PP framework by inferences (realised as error prediction minimisation) over modality-specific representations with varying degrees of abstraction.

Because of the commitment to LOTH, Carston (2018) rejects that imagery and affective states are part of the metaphorical meaning. This seems to be a significant point of conflict with the PP account. But it all depends on what one means by "metaphorical meaning". Carston is concerned with a notion of meaning that relates to truth-conditional semantics. However, PP is concerned with the "cognitive meaning", i.e., with the cognitive-psychological significance of conceptual representations. This distinction is important, given also that they are arguably orthogonal aspects (see Machery, 2009; Löhr, 2020).

The notion of concept relevant for PP corresponds to those entities that drive cognitive processing. As imagery does contribute, as e.g., Carston also admits, to cognitive processing (for the cognitive significance of imagery see, e.g., Ifantidou, 2021), it would seem unmotivated to keep it in a cognitively oriented theory as a merely ancillary phenomenon of metaphor. This does not mean that there might not
be another notion of "meaning", namely the truth-conditional one, for which those cognitive-psychological aspects are indeed irrelevant.

Let me draw attention to the fact that the exact cognitive and neurally plausible mechanisms underlying the trade-off between relevance and cognitive effort are left largely unspecified in RT. Also, the notion of "relevance" is highly contentious (see, e.g., Sperber & Wilson, 1996; Chiappe & Kukla, 1996). According to RT, the "relevance" of some representation depends on the implications that can be drawn from it. However, one can draw an infinite number of implications from any statement. What is needed, of course, are "relevant" implications. But this leads to a regress. Of course, relevance is a technical term and should not be understood in the common-sense way, but as some label for "cognitive effect". Also, RT is explicit that cognitive effort and effect are not being "calculated", but somehow arise from electro-physical processes (Sperber & Wilson, 1996). While this effort/effect trade-off seems intuitively highly plausible, its mechanistic implementation remains largely unclear.

PP has the advantage that it hypothesises just one simple computational principle, namely prediction error minimisation and it is perspicuous how it could be implemented in principle in a neurally plausible fashion. "Relevance" does play a crucial role in PP, though this notion should be understood in a much broader and general way than in RT. RT's understanding of relevance seems to correspond to a narrower, intellectualised notion where relevance is propositionally expressible. The inferences can then be rationalised and explained. "Relevant" in the very inclusive understanding within PP is everything that contributes to prediction error minimisation. It is the precision weighting mechanism that suppresses irrelevant (incl. unreliable) signals and in this way steers the inference process. The "knowledge" encoded in the precision weighting mechanism can itself be learned by the sort of hierarchical Bayesian model that PP posits. Knowing" what is relevant information is represented as "meta-knowledge" in higher levels ("hyper-priors") of that model.

In conclusion, RT and PP share the CIV spirit and could be allies, not least because they see cognition as massively inferential and concepts as highly dynamic and flexible. As PP has algorithmic and implementational level commitments, it could contribute to a joint project. However, the commitment to very different paradigms
that imply different representational formats (namely amodal versus modality-specific ones) might be an obstacle to such a combination.

9.5. PP and Conceptual Metaphor Theory

In this section I argue that certain aspects of the ICV do not sit well with some of the commitments of PP. Therefore, I conclude overall that PP supports the (and specifically the Property Attribution Model) better than the ICV.

I will use Lakoff's Conceptual Metaphor Theory (CMT) as the arguably most influential ICV account. According to CMT, abstract concepts are represented as structural mappings between a source and target concept. Abstract concepts like \textit{ARGUMENT} are represented and understood in virtue of an underlying fixed mapping between, e.g., \textit{ARGUMENT} and \textit{BUILDING} (the conceptual metaphor "\textit{ARGUMENTS ARE BUILDINGS}"). The central idea is that when we are thinking about arguments, we do this in virtue of exploiting a structural analogy between building and argument properties, like \textit{FOUNDATIONS = PREMISES}, \textit{SOLIDITY = VALIDITY}, etc. A \textit{BUILDING} is supposed to be more "concrete", i.e., it is represented in the form of the specific experiences based on physical interaction and the perception of buildings.

CMT emphasises thematic clusters in metaphorical language. For instance, the conceptual metaphor \textit{LIFE IS A JOURNEY} gives rise to many related journey-based analogies in the life domain: "We are at the crossroads of our lives", "This marriage is a roller-coaster", and so forth. A strength of CMT is therefore that it can account for this sort of systematicity and productivity in the form of inferences from an underlying mapping represented as a conceptual metaphor.

CMT has been criticised as being theoretically and empirically inadequate on various grounds (see, e.g., Glucksberg & McGlone, 2001, Chapter 6, for a critique of CMT, Gibbs, 2017, for a discussion and defence of CMT). I cannot possibly discuss and evaluate the debate here, but this is also not necessary for the current purposes. Rather I will focus on making the point that one crucial aspect of CMT seems to conflict with the PP based account of metaphor proposed here. This is enough to show that PP does not straightforwardly underpin CMT.

A very central point of CMT has to do with the nature of concrete versus abstract concepts. According to CMT, abstract concepts are crucially dependent on concrete
concepts. It is through concrete concepts and their mapping onto abstract concepts that we understand the abstract concepts. That is, it seems that we cannot think in terms of abstract concepts without the help of concrete concepts.

In contrast, in the PP picture of concepts, as I have framed it, abstract concepts (which are implemented as higher-level PUs) can be represented independently from lower-level sensorimotor PUs. More importantly, there is no dichotomy between concrete and abstract concepts in the first place (see Michel, 2020b). The representations close to the sensorimotor periphery are abstracted in multiple layers to more abstract representations. It might be that the creation of abstract concepts relies on the compression/abstraction and convolution of more concrete concept units (lower-level priors), but once a prior is created as a PU it can have its own life. Analogies represented as mappings in the CMT model can be fleshed out in the PP model as categorization relations, e.g., relations between nodes and sub-nodes. According to the PP proposal, the mind can represent, for instance, ARGUMENT without a representation or co-activation of BUILDING in the form of a mapping. ARGUMENT corresponds to a PU that has abstracted over all kinds of experiences in connection with arguments, including linguistic information and descriptions (which provide an indirect sensorimotor grounding). ARGUMENT is just a PU that allows for recognizing certain complex patterns of its input nodes (its "receptive field"). Those do not require the concrete concept BUILDING at all. However, it turns out that there are fruitful analogies between buildings and arguments. Those analogies (or pattern) are captured by higher-level PUs for which argument and building can be considered instances. We do not have verbal labels for them. There is a PU representing the category for which ARGUMENT and BUILDING are instances. As we have no verbal label for it, we conveniently use BUILDING as a label. Exemplars of buildings are represented closer to the sensorimotor periphery than exemplars of arguments.

Admittedly, it is plausible that the source concept (e.g., BUILDING) is in some form cognitively privileged over the target concept (e.g., ARGUMENT), but not because the target concept is dependent on the source concept in the strong sense postulated by CMT. The way the source concept is privileged over the target concept has to do with the fact that the source concept can be more easily shared or communicated. I can share a colour, by pointing to it, because it pertains to the domain of exteroception. Sharing/communicating the concept FREEDOM or ANGER is more
difficult (but it is not more difficult to represent and hence "understand" the concept), not least because abstract concepts plausibly involve affective states in the form of interoceptive representations. To share ideas, the speaker needs to externalise concepts, by translating them into sensorimotor cues, and the hearer then needs to translate those back into concepts that are in higher levels of his model. So, any concept that is already closer to the sensorimotor periphery (is "more concrete") has a head start in this process.

Notice that I do not deny the existence and exhaustive use of mappings in thought as posited by CMT. It is plausible that such mappings, which I suggested can be implemented in the CIV as categorisation relations, are pervasive because they are good cognitive heuristics. Such heuristics consist in translating a problem into a structurally analog domain that is "more concrete", solving the problem there, and then translating the solution back into the more abstract domain. I merely suggest that according to the PP picture, we do not need to assume that the comprehension (or representation) of a target concept depends in some stronger form on the source concept.

Despite this central discrepancy, note that both CMT and PP share a very crucial point (which is not shared with RT). CMT is largely responsible for the success of the embodied cognition paradigm more generally. The idea that the body significantly shapes our conceptual apparatus is fully taken on board by PP, however in a different guise. The proposed PP account posits a continuous abstraction/convolution hierarchy grounded in the lowest sensorimotor layer denying a dichotomy between the concrete and abstract. CMT, in turn, speaks of a mapping between two qualitatively different formats: concrete (i.e., sensorimotor grounded) and abstract concepts with a dependency relation of the abstract on the concrete. In sum, there are some tensions with RTM and the ICV, specifically CMT. Overall, PP seems to better support a version of the CIV, namely Glucksberg et al.'s Interactive Property Attribution Model.

9.6. Hybrid theories of metaphor

I have pointed out that the available evidence only slightly favours CIV (Holyoak & Stamenković, 2018; Rodríguez Ronderos, 2021). Given the lack of a clear winner so
far, some proposals for CIV-ICV hybrid theories have emerged. If PP does not underpin ICV well and a hybrid is needed, does this speak against the PP account? Here I argue that the PP account can accommodate the features that have motivated the CIV-ICV hybrids. In other words, I argue that the PP account of metaphor allows for avoiding a hybrid account.

Note that the only hybrid account that has been proposed that specifically involves the Interactive Property Attribution Model is the "Career of a Metaphor" theory (Bowdle & Gentner, 2005), which combines it with the ICV. According to this view, as a novel metaphor is conventionalised, the mode of processing shifts from comparison/mapping to categorisation (p.94). The Career of Metaphor theory is, however, judged to have weak empirical support (Holyoak & Stamenković, 2018), so I will not discuss it further here.

Other available hybrids, however, combine RTM as an instance of CIV. As the instance for ICV they use CTM (Gibbs & Tendahl, 2006; Tendahl, 2009; Tendahl & Gibbs, 2008; Stöver, 2010). I suggest that the PP account can provide a unified underpinning for those aspects that are thought of in the RTM/CTM hybrids as dichotomic and complementary. PP is, of course, not equivalent to such a hybrid (the above-mentioned tensions with both RTM and CTM still remain), but it can accommodate those aspects that have motivated an RTM-CTM hybrid. In the RTM-CTM hybrids, in a nutshell, two complementary aspects are emphasized. Firstly, RTM focuses on the context-sensitive modulations, while CMT focuses on entrenched conceptual metaphors. Secondly, RTM accounts for propositional representations (i.e., amodal ones), while CMT accounts for sensorimotor image-schema (i.e., modal representations).

Regarding the first aspect, as already emphasised, the PP model can account well for context-sensitive modulations of the senses of concepts. But entrenched mappings between concrete and abstract representations can also be accounted for, namely, as sketched above, as categorisation relations (priors) with relatively higher prior probabilities compared to novel or less entrenched mappings. So, PP can cover both elements, the flexibility of ad-hoc categorisation and more stable mappings that in the RTM-CTM hybrids require two different theories of metaphor.
Regarding the second aspect, the PP model is one that builds on modality specific representations, so what about amodal representations, i.e., linguistic representations? The purpose of the RTM-CTM hybrids is also to accommodate the fact that metaphors can arise both from thought (CTM) and from language (RTM) (see Stöver, 2010). But, as already mentioned, the PP model of metaphor can also account for linguistic representations of the sort emphasized by RT. An account of how this works in detail would be beyond the scope of this chapter, because we would have to lay out an account of language within PP. Therefore, I limit the response to a brief sketch that builds on Michel (2019, 2020b) and Rappe (2022).

The idea is, following dual processing models (e.g., Paivio, 1990; Simmons et al., 2008; Dell & Chang, 2013), to posit two modality-specific hierarchies of PUs, one perceptual-conceptual one (consisting of modality-specific world knowledge, both perceptual and conceptual) and one linguistic one (consisting of modality-specific representations of linguistic forms, which includes representations of letters, word forms, syntactical rules, distributional statistics, etc.). The hierarchies are laterally connected such that we get "symbolic pairings" of, for example, word form and word meaning. For instance, the PU representing the word form "cat" from the linguistic model is then connected to the PU representing the concept CAT. In this way, we can represent sentences, which express propositions. Note that the horizontal mapping between the two hierarchies is not complete, because there are PUs in the perceptual-conceptual model that have no lexical symbol (e.g., consciously not accessible non-lexicalised sub-symbolic representations), and there are PUs in the linguistic hierarchy that have no perceptual-conceptual partner (e.g., gibberish words and jabberwocky sentences). The important point to stress here is that PP can provide an account of linguistic representations, which are considered generally to be amodal, merely with modality-specific resources.

I conclude that the features that are dichotomic and complementary in the RTM and CTM account, and require an RTM-CTM hybrid approach, are available in a unified way in the PP account of metaphor. Hence, we can avoid a hybrid approach to metaphor processing.
9.7. Conclusion

This chapter has aimed to contribute to metaphor research by linking the dominant theories of metaphor processing with an emerging but increasingly popular neurocognitive-computational framework, predictive processing (PP). I have argued that a specific account of concepts within the PP framework supports the Category Inclusion View (CIV) more than the Implicit Comparison View (ICV) as a cognitive-computational underpinning.

This finding has various implications, which are relevant both for metaphor research and the PP research programme. Firstly, PP and CIV mutually support each other. Whatever independent support accrues for CIV, it makes PP more plausible because no theory of cognition that aspires, as PP does, to be a unified account of cognition can ignore the higher-level cognitive phenomenon of metaphor; and vice versa, as more evidence for the PP model accrues, this supports the CIV given that PP can serve as its cognitive-computational underpinning. PP also provides resources to extend the narrow scope of metaphor types for which CIV has been formulated so far. The PP approach has the additional and related advantage that it makes unnecessary a hybrid approach. Secondly, by embedding CIV into the PP framework, new ways are made available to approach metaphor research, by leveraging the concepts, resources and constraints provided by PP. For instance, the PP framework can contribute by making specific neuro-architectonical predictions about how concepts are represented and modulated during metaphor processing, because it includes commitments on the neuro-implementational level.
Chapter 10. Copredication in context: a predictive processing approach

(co-authored with Guido Löhr)

Abstract

We propose a cognitive-psychological model of linguistic intuitions about copredication statements. In copredication statements, like 'The book is heavy and informative', the nominal denotes two ontologically distinct entities at the same time. This has been considered a problem for standard truth-conditional semantics. In this chapter, we discuss two questions that have so far received less attention: What kinds of word representations and cognitive mechanisms are responsible for judgments about the felicitousness of copredication statements? Relatedly, why can similar copredication statements have different degrees of felicitousness? We first propose a cognitive-computational model of copredication within the predictive processing framework. We then suggest that certain asymmetries in felicitousness judgments can be modeled in terms of a set of expectations that are influenced by higher-order priors associated with discourse context and world knowledge.

Keywords: copredication; linguistic intuition; felicitousness judgment; predictive processing

10.1. Introduction: what is copredication and what is the problem?

The term 'copredication' captures the phenomenon that we can use a single nominal to denote two or more distinct kinds of entity in the same statement. To illustrate, consider the following examples:

(1) The manager entered the bankrupt bank.
(2) The heavy book is informative.

In the case of (1), a single noun 'bank' is copredicated by 'entered' and 'bankrupt'. While the predicate 'entered' is intuitively taken to apply to the building of the bank,
the predicate 'bankrupt', in this context, is meant to apply to the more abstract financial institution whose existence conditions are independent of the concrete building that hosts it. In the case of (2), the physical copy of the book is said to be heavy while only its content can be informative.

The phenomenon of copredication poses a challenge to standard truth-conditional semantics (cf. Chomsky, 2000; Collins, 2017; Pietroski, 2018). It is not clear what the denotation of a term like ‘bank’ could be such that it refers to the building of the bank if combined with ‘entered’ and the abstract institution if combined with ‘bankrupted’. If we restrict the meaning of ‘bank’ such that it refers only to the institution, it ceases to be clear what the truth conditions of (1) could be, considering that we cannot literally walk into an abstract entity. Similarly, it is not easy to see what the reference of ‘book’ in (2) could be, considering that a physical object (the book with pages and cover) cannot literally be informative, and the informational content of the book cannot literally be heavy.

It is currently debated what the phenomenon of copredication tells us about the meaning of words like ‘bank’ and ‘book’ and how these terms relate to truth-conditional semantics. One option is that nominals like ‘book’ refer to objects that can literally be informative as well as heavy (Liebesman & Magidor, 2019). Another option is that ‘book’ refers to a complex object (e.g., Gotham, 2017) consisting of the informational content of the book as well as a physical object with pages and a cover. A third option is to reject the claim that words have a reference, and that linguistics can tell us anything about the nature of books or banks (Pietroski, 2018; Chomsky, 2000).

While there has been significant recent interest in copredication with respect to semantics and ontology, an important question is much less discussed (cf. Ortega Andrés & Vicente, 2019): What are the cognitive psychological mechanisms underlying the processing of copredication statements that give rise to acceptability intuitions? An answer should allow us to address questions like: How can we model or account for certain asymmetries in felicitousness judgments? For instance, why do two copredication statements involving the same pair of senses of the nominal produce different intuitions? Why can the order of the predicates or the discourse
context alter our felicitousness judgments in the case of copredication?\textsuperscript{102} Note that by addressing those questions, we do not aim at a philosophical theory of linguistic meaning. Instead, we develop a plausible mechanistic cognitive-computational model of our felicitousness intuitions with respect to copredication.

The key question we pursue in this chapter is then this: How can we best model our linguistic intuitions based on which philosophers and linguists often draw conclusions about (psychologically significant) word and sentence meanings as well as speakers’ ontological commitments? We will build on work by Ortega Andrés & Vicente (2019) who argue that copredication statements sound felicitous to us because words elicit a body of information or a "coactivation package" that contains the information needed to understand words and sentences. The reason why (2), for example, sounds felicitous is that ‘book’ makes available information about both the abstract content of the book and the physical object that contains the content. We pick up on this idea but supplement it in two important ways.

First, we integrate it into the so-called "predictive processing" framework. Predictive processing (PP) pictures the mind as a "prediction machine". Contrary to a traditional view of cognition, the predictive processing model construes perception and cognition in general and linguistic understanding specifically (see, e.g., Pickering & Garrod, 2013), not as merely a passive interpretation of sensorimotor input but as involving active predictions of this input. We argue that information packages can be understood as "expectation hierarchies". These are complex networks of representations that correspond to expectations at different levels of complexity and abstraction. We call those expectations "priors". Based on this view, we propose a model of the mechanism underlying linguistic acceptability intuitions.

Second, we go beyond Ortega Andrés & Vicente (2019) by tackling a problem with their information package approach: Their approach has difficulty accommodating asymmetries with respect to acceptance intuitions. Why do some copredication statements seem to be grammatical, and hence are grammatically acceptable, but it would require further (possibly contentious) assumptions to relate semantic acceptability and felicitousness. This question is highly interesting, but we need not to commit for current purposes to any view of how the semantic and the psychological notions of acceptability are related. Thanks to an anonymous reviewer for recommending distinguishing those notions more clearly.

\textsuperscript{102} Note that we focus on "felicitousness judgements", i.e., one specific, psychological, notion of acceptability. There are others, namely grammatical and semantic (in the sense of truth evaluability) acceptability. The relation between the three notions of acceptability is a highly contentious question. Copredication statements seem to be grammatical, and hence are grammatically acceptable, but it would require further (possibly contentious) assumptions to relate semantic acceptability and felicitousness. This question is highly interesting, but we need not to commit for current purposes to any view of how the semantic and the psychological notions of acceptability are related. Thanks to an anonymous reviewer for recommending distinguishing those notions more clearly.
statements sound infelicitous even though the same information is available as in similar felicitous statements? Our model improves the information package approach by introducing different prediction layers and by modelling discourse context. So, while, Ortega Andrés & Vicente (2019) assume that information packages are relatively rigid conceptual structures, our model allows for significant context sensitivity, which is not only independently supported by empirical evidence but also accounts better for our linguistic intuitions.

This chapter is structured as follows. In Section 10.2, we discuss the information package approach to copredication in more detail and point to a drawback of the account of Ortega Andrés & Vicente. In Section 10.3, we introduce the core ideas of the predictive processing framework and propose an account of "information packages" in the form of "expectation hierarchies". In Section 10.4, we cover cases of felicitous and infelicitous copredications and discuss the context-sensitivity of felicitousness judgments. We also briefly discuss coprediction order effects (Murphy, 2021a, 2021b) and how they could be modelled with our framework.

10.2. The information package approach to co-predication

It is largely uncontroversial – and we do not depart from this view – that word forms make available so-called "bodies of information" or "information packages", which inform our linguistic intuitions (e.g., Machery, 2009; Vicente, 2021). When the hearer encounters a copredication statement, her cognitive system must combine these information packages in a way that generates intuitions about sentence meanings. We call this the information package view. A philosophical theory of copredication explains how this body of information relates to linguistic meanings. In contrast, a psychological model of copredication specifies the structure and the cognitive content of such information packages and how the system processes them.

Ortega Andrés & Vicente (2019) recently applied the information package view to copredication (see also Vicente, 2021). According to the authors, words make available so-called "coactivation packages" that contain "senses" or pieces of information, which are closely related by "explanatory, realization relations" (2019, p.14). For example, the word 'book' allows for copredication with the content/physical object alternation in "The heavy book is informative" because both senses are
available in the same coactivation package of ‘book’. The interpretation of copredication works then by pairing each predicate with one of the two co-activated senses (2019, p.5).

The information package view addresses at least a part of the question of why some combinations are felicitous while other combinations are not. Why can we say that the book is heavy and informative? The answer is that the information immediately available after hearing ‘book’ in each context contains both relevant senses: the book as content and as a physical object. This is also the reason why we do not consider ‘heavy’ in this context to be used metaphorically, e.g., heavy in the sense of sad or intense. Since we know that books have both content and a physical realization, we assume, in the right context, that the speaker means that the physical book is heavy, not its content.

While we do not disagree with the overall approach of the coactivation package view to copredication, there remain two crucial open questions that we want to address: What determines which senses or information are selected and combined in any given context? Why does the same information package allow for some combinations but render other similar combinations infelicitous? Consider the following statements:

(3a) The newspaper has been attacked by the opposition and was publicly burned by the demonstrators.

(3b) The newspaper has been attacked by the opposition and fell off the table.

(4a) The bank went bankrupt and issued a statement.

(4b) The bank went bankrupt and flew to the Cayman Islands

According to the coactivation package view, (3a) and (3b) should be equally felicitous. ‘Newspaper’, in this view, makes available in both statements the same coactivation package consisting of the senses newspaper-producer and newspaper-copy. But why does (3a) sound more felicitous than (3b)?

104 Similarly, why does (4a) sound better than (4b)?

---

103 This example is from Geoffrey Nunberg (in Copestake & Briscoe, 1995, p.55). It is a widely discussed example in the copredication literature. See also Ortega Andrés & Vicente (2019).

104 A referee pointed out that (3a) does not seem to be much better than (3b). As we will make clearer later, we do not deny that there are interpersonal differences in felicitousness judgments, we even expect this to be the case. The example (3a)/(3b) is a standard example discussed in the literature by
sound better than (4b) if in both cases the relevant coactivation package is the same and, therefore, the senses bank-institution and bank-staff are both readily available?

This problem has recently been observed by Collins (2017, p. 691) in response to theories that take polysemous words to refer to complex objects or dot-objects:

*The Times*, let’s suppose, has the dot-object [material-institution], but *The Times made most of its revenue from advertising and blew away* is badly zeugmatic. I take it to be an outstanding problem for any account of polysemy in general and copredication in particular why some constructions are acceptable, finding a ready interpretation, whereas others are zeugmatic.

The question of which statements sound acceptable and which statements sound odd is a psychological one – whatever the meaning of words turns out to be. So, we focus here on cognitive processing and subjectively experienced linguistic intuitions and do not discuss the nature of linguistic meaning understood more abstractly (e.g., as meaning as it figures in truth-conditional semantics). In other words, we take a cognitive psychological stance and do not commit to any philosophical theory of meaning. We argue that the predictive processing framework, which is an increasingly influential framework in cognitive science, can help us to model the cognitive processes underlying the processing of copredication statements.

We also worry that Ortega Andrés & Vicente’s coactivation package view tends to rely too much on a rather traditional view of linguistic processing. First, linguistic stimuli are processed in a relatively passive way by feature aggregation, for instance, from basic visual features (say a pixel pattern in the retina) to more complex visual shapes, to some semantic representation. Second, Ortega Andrés & Vicente’s account uses a relatively rigid knowledge structure consisting of discrete pieces of information, namely senses or concepts, as components that are coactivated by default (see Machery, 2009, for this kind of invariantism).105 This view arguably leaves little room for context sensitivity. The mainstream in cognitive science, however, is moving toward more flexible, context-dependent structures (Pulvermüller, 2013; Casasanto & Lupyan, 2015; Kiefer, 2018; see also Löhr, 2017; 105 Note that Vicente and Ortega-Andrés do not commit to this picture and rather take it as an initial working hypothesis (A. Vicente, personal conversation).
Moreover, expectation or prediction-based models, as advocated by predictive processing, are becoming more and more influential in language processing.\textsuperscript{106}

Considering recent evidence and theoretical support for context dependent bodies of information (e.g., Barsalou, 2009, 2011; Hoenig et al., 2008; Lebois et al., 2015; Ludlow, 2014; Ludlow & Armour-Garb, 2017; Michel, 2020\textsuperscript{a}), we suggest that different senses are not merely simultaneously activated. Rather, we argue that an information package can be understood as an expectation hierarchy defined by its \textit{root node}. Different subparts of that information package are made available depending on the context. Copredicated nominals then do not select different senses, each of which is adequate for one predication. Rather, a certain portion of the information package that is selected can be used for both predications. In other words, we suggest that successful copredication does not involve pairings of co-activated senses with predicates. Instead, it involves the selection of one single and sufficiently abstract representation that is compatible with both predicates.

\section*{10.3. Predictive processing and information packages}

\subsection*{10.3.1. The PP framework}

Predictive processing (PP) is a neurocognitive-computational framework that construes the mind as entertaining a \textit{hierarchical probabilistic generative} model of the world with which it continuously predicts its sensory input (Clark, 2013, 2016; Hohwy, 2013, 2020; Friston, 2010). In contrast to more traditional views of cognitive architecture, the mind is not viewed as a mere passive analyst of incoming stimuli, but as an active prediction machine. This mental model is continuously being improved based on the processing of its prediction errors with the aim of minimizing these prediction errors in the long run and on average.

The PP model is \textit{hierarchical} in the sense that it is composed of various layers of representations. Each layer generates expectations or predictions, which are compared to the signals from the level immediately below. This prediction cascade

\textsuperscript{106} By “language processing” we mean processing pertaining to language generally construed: syntax, semantics, and pragmatics.
extends from the highest levels in the neocortex down to the lowest-level sensorimotor areas. The higher-level layers make more abstract predictions, i.e., predictions with less detail corresponding to more compressed information, and represent patterns of larger temporal and spatial scales. In this way, the model tends to replicate the causal structure of the world, which is the source of the sensory signals. Because the brain's model generates proactive hypotheses about the sensory causes, it is also called a *generative* model.

The layers contain *prediction units* each of which consists, on a neural level, of a pair of a so-called *representation unit* and an *error unit*. The prediction unit generates a prediction signal that is fed downwards and an error signal that is fed upwards in the hierarchy of representation layers (see, e.g., Kanai et al., 2015; Bastos et al., 2012; Keller & Mrsic-Flogel, 2018; Weilnhammer et al. 2018). The error signals provoke updates of the model to reduce future prediction errors.

The system also contains an *error weighting mechanism*, which uses estimates of the precision of the signals to tune the error signals up or down. We do not want the brain to update the model based on unreliable sensory information; therefore, this mechanism can suppress error signals generated by unreliable input. For instance, in a foggy environment, we can rely less on our visual input (we might more easily mistake a cat for a dog). Therefore, we need to give greater consideration to prior experiences and knowledge (i.e., expectations) in our judgments. The error weighting mechanism therefore regulates how we should balance expectations and sensory evidence depending on the situation.

The model is *probabilistic* because it represents states of affairs as probability distributions and carries out approximate Bayesian inference (which is realized by prediction error minimization). Representations at level N play the role of predictions or priors for representations at level N-1. Through Bayesian inference, prior beliefs are updated to posterior beliefs to better match the evidence. Then the posterior beliefs become the new prior beliefs. If the evidence matches the predictions (priors) in all layers, then the prediction error is minimized, and the brain achieves a temporal error minimizing state.

The revisionary element of PP is that what we perceive, grasp, or represent as being the case (linguistically or non-linguistically) is materialized as predictions. I perceive
an apple on the table, and grasp (and believe) this fact, because my model has predicted an apple on the table – a situation that is compatible with the incoming sensory input and other prior beliefs. We could say that the brain constantly hallucinates, but those hallucinations turn out to match the environment well (under normal circumstances). Similarly, if I hear that you said that you have a new cat because given the context and the stimuli received, this interpretation is most coherent with my network of expectations.

10.3.2. Information packages as expectation hierarchies

We want to apply the PP framework described in the previous section to model copredication sentences. Central for our proposal is how "information packages" associated with a word are structured. Once we have such an account in place, we can show how felicitousness intuitions arise via the violation of the expectations encoded in the information package of the relevant nominals.

We argue that the PP view provides a novel and empirically plausible model of information packages. An information package associated with a word on the PP picture, as we propose, consists of nodes (in squared brackets) that are connected in a hierarchical tree-like structure that we call the "expectation network" (see Figure 10.1). The expectation network is identified by its root-node, i.e., the prediction unit "at the top" and from which all other lower level (i.e., more specific and concrete) child-nodes are connected forming the hierarchical structure. All these nodes correspond to the above mentioned "prediction units" implemented as neural assemblies. Let us explain more slowly the structure of an expectation network with the example 'book'.

Assume that we have a node [BOOK] in the highest (most abstract/compressed) layer of the hierarchy. This node is the root-node of the whole information package related to the entity denoted by the word 'book'. All (lower level) child nodes emanating from [BOOK] form part of the information package of 'book'. For instance, assume [BOOK] is connected to the lower-level child nodes [PAPER_BOOK] and [E_BOOK]. The node [E_BOOK] is the root-node of the more specific information package E_BOOK. The node [is heavy] is the root-node of the information package of being heavy, which is connected to the node [PAPER_BOOK]. And so forth.
The crucial feature of the information package structure corresponding to the word 'book', i.e., the root-node [BOOK], is that it can relate to the lower-level nodes [E_BOOK] and [PAPER_BOOK] in terms of different probabilities. In a sufficiently neutral context, both nodes might be similarly probable. Different contexts or higher-order priors can modify this probability relation. E-books are expected more at an e-book fair while they are probably not expected in a traditional university library (again, this depends on one's previous experience and world knowledge). The higher the probability for a more specific kind of thing or event is, the more “concrete” the expected stimulus becomes.

When hearing ‘book’ in a sufficiently neutral context, our expectations remain rather “abstract”. For instance, again, if the relation between the lower-level nodes and the root-node of [BOOK] is rather balanced, then we have a very schematic representation of ‘book’. In specific contexts, [BOOK] might increase the expectations of specific sub-packages like [PAPER_BOOK] or [E_BOOK]. These sub-packages can be made even more specific by including additional nodes further down in the hierarchy. They become more specific once the expectation relation increases for them, while decreasing for the other nodes.

In PP terms, once [BOOK] is predicted, we are not surprised to hear or read about e-books or paper books. At the other extreme, if the path [BOOK]/[E_BOOK]/[occupies Bytes] is predicted, we have a more specific representation for 'book' where the features [E_BOOK] and [occupies Bytes] are cognitively salient. The depth of
prediction is context dependent. Context is represented in the model as a set of priors as well. When visitors at an e-book fair use the word 'book', the node \([E\_BOOK]\) receives a higher probability than \([PAPER\_BOOK]\). This is because some situational higher-level prior (e.g., \(BEING\_AT\_AN\_E\_BOOK\_FAIR\)) represents our awareness that we are at an e-book fair and makes available an expectation about e-books, not physical books. In a library, the librarian's use of the word 'book' might increase the probability of the prior \([PAPER\_BOOK]\), because the awareness is represented by, e.g., \(BEING\_IN\_A\_LIBRARY\), which is a prior that makes us expect paper books. When someone is wondering how much storage is needed, the sub-package \([E\_BOOK]/[occupies\_Bytes]\) receives a higher probability, and so forth. Any given information package, like \([E\_BOOK]\), is embedded in the overall network of priors constituting the brain's model of the world. Priors outside the information package \([BOOK]\) (like \(BEING\_AT\_AN\_E\_BOOK\_FAIR\)) constitute a context-sensitive influence on the package.

At this point, a question might arise about how those expectation hierarchies are constructed.\(^{107}\) This question reduces to the more general empirical question of how the PP model of any individual is built. What can be said here is that PP has certain empiricist tendencies in the sense that the mental world model is constantly adjusted to the sensory input from the world. One might expect that over time it will structurally correspond to the world, at least in aspects relevant for the survival of the cognitive agent. The world contains regularities on different spatiotemporal timescales that cause the barrage of signals that hits our sensory periphery. The hierarchical generative model is then a model in the form of interconnected prediction units representing the model variables relevant to predicting the sensory inflow (those variables we call priors or expectations). Despite the empiricist tendency, PP is also compatible with the view that many expectations are innate. After all, it is difficult to see how a model can get off the ground without at least some initial biases.

The core idea that we will develop in the rest of the chapter, is that copredication statements are felicitous if we can minimize the prediction error in the information package structures involved in the processing of these statements. But before we

---

\(^{107}\) Thanks to a reviewer for suggesting addressing this question.
can characterize in detail the felicitousness of copredications, we need to provide additional background on how prediction error minimization in language processing works according to the proposed PP model.

### 10.3.3 The holistic nature of prediction error minimization

It is crucial for our proposal to understand the interplay of top-down and bottom-up information flow\(^\text{108}\) in the PP model. There is a constant process of adjusting priors and suppressing error signals, such that the brain reaches a prediction error minimizing state in the *entire hierarchy*. Such a state then corresponds to a certain mental state, like a perception or a belief.

This view can best be illustrated by considering how our brain processes visual input. We must deal not only with incoming stimuli from the retina but also with expectations or priors (e.g., Rao & Ballard, 1999). Both types of signals must be related in such a way that they match. If a reliable external stimulus is unexpected, i.e., inconsistent with the priors, the priors are likely to be adapted such that the error signal will be minimized. If the external stimulus is estimated not to be reliable (e.g., in a dusty environment or when it is dark), the priors will be given priority, and the error signal will be suppressed.

Take as an example the visualization of faces. To recognize something as a face, information passes through various stages in the neural abstraction/compression hierarchy of the brain. Imagine looking at a screen that presents you with pictures of different faces. The brain will immediately make predictions as to what kind of stimuli you will be presented with. Here different layers of the brain are representing stimuli at different degrees of abstraction/compression. An initial neural layer in the retina represents a pixel field. In a subsequent layer, neuron assemblies can recognize pixel-patterns as elementary edge-forms. Higher in the hierarchy, we have representations of more complex lines and shapes. Finally, there is a layer in the visual cortex with neurons sensitive to faces (regardless of, e.g., specific light conditions, head positions, etc.) and that “assumes” that the incoming information is

---

\(^{108}\) With "top-down" and "bottom-up" information processing we mean the following. "Top-down" refers to the flow of predictions from higher to lower levels of the model hierarchy. In other words, it corresponds to the flow from levels with more abstract and compressed representations farther away from the sensorimotor periphery to layers closer to it. "Bottom-up" information flow corresponds to the error signals that project from layers closer to the sensorimotor periphery to layers farther away.
about faces. The face detector is so keen on detecting faces that it can make you see faces where there are none.

One may erroneously see a face when briefly exposed to a vague visual stimulus, e.g., a cloud. This can be explained by an expectation effect (e.g., Barik et al., 2019; Kok et al., 2013; Salge et al., 2020). At some point, however, you realize that what you are seeing is not a face given that you have background beliefs (priors), according to which it is highly unlikely that a cloud literally has a face. The two contradicting predictions (the sensory information that it is a face, and the prior expectation that it is not a face) produce an error signal that the brain needs to reduce. Given that one did not look carefully enough, the precision of the perception ("it is a face") is estimated to be low and the error signal is suppressed, and the prediction ("it is not a face") now prevails.

The representational structure and mechanism of predictive processing can also be applied to higher cognition like language and conceptual thought. Word recognition works like face recognition. Once a stimulus, say a certain two-dimensional shape drawn with a pencil, is recognized not as a face but as a familiar word, it immediately changes the probability of future perceptual stimuli. Clearly recognizing a shape as the printed word 'bird' in a certain context, say when sitting in a psychological experiment staring at a screen, will prime you to generate certain expectations, for example, of seeing other bird-related words or seeing or hearing bird sounds. The word 'bird' serves as a label for the corresponding information package [BIRD] and hence plays the role of a prior that generates expectations related to birds. Translated to the PP model, this means that an adjustment of priors is happening such that when something is seen after exposure to the word 'bird' and a bird is seen, the error signal is minimal. When something is seen and it is a horse, the error signal will likely be higher and the prior on a higher level needs to be changed to [HORSE]. This assumes, of course, that the horse is clearly seen, and the visual stimulus is therefore assigned a high precision estimate. The expectations related to the word stimulus 'bird' are strongly constrained by the mental expectation model, and of course, not everything bird-related is expected. When the screen suddenly disappears in smoke and a huge peacock appears instead, the surprise (and hence the prediction error signal) will be large.
Importantly for the present chapter, predictions are made all the time when processing statements in incremental steps. When having processed a statement partially, and having, for instance, recognized the incomplete sentence fragment, 'The train is coloured ...', the chances are high that a Dutch person will expect the next word to be 'yellow'. This is because in The Netherlands trains are typically yellow and, therefore, the Dutch person has a mental model which contains an information package [TRAIN] with the colour feature [YELLOW] as being highly probable. When the statement is continued with 'red', a Dutch person will likely be slightly surprised, i.e., a small error signal will be generated because [RED] is not a feature of the information package of [TRAIN] which has a high probability associated. The error can easily be minimized because, of course, a train can be red (e.g., in a different country, or in The Netherlands when we deal with a Coca-Cola promotional train). An even larger error signal should arise when the statement continues with 'sour' because while a train can be red even for a Dutch person, it cannot be sour (Hagoort et al., 2004). In this case, the information package [TRAIN] does not even have a feature representing smell properties, because people in general do not have taste experiences with trains. Given that we have here a sort of category mistake ('sour' cannot be applied to 'train'), we deal with a prediction error that is difficult to resolve. This example merely serves to illustrate the core idea of processing statements based on expectations, and the mechanism as applied to copredication will be explained in more detail later.

We cannot review here all of the available evidence for the general PP model of cognition. However, there is increasing support for the idea that predictions and expectations play a critical role in linguistic and non-linguistic thought. Neuroscientists have suggested that there is a neural marker, the so-called N400, which has been observed for violations of semantic and world-knowledge expectations (e.g., Hagoort et al., 2004). One advantage of the PP model is that there is no principled distinction anymore between semantic and world knowledge violations, which matches those findings and makes it a parsimonious account. More generally, there is mounting evidence that language is underpinned by predictive mechanisms as posited by predictive processing, on all levels of the linguistic

109 However, the interpretation of the N400 signal is much more complex than is suggested here (see, e.g., the overview in Kutas & Federmeier, 2011).
hierarchy: phonemes, words, sentences, and discourse (see, e.g., Kuperberg & Jaeger, 2016; Chow et al., 2018, p.804).  

At this point, we should emphasize two unique and critical features that the PP framework contributes to the information package account proposed. The PP framework supplies a model of constraints that we will need for our copredication account and for how they work computationally, namely by prediction error minimization in the network of priors. Further, PP is a holistic approach to cognitive processing, which naturally provides resources for an account of the context-sensitivity of the felicitousness of copredications. Context, as we will see, is operative in the form of priors outside an information package.

Before we turn to the PP account for intuitions of copredication statements, we need an additional critical and PP-specific ingredient. It consists of the assumption that we can read or listen to a statement in two processing modes, namely a "shallow" and a "deep" one.

10.3.4. Shallow and deep processing of a statement

Remember that one of the central commitments of the PP framework is a counter-current information flow: top-down predictions and bottom-up "evidence". Furthermore, PP posits a mechanism to regulate the influence of either of those two directions of processing. If incoming information is estimated to be unreliable (or irrelevant), then prior knowledge has more weight in the predictions. If sensory information is precise but does not correspond to the predictions based on prior knowledge, then the system tends to modify/update the higher-level predictions.

An idea central to PP is that the cognitive system can regulate whether it prefers updating the higher-level predictions or the lower-level prediction that serves as "evidence" for the higher-level prediction. This leads us to posit two distinct modes of processing a statement within our PP framework: shallow and deep processing. This distinction is inspired by the influential "levels of processing" framework in memory.

110 But we also acknowledge that PP is far from being a confirmed and mature neurocognitive framework and is still very much in the making. See, e.g., Walsh et al. (2020) for a review of the neurophysiological evidence for PP. The authors conclude that "Although the debate about PP's empirical grounding is currently unsettled, the theory can nevertheless be regarded as a milestone in cognitive neuroscience, spurring efforts to recognize the importance of backward connections in the architecture of the neocortex and the role of prediction in sensory processing" (p.262).
research (e.g., Craik & Lockhart, 1972; Craik, 2002). Depending on the specific task, the semantic information accessed when processing words can be more or less rich or detailed. This idea has also been taken up by Barsalou et al. (2008) (see also Simmons et al., 2008), who distinguish shallow, i.e., merely syntactic processing of language from deeper processing involving richer sensorimotor simulations.

Within the PP framework we posit a shallow mode of processing of a statement, in which the overall understanding of the situation expressed by this statement is prioritized. An overall situation is "understood" when we settle on a prediction in the form of a higher level situational prior. In this mode, we might reduce the influence of certain evidence to minimize the overall prediction error.

In the deep mode, what is prioritized is the priors representing the lower-level evidence for the higher-level hypothesis, here in the form of words and phrases. In this mode, we tend to hold fixed the lower-level evidence and update the higher-level prediction, to minimize the prediction error. In other words, in the shallow reading mode, we care about the overall gist of the situation described. In the deep reading mode, we care about the detailed understanding of the words, their denotations, and how they fit together into phrases, etc.

The PP apparatus supplies tools for modelling those two modes on a cognitive-computational level through an attention mechanism. Attention is often cashed out in PP as increasing the error-signal sensitivity of the relevant domain (see, e.g., Feldman & Friston, 2010; Hohwy, 2012, 2013). If we attend to the individual words and phrases, we increase the error-signal sensitivity on the level at which words are represented. If we attend to the situational gist, we increase the error-signal sensitivity on the level where situational patterns are predicted.

This distinction is psychologically plausible and receives further support from Kahneman's findings that the brain "operates as a machine for jumping to conclusions" with the aim of creating a coherent overall story (e.g., 2012, p.85). In fact, one can compare this to Kahneman's famous distinctions between "System 1" and "System 2" thought processes. System 1 is unconscious and quick and might correspond to the shallow processing mode. The objective is to quickly "jump to conclusions" about the overall situation. System 2 is conscious and effortful and

---

111 Thanks to Beate Krickel for bringing this to our attention.
might correspond to the deep reading mode. In this mode, we pay careful attention
to the relevant distinctions that give rise to the impression of polysemy.

Characteristic of shallow reading of a statement is that certain details concerning
words (or grammar) are disregarded given that what is prioritized is the gist of the
situation. This has various important implications. Firstly, it allows us to perfectly
understand statements with wrong words (malapropisms, e.g., Davidson, 1986) or
with grammatical errors. Those errors often go unnoticed. For instance, consider
"Moses sentences" (see Erickson & Mattson, 1981, or Barton & Sanford, 1993).
People tend to answer the question "How many animals of each kind did Moses take
on the Ark?" with "two". They overlook that it should say "Noah". It seems that when
shallowly reading the Moses sentence, "Noah" is represented rather "sloppily" (in
more neutrally terms: "flexibly" or even better "abstractly") as "some biblical person".
This more abstract interpretation of "Moses" is enough to grasp the gist of the
situation.

The tendency to "jump to conclusions" on a situational level is highly natural and
probably essential given that we are embodied minds that need to survive in an
uncertain environment, and, hence, need to deal with all kinds of situations all the
time. In the deep reading mode, on the other hand, the details of a statement, i.e.,
words and phrases, are prioritized over its overall gist. We read more carefully and
conscientiously with awareness, e.g., of denotational nuances of words. But then it
can happen that we do not manage to integrate the words into an overall sentence
meaning. We might understand each word, but we do not understand the whole
sentence.

What the system tries to achieve is an optimal balance between the two modes of
processing. If too much focus is placed on the detail, i.e., words and their exact
"sense" and how they combine with predicates, one might not see the forest for the
trees, i.e., one might not comprehend the complete sentence. If we are too sloppy
with respect to the details, the way we end up interpreting the statement might have
little to do with what the statement actually says, which may harm
communication. When we read difficult texts, we often alternate between the two

---

112 See also Fillmore (1975) for the idea that we grasp a scene description holistically, before we
grasp all the details.
modes. We try to understand in detail some complicated sentences and then step back to grasp the overall message or big picture. The same is true on the statement level with regard to words.

We now have the two essential ingredients, derived from the PP framework, namely the expectation hierarchy structure and the two modes of reading in place. These allow us to formulate a model for intuitions about the felicitousness of copredication statements.

### 10.4. A predictive processing approach to copredication

In this section, we work out the PP-based approach to copredication. We proceed in three steps. First, we discuss cases of felicitous copredications and provide a characterization of felicitousness (4.1) within the PP framework. Then we discuss examples of infelicitous copredications (4.2). Finally, we discuss how the PP framework can model the fact that felicitousness intuitions are context-dependent (4.3). This suggests an answer to the problem of asymmetric felicity intuitions that is unanswered by Ortega Andrés & Vicente's coactivation package account.

Notice that this section presents a model of a cognitive architecture underpinning copredication. The plausibility of such a model does not depend on whether the reader finds the reported acceptability intuitions convincing. We focus on examples from the literature, and it may be that some readers have different intuitions. This should be reflected by the way their individual acceptability intuitions are modeled.

On an abstract level, felicity judgments are based on world knowledge and innate constraints (or an interplay of the two), which are reflected in our individual cognitive architecture. Note also that intuitions regarding acceptability are not always clear-cut. We incorporate this idea by referring to different degrees of acceptability (see, e.g., Murphy, 2021a).113

**10.4.1. Felicitous co-predication**

Consider the following example:

---

113 Murphy provides a wealth of empirical data about the acceptability ratings of many variations of copredication statements. Note that Murphy focuses on aggregate statistics. We are here interested in a mechanistic cognitive account that can also accommodate differences of intuitions across individuals.
(5) The school caught fire while it was celebrating 4th of July.

We argue that copredication statements are felicitous in the PP model if our mind can adjust the network of priors such that prediction errors are suppressed. In the case of (5), for the nominal 'school', we apply a single and more abstract prior that can be combined with, say, 'caught fire' and 'was celebrating'. Therefore, 'school' is not interpreted as two different entities but as a single prior that is more abstractly represented and that allows us to expect the two more specific priors. Only after more careful deliberation and conscious analysis (what we call "deep reading") do we realize that 'school' might denote two different entities: one is a school building that can catch fire – the other is an institution whose anniversary it is and whose members can celebrate.

Fleshed out in more detail, we take there to be an information package with the root-node [SCHOOL] that is the prior of the two more specific sub-nodes [SCHOOL_BUILDING] and [SCHOOL_INSTITUTION]. Let us assume that [CAN_BURN] and [CAN_CELEBRATE] are child nodes of [SCHOOL_BUILDING] and [SCHOOL_INSTITUTION], respectively. When we read 'school' in the shallow processing mode, the error-sensitivity (i.e., attention) is increased for the more schematic prior [SCHOOL]. It is decreased for the more specific [SCHOOL_BUILDING] and [SCHOOL_INSTITUTION]. [SCHOOL] is a prior for its child nodes, which in turn are priors for the nodes representing the predicates. Hence the copredication is successful as no expectations are violated. [SCHOOL], [CAN_BURN] and [CAN_CELEBRATE] are part of the same expectation hierarchy and [SCHOOL_BUILDING], [SCHOOL_INSTITUTION] and — by "transitivity" — the predicates [CAN_BURN] and [CAN_CELEBRATE] are expected to a similar degree in this mode (this is because of the reduction of the prediction error-sensitivity below [SCHOOL] which creates a degree of "indifference" among the child nodes). By processing in the shallow mode, we quickly get an overall “good enough” (Ferreira & Patson, 2007) understanding of the statement (here: a vivid scene where a fire interrupts the school's celebration). This scene is represented as a strong higher level situational prior that now influences word processing.

Now turn to a more careful and detailed word-by-word reading or what we call "deep processing". What changes compared to the shallow mode of processing is simply that we focus more carefully on the words (nominals and predicates) and now realize
that two different kinds of entity need to be combined with each predicate. In the deep processing mode, the reader focuses her attention on individual words/phrases and each combination of the nominal/predication. This increases the influence of the individual words and de-emphasizes the overall situational understanding of the statement. In the PP literature, the focus is, again, often cashed out as increasing the error-signal sensitivity of the relevant domain (see, e.g., Feldman & Friston, 2010; Hohwy, 2012, 2013). When the reader detects the word 'school', the focus is on the word 'school' and its combination with the first predicate, that of a fire. In this context, hearing the word 'school' strongly increases the error-signal sensitivity of \[\text{SCHOOL}_\text{BUILDING}\] and reduces the expected probability of hearing about schools as an institution. This means that what is expected next is something that has to do with a building.

When we continue reading, we encounter the predicate denoting the celebration. This is unexpected because \[\text{SCHOOL}_\text{BUILDING}\] is still the most expected prior without being a prior of \[\text{CAN}_\text{CELEBRATE}\]. 'School' now needs to be modulated to the prior \[\text{SCHOOL}_\text{INSTITUION}\] such that we expect this predicate (minimize prediction error). This process is sometimes (in especially unexpected cases) manifested by a feeling of oddness.

But even in the deep mode of reading, statement (5) is not entirely incoherent and manifests some degree of felicitousness. This, as we have already suggested above, is because the human mind tends to aim at a situational, i.e., statement-level understanding, and a detailed word-level understanding is only instrumental. This is plausible in the PP framework. According to some PP theorists, mental predictions serve only one purpose: to secure the survival of the mind-body system which is thrown into and interact with an uncertain world (Clark 2016; Friston, 2010). The dominant level of representation must therefore be the level of situations. Even when we read (5) in the word-by-word deep mode, i.e., extremely carefully and very reflectively, we cannot escape the automatic force that drives us to interpret the statement on a situation level; we grasp intuitively and immediately that we are dealing with a scene of a celebration that is spoiled by an unfortunate fire.

Let us examine another example of what is considered to be a felicitous co-predication to further illustrate the model (Ortega Andrés & Vicente, 2019, p.16):
(6) Brazil [place] is a large piece of land & Brazil [people] is Portuguese-speaking & Brazil [government] is a republic & Brazil [economic system] is very high in inequality & Brazil [football team] is always first in the FIFA rankings.

(6) is perfectly well understood, and—if special attention is not drawn to them—the fine distinctions in terms of the "senses" of "Brazil" indicated in brackets go unnoticed. Furthermore, the following copredication statement, which only differs from (6) by mentioning the nominal "Brazil" only once, appears to be felicitous:

(7) Brazil is a large piece of land, Portuguese-speaking, a republic and is very high in inequality and always first in the FIFA rankings.

According to our model, when hearing (7), the mind engages in shallow reading and is not aware at all of the fine-grained sense distinctions. In the shallow reading mode, the influence of the lower-level priors of the prediction is reduced (less attention is paid to the details) and the sensitivity of the [BRAZIL] node and higher-level situational priors is increased (more attention is paid to the big situational picture). If the sensitivity of the lower level nodes [BRAZIL_PLACE], [BRAZIL_PEOPLE], etc., were high, then the mind would struggle to integrate the statement (in PP terms: to predict the overall situation). Only by careful reflection and word-by-word analysis of the statement, i.e., in the deep reading mode, might 'Brazil' be modulated into the more concrete and specific priors [BRAZIL_PLACE], [BRAZIL_SOCCER-TEAM], etc.

We can summarize the case of felicitous copredications with the following condition\textsuperscript{114} (generalizing from two to \( n \) predications):

\textbf{Felicitous Copredication Condition (FCC):} There is a prior [N] corresponding to the nominal which has \( n \) child nodes [N1], [N2], ..., [Nn]. Those \( n \) child nodes in turn serve as priors for the \( n \) predicates respectively.

When FCC is fulfilled for a copredication statement, shallow reading can succeed and provides us with a felicitousness intuition. However, note that sometimes shallow reading fails even if FCC is met, namely in cases where the statement expresses saliently a \textit{spatial or temporal separation} of the entities that the different

\textsuperscript{114} This condition is an informal summary and is not meant to be a "formalization" of our account in terms of the precise necessary and sufficient conditions.
senses represent (see example 5b in Section 10.4.2). But this is exactly what we should expect to happen.

10.4.2. Infelicitous copredication

From FCC one can derive a condition for infelicitous copredications, namely simply by its negation: if FCC is *not* fulfilled, then a copredication is infelicitous. As it turns out, copredicative statements might be infelicitous for a range of reasons. However, we want to focus only on the interesting cases where infelicity has to do with the existence of different "senses". Consider again the following copredication statements that are considered infelicitous in the literature (e.g., Collins, 2017; Vicente, 2021).

(3b) ?The newspaper fired the editor and fell off the table.
(5b) ?The school caught fire when it was on excursion.

In the case of (3b), the infelicitousness depends on the existence of two different senses of "newspaper", newspaper-institution, and newspaper-copy. On our account, what is missing here is a common, more abstract parent prior [NEWSPAPER] for [NEWSPAPER_institution] and [NEWSPAPER_copy] (contrary to [NEWSPAPER_copy] and [NEWSPAPER_content] which do have such a prior). When we think of a newspaper as an institution, the scenario that its product falls off the table is highly unexpected, and no obvious adjustment of priors is possible to minimize the prediction errors.

Similarly, statement (5b) is infelicitous because once the first part has been grasped the second part is unexpected. The phrase ‘caught fire’ is expected by the prior [SCHOOL_building], while ‘was on excursion’ is expected by priors like [SCHOOL_faculty] and [SCHOOL_students]. Those three priors are child nodes of [SCHOOL]. So, the statement fulfills FCC. However, a modulation of 'school' towards the more abstract [SCHOOL] through shallow reading is blocked here, as opposed to example (5). We cannot easily ignore the sense distinctions through shallow reading. It is precisely by grasping the overall situation/scene (which is the whole objective of shallow reading) that we become aware of the two different senses. It is salient in the statement (5b) that [SCHOOL_building] and [SCHOOL_students] are separate entities, precisely because the statement's content explicitly expresses spatial separation.
So, we cannot suppress that 'school' refers to different things (building versus people) by just representing the more abstract [SCHOOL] and using it for both senses at the same time.

We suggest then that a co-predication is *infelicitous* if a conflict between the expectations evoked by the statement cannot be resolved, i.e., a significant prediction error remains. This persisting prediction error exceeds a threshold such that we become aware of it, leading to the infelicitousness intuition. In *felicitous* copredications, we shallowly process a structure consisting of a more abstract parent prior and two (or more) different child nodes which in turn are priors for the predicates. The modulation of the attention towards the parent, rather than the child priors "resolves" the clash of expectations.

Couched in PP-terms, infelicity is a consequence of the failure of the brain to settle the network of priors in an error-minimizing equilibrium. Error signals can be minimized by changing priors in adequate ways. However, priors cannot be modulated arbitrarily because the relations between the priors on different levels are expectations, i.e., constraining relations. Certain priors with high precision estimates are less "flexible" than others. Therefore, the configuration of priors can turn out to be such that prediction errors above a threshold remain because of inconsistencies between priors. This generates a phenomenology of unexpectedness or oddness. This cognitively unsatisfactory situation will usually lead us to undertake further efforts to suppress conflicts of expectations by adjusting priors at an even higher level. There are different ways to do this. We could explain the error away by hypothesizing that the speaker has not expressed herself correctly or lacks linguistic capabilities. If that were the case, this high-level situational prior would lead to a suppression of the lower-level error signals, because we cannot rely on the sentence being correct. Or we might think that we have not understood well and ask for a clarifying statement.

To summarize, in the proposed PP view, a copredication is infelicitous if there is no prior such that the two different priors evoked by the nominal within the context of
each predicate are child nodes of that prior. A copredication is hence infelicitous if it fulfills the following simple *Infelicitous Copredication Condition (ICC)*:115

**Infelicitous Copredication Condition (ICC):**

The priors \([N_1]\) and \([N_2]\), denoted by the same word form, lack a common parent prior \([N]\).

We must add here that a copredication is also infelicitous, as already said in 4.1, when it fulfils FCC but expresses saliently a spatial or temporal separation of the entities represented by the different priors corresponding to the senses.

**10.4.3. Context-dependency of felicitousness**

So far, the story is relatively similar to the one offered by Ortega Andrés & Vicente. The authors suggest that in felicitous copredication two senses of the nominal are "activated" simultaneously as a coactivation package and hence made available to be combined with each predicate. We, on the other hand, have suggested that in felicitous copredications a more abstract prior makes available an information package in the form of an expectation hierarchy. This more abstract prior can be combined with both predicates.

How can we model the finding that some copredications sound odd while others sound better even if in both cases the nominal is associated with the *same coactivation package*? Ortega Andrés & Vicente do not provide a general solution to this difficulty due to asymmetries.116 More recently, Vicente (2021) expressed awareness that a more flexible approach than default coactivation of the senses is necessary.

Let us analyse again the example (3) from Section 10.2, which is problematic for the coactivation approach. It consists of a pair of statements that arguably involves a

115 Note that our account is silent concerning what exactly determines whether the required parent prior is available or not. Various authors have discussed patterns that favor felicitousness, like metaphysical relations between the involved senses (see Vicente, 2021). We are skeptical about the possibility of positing a comprehensive system of rules or generalizations that capture those patterns because of the context-sensitivity of felicitousness. This does not rule out that some generalizations can be identified that are descriptively useful.

116 However, for some specific cases of asymmetries, e.g., pairs of sentences involving a word that has a content and a container sense (like 'glass' or 'beer') and that appear in a different order, the authors point to *metaphysical dependency relations* as the reason of the difference in felicitousness (e.g., Ortega Andrés & Vicente, 2019; Ortega Andrés, 2020, especially Chapters 6.3 and 6.4).
single coactivation package ([NEWSPAPER_institution] and [NEWSPAPER_copy]) but produces different felicitousness intuitions.

(3a) The newspaper has been attacked by the opposition and publicly burned by demonstrators.

(3b) ?The newspaper has been attacked by the opposition and fell off the table.

Both statements arguably involve the same coactivation package as suggested by the felicity of (3a). Why then is (3b) clearly infelicitous? Positing coactivation packages of "senses" associated with the nominal cannot be the whole story.

We suggest that we need to take the background beliefs into account which set up a context in which the information packages are processed. The background beliefs have an influence on what part of the package is actually "activated". This is where the PP model of copredication can play out its strength in modelling context-sensitivity. The context-sensitivity of information packages is naturally available in the PP model. The priors associated with the information package of the nominal are embedded in the overall network of priors that constitutes the brain’s prediction model. We argue that higher-level priors outside the information package can serve as contextual priors modulating the pieces of information to be "activated" (or in PP-terms, have their probability increased). Contextual priors are part of the complete expectation network that needs to be brought into a global optimum for the task of sentence comprehension. When I am in a restaurant, a situation prior is represented in the brain that generates (mostly tacit) expectations of hearing sentences like 'Today we have fresh salmon' rather than 'Today we offer 10% off on tire changes'.

The reason why (3a) sounds at least better than (3b) may be that in (3a) we grasp a situational discourse context in the form of a contextual prior that represents a familiar prototypical scene of a specific way of protesting. Often, symbols of the object/person against whom the protest is directed are burned. We have all seen pictures and videos of flags, books, or photographs that are being burned by an upset crowd. This prototype of protest is a high-level conceptual pattern represented as a high-level prior in our model. Once we grasp that the statement is about a protest, a prior is activated, let us say [BURN_A_SYMBOL_AS_PROTEST]. And this prior generates an expectation that the mentioned newspaper is being burned. Under the influence of the [BURN_A_SYMBOL_AS_PROTEST] prior, we expect a newspaper-copy to
be burned in virtue of it being a symbol of the newspaper-institution. For (3b) we do not have such a prior. In (3a) the second part of the statement is easily conceptualized as part of the same event, while in (3b), the second part introduces an event that appears to be entirely unexpected.

Consider again example (5). Vicente (2021) has provided a version of it that is of the form we just discussed to illustrate that the coactivation account needs finetuning. It consists of two statements that invoke the same coactivation package but have different degrees of felicitousness:

(5a) The school caught fire when it was celebrating 4th of July.
(5b) ?The school caught fire when it was on excursion.

We have argued above why (5a) is felicitous (at least to some degree). Why is (5b) less felicitous according to the PP account of copredication? According to our model, in (5a), we have a strong situational prior (a party interrupted by a fire) that allows us to expect a building and a celebration (namely in or at the building). But in (5b) we do not have a single situational prior that makes us expect both predications. On the one hand, we have a fire in a building, while on the other we have a disconnected situation of a group of people (who happen to be people from the school) on some excursion at whatever location. There is no single situational prior that we can think of that makes us expect a school on fire and a group being on an excursion. After 'school caught fire' we can expect people, but people in the school, not people somewhere else who happen to be from the school. As we already suggested, the spatiotemporal relation plays plausibly a role here (see also Vicente 2021, p. 351). In the felicitous case, we have spatiotemporal coincidence and in the less felicitous case, we have spatial separation. This spatial separation makes very salient that the two senses represent different entities. In this case, the mind cannot represent both senses with the same more abstract prior.

Finally, for another illustration of how package-external priors can influence felicitousness judgments, consider so-called "predicate order effects" in copredication (e.g., Murphy, 2021a, 2021b; Ortega Andrés, 2020; Ortega Andrés & Vicente, 2019). Such order effects involve two statements with the same predications (and hence associated senses of the nominal) but in inverted orders. The statement expressed by one sentence turns out to be more felicitous than the
other. There is no space for a full discussion of the whole range of order effects, but we want to show how our account could model an especially interesting kind of order effect involving concrete and abstract senses of the same word.

Consider the following examples discussed by Murphy (2021a, p.16):¹¹⁷

(8a) The city has 500,000 inhabitants and outlawed smoking in bars last year.
(8b) The city outlawed smoking in bars last year and has 500,000 inhabitants.

(9a) The White House is being repainted and issued a statement concerning taxes.
(9b) The White House issued a statement concerning taxes and is being repainted.

Murphy (and Asher in case of (8)) take the b-statements to sound less acceptable than the a-statements. Again, the problem for the coactivation account is that the same body of information should be available in both cases. As before, we would like to emphasize that we can avoid the question of whether most people agree with the intuitions pointed out by Murphy and Asher. We recognize that not all readers have the same linguistic intuitions and, in particular, experts may have significantly unusual linguistic intuitions given their unique exposure to certain sentences. At least one of us does not hear any difference. Still, none of this is critical as we wish to offer a cognitive mechanism that can be used to model both felicitous and infelicitous intuitions even in the case of individual differences. An explanation of the differences in intuition that our model captures likely boils down to differences in experience, e.g., exposure to this or similar sentences.

Murphy (2021a,b) has suggested that we can explain this asymmetry by parsing preferences: we prefer to process first simpler and then more complex semantic structures.¹¹⁸ This suggestion is prima facie quite plausible (not least because it follows some common-sense principle of "incremental effort"). However, various observations suggest an alternative interpretation is worth considering. First, take the word ‘school’ and one of its abstract (institution) and concrete (building) senses.

¹¹⁷ Example (8) is taken from Asher (2011, p.63).

¹¹⁸ Murphy’s experimental data (2021a) indeed shows a simple-complex order bias for the cases that he has investigated. The question is whether the variable "complexity" plays a central role in a mechanistic cognitive-computational account of copredication. In other words, Murphy’s interpretation might be descriptively fully adequate for the experimental results, but it is still an open empirical question whether complexity is a causal-explanatory variable. For instance, an interesting question is why the parser should prefer a simple-complex ordering.
According to Murphy, the preferred ordering of senses should be SCHOOL-BUILDING and then SCHOOL-INSTITUTION. But it seems that the notion of SCHOOL-BUILDING is not simpler than SCHOOL-INSTITUTION. The former seems to require the more abstract institutional sense of 'school'. We are dealing with a school building, not any arbitrary building. We suggest that what might be instantiated in cases like (5) is the concept SCHOOL-BUILDING not BUILDING simpliciter.\textsuperscript{119} In any case, it is far from obvious whether a linear order can be established at all in all cases; the here proposed approach has the advantage that it does not need this assumption.

Secondly, it has been suggested that children can quickly grasp more abstract features and generalizations before more specific exemplars are presented (e.g., Kemp, Perfors & Tenenbaum, 2007; Keil, 2021) and that perception of complex objects/scenes often involves first grasping the overall gist before concrete details (e.g., Barrett & Bar, 2009; see also Fillmore, 1975). This at least suggests that conceptual processing, including the processing of concepts within copredication statements, might not necessarily follow a simple-complex order preference (see also Rappe, 2019, for a PP-based account of sentence processing in this sense). Now, Murphy explains order effects not by appeal to abstraction but semantic complexity.\textsuperscript{120} However, abstraction and complexity are plausibly often correlated; abstract concepts seem to be more complex in many cases (e.g., SCHOOL is more abstract but also more complex—in the here relevant sense—than BUILDING). In any case, our aim here is not to argue against Murphy's account, but merely to illustrate how the proposed PP model can support alternative hypotheses about interesting copredication phenomena.

The question that interests us is how to model the cognitive underpinnings of linguistic intuitions about copredications. Our model offers a neat implementation of the intuitions about (8) and (9). (8b) simply sounds odd because when reading 'The city outlawed' our prediction model "jumps" from the parent node [CITY]—which is connected to child nodes like [CITY\_government], [CITY\_population], [CITY\_geography], etc.—to the child node [CITY\_government]. This is plausible given

\textsuperscript{119} This discussion benefited from a conversation with Agustin Vicente.

\textsuperscript{120} Semantic complexity is defined by Murphy via the number of cognitive modules involved in the conceptual processing. A module is here understood as a specific domain of reasoning, like number sense, folk psychology, social reasoning, geometric reasoning, etc. (see 2021a, pp.54-56).
that it is usually the government of the city that makes the laws, hence [CITY\_government] has a high conditional probability given [CITY] and the context of lawmaking. From the sub-node [CITY\_government], the predicate 'has 500,000 inhabitants' is then less expected. Hence, the statement sounds odd.

This is not the case for (8a) given that now we are talking about the city in its more abstract and inclusive sense ([CITY]) having 500,000 inhabitants. We now expect that more will be said about the city with all its different aspects rather than just one specific aspect of the city as in (8b) (see Figure 10.2).

Note that we can expect the predicate 'has 500,000 inhabitants' directly from [CITY], but not from [CITY\_population]. In fact, 'The city marched on the Capitol building and has 500,000 inhabitants' sounds odd. In this sentence 'city' is used in the [CITY\_population] sense with 'marched' and therefore, having a certain number of inhabitants is expected based not on [CITY\_population], but on the more abstract prior [CITY] (see Figure 10.2 A).

![Expectation hierarchies for the 'city' (A) and 'White House' (B) examples from statements (8a,b) and (9a,b).](image)

**Figure 10.2:** Expectation hierarchies for the 'city' (A) and 'White House' (B) examples from statements (8a,b) and (9a,b).

The case in (9) is slightly different, given that 'White House' is an intrinsically metonymic expression as opposed to 'city'. In statement (9a), 'White House' with 'is being repainted' evokes the building sense [WHITE-HOUSE\_building]. The second predication 'issued a statement' is not expected from [WHITE-HOUSE\_building];
nevertheless (9a) sounds acceptable. This, we suggest, is due to the availability of a building-for-institution metonymy that serves as a higher-level prior. A building-for-institution metonymy is nothing more than an expectation relation that allows us to expect an institution sense given a building sense. In (9b), first the [WHITE-HOUSE\_government] sense is evoked and then the unexpected [WHITE-HOUSE\_building]. It is unexpected because there is (as a matter of linguistic fact) no institution-for-building metonymy that can serve as a parent prior (see Figure 10.2).

However, these are hypotheses and much more needs to be said about order effects elsewhere. We also would like to stress again that the explanation of why we have the expectations we have is only a secondary aim of this chapter. The primary aim is to provide a model that allows for formulating such hypotheses. In many cases, there might be no "explanation" at all, and the generation of different expectations is simply a consequence of how our cognitive system has evolved to optimally adjust itself to the environment by prediction error minimization. We do think however that there is room for interesting generalizations about expectation relations (for instance, metonymy research can be interpreted as contributing to this enterprise). So, there are interesting—but (possibly many) different—projects that can provide a taxonomic inventory of all of the possible expectation relations that might do "explanatory" work in a different sense as we have aimed to provide here.121

10.5. Conclusion

We have put forward a cognitive-computational model for felicitousness judgements of copredication statements. The account further develops Ortega Andrés & Vicente's psychological "coactivation package" approach. In their account, the senses of the nominal are available in the coactivation packages. In this way, both senses are available for the predications and can be selectively applied. However, this account cannot accommodate cases where the same coactivation package is involved but the felicitousness intuitions are different. What is missing is factoring context-sensitivity into an account of the felicitousness of copredications.

121 Other examples of relevant types of expectation relations are coherence relations (e.g., Murphy, 2021a, 2021b) and metaphysical realization relations (e.g., Ortega Andrés & Vicente, 2019; Vicente, 2021).
We use the framework of predictive processing (PP) to provide a model of the structure and context-sensitive processing of information packages. In PP, cognition is continual prediction making in a hierarchically organized model of the world. The key mechanism for inference and model improvement is prediction error minimization. We have characterized the information packages associated with words as expectation networks consisting of priors (=expectations) at different levels of abstraction. The representations at higher levels serve as predictions or expectations for the representations at lower levels. The information packages are embedded in the huge hierarchical expectation network that constitutes the brain’s prediction model.

The core idea is that felicitous copredication is possible because we can make available a single higher-level prior that is compatible with both predications. This higher-level prior is a parent prior of the two different priors denoted by the nominal. Infelicitous copredications lack a unifying parent prior. In some cases of apparently infelicitous copredications, higher-level situational or context priors can contribute to making the copredication felicitous. We have also argued that metonymic constructions can play the role of higher-level priors, which allows us to explain some of the effects of the order of the predicates on felicitousness judgments.

We would like to close by emphasizing that our focus has been predominantly theoretical and based on explaining linguistic material. More work needs to be done to provide further empirical support for the model presented here, especially the existence of the expectation hierarchies corresponding to priors. Our proposal is empirically testable in principle as we build on a specific PP model of neural implementation. A possible way to proceed, in principle, would be to localize expectation nodes, correlate them with specific interpretable expectations and intervene on them (e.g., by TMS-like techniques) to study the impact on felicitousness judgments. But this would, of course, require further progress in the study of the way the neurons are connected in the brain (the connectome) and very precise, localized neural monitoring and intervention techniques.

Another approach that could provide further empirical confirmation would be behavioural studies, for instance, cross-personal and cross-cultural studies of the variability of felicitousness judgements. As an example that is related to copredications and that involves metonymies, one could try to identify non-universal
metonymies and then compare felicitousness judgments of copredication statements involving metonymies across different individuals or language communities. Also, cognitive developmental data could be used to test our view. For instance, one could study how the felicitousness judgements of an individual might change after the acquisition of certain metonymic constructions.

Acknowledgements

We would like to thank François Recanati, who introduced GL to the topic of copredication. We are very grateful to Marina Ortega Andrés, Martin Pickering, Agustín Vicente, and various anonymous reviewers for their valuable comments on earlier drafts, as well as the audiences of the 47th Annual Meeting of the Society for Philosophy and Psychology (SPP) and the 28th Conference of the European Society for Philosophy and Psychology (ESPP), to whom we have presented earlier versions of this chapter.
Conclusion of Part 4

I hope to have convinced the reader that PP has interesting things to say about outstanding linguistic and semantic puzzles considering them to be phenomena of higher cognitive processing.

I have treated the Liar Paradox not as a logical formal problem, but as a problem related to a failure in cognitive processing. Based on a dual processing model I have suggested that the paradox arises as a failure of synchronized predictions in the world sub-model and in the linguistic sub-model. The world sub-model contains the world knowledge and the linguistic sub-model, the linguistic knowledge. Both must work in a synchronized way. The Liar Paradox appears, so I suggest, because we "sloppily" suppress subtle conceptual constraints when processing the situation evoked by the liar sentences. The liar situation cannot arise "in reality" (represented in the world sub-model). Therefore, there is no paradox, only a failure in mental processing. The reason why the Liar manifests itself as a paradox is, therefore, not that classical logic is inadequate, or our concept of truth is defective but that we neglect conceptual constraints, which need to be carefully considered when forming sentences. As pointed out in Chapter 7, thought is not unrestrictedly compositional, i.e., not all grammatically acceptable sentences are meaningful. As the constraint in the case of the Liar is very subtle, it is easy to ignore it.

Regarding metaphor, I have argued that PP with the here proposed model of concepts can cognitive-computationally underpin one of the two main rival views, namely the Category Inclusion View (and specifically the Interactive Property Attribution Model) better than the other (the Implicit Comparison View). This is relevant because whatever support accrues for PP, it increases the plausibility of the PP-supported view of metaphor—and vice versa. Furthermore, the PP vision adds new elements to metaphor research by providing commitments on the level of neural architecture, that can at least in principle been empirically tested. This might allow for deriving new testable hypothesis that could tip the balance towards one or the other account of metaphor processing. It also turns out that if a hybrid account for metaphor were needed—such a view is a current trend in metaphor research—the
PP account can already provide the necessary features that the complementary account (Implicit Comparison View) would add to the Category Inclusion View.

With respect to the phenomenon of copredication, we have suggested that a PP account can improve the existing "information package" account by covering context effects. We have also shown tentatively that so-called "order effects", i.e., effects where the acceptability of a copredication sentence depends on the orders of the predicates, can be modelled by the PP approach.

In sum, PP with its specific concepts and mechanisms provides new ways of thinking about those conceptual-linguistic phenomena by looking at them from a cognitive-computational perspective. By grounding theories and approaches to those problems in an existing and promising cognitive paradigm, which makes commitments on all three level of Marr, one might gain new angles to progress towards more robust solutions. Such a grounding might also help in particular in developing new testable hypotheses (even on the level of neural processing) that might contribute to dissolving the characteristic empirical deadlocks in long-standing debates around those phenomena.
Part 5 - Some challenges, further work, and concluding remarks

Chapter 11. Addressing Williams' objections

Williams (e.g., 2019, 2020) has put forward various interesting objections on theoretical grounds that cast doubt on the suitability of PP as a framework for higher cognition. In this chapter, I try to synthesize his concerns and provide some tentative responses. I conclude that Williams quite correctly points out that some of the assumptions that some PP theorists make are difficult to maintain. Also, Williams points to issues with PP commitments that require more careful formulation or qualifications. However, his critique should be seen in a constructive light. In fact, I have already tried to circumvent some of his concerns in this dissertation.

I will focus on three blocks of objections: the objection from off-line processing, the objection from expressive power, and the objection against the unified hierarchical structure of the PP model (which in turn consists of three more specific objections).

11.1. The objection from off-line processing

A first issue mentioned by Williams (2020) concerns the off-line nature of a lot of higher cognition. Here, "off-line" cognition means simply that the brain can think without a constant and immediate intake of sensory data; in other words, it can decouple itself from its environment and cognize autonomously. The objection can then be articulated as the question: What, for instance, has philosophizing or mathematical research to do with the ongoing prediction error minimization of incoming sensory input?

But PP has the resources to provide a plausible response to this prima facie objection. Namely, the model can be run in a mode where bottom-up signals from the sensorimotor periphery obtain low precision estimates and are suppressed. The usefulness of this processing mode can be appreciated if we consider insights from reinforcement learning (see, e.g., Sutton & Barto, 2018). The PP model might be optimized for long-term error minimization if it is run off-line from time to time, i.e., the brain shuts off from external sensory exposure and keeps improving its model by exploring scenarios and consequences in simulation mode and makes it more coherent, also on the level of lexicalized concepts. For instance, philosophizing could
be seen as trying to clean up conceptual inconsistencies and obtain a better big picture model—a worldview or outlook. One such type of offline clean-up (reduction of model complexity), by the way, is supposed to happen in the hippocampus during sleep (e.g., Born & Wilhelm, 2012). Furthermore, it has been emphasized that counterfactual modes of cognition—in the form of the (off-line) evaluation of alternative courses of action, for instance—are crucial for more sophisticated cognitive agents according to PP (e.g., Corcoran et al., 2020; Parr et al., 2022). It is obvious that learning via counterfactual exploration is in many situations less risky and costly than learning by doing.

Certain higher cognitive human undertakings not focused on online processing of immediate sensory input, like theoretical research, might also be seen as contributing to long-term prediction error minimization on a collective and long-term scale. After all, the development of science has dramatically augmented the predictive capability of mankind. Science consists of publicly shared models of certain aspects of the world that can be exploited both to predict phenomena (i.e., to carry out perceptual inference)—e.g., to make weather forecasts—but also to change aspects of the world conveniently (i.e., to carry out active inference)—e.g., to heal ill people. Here, PP might be given a perspective of an improvement of collective long-term prediction. This is also consistent with the view of the importance of language in the prediction economy as alluded to before: language is a vehicle for publicly sharing and improving on mental models and predictions that allows for the coordination of predictive agents for improved collective prediction making.

11.2. The objection from expressive power

Furthermore, Williams (2020) has reservations about the expressive power of the PP model. The concern is that the commitment of PP theorists to so-called Probabilistic Graphical Models (PGMs) prevents PP from being "richly compositional". The argument is that the network structure where nodes correspond to propositional facts limits the combinatorial possibilities. With PGMs we seem to lack the possibility of representing more fine-grained representations like objects and relations. Each node is merely a proposition, so we only obtain the expressivity of propositional logic, which indeed is not enough.
Williams makes a valid point here and I agree with him. Simple PGMs do not seem a suitable way to formalize the PP model because they are not fine-grained and flexible enough. Also, they do not contemplate the addition or deletion of nodes,\textsuperscript{122} so neurally they seem not very realistic. Modelling cognition with the help of PGMs might, of course, be useful for some idealized aspects of PP, but they cannot serve as a full model for PP-based higher cognition. How do we get the high degree of flexibility in combining nominals and predicates, for instance? Indeed it seems that we need more powerful representational structures.

One possible avenue, if one insists on mathematically formalizing the PP model, is to generalize Probabilistic Graphical Models towards others with more expressive power. For instance, Getoor et al. (2001), describe "Relational Probabilistic Models" which generalize Bayesian probabilistic graphical models achieving more expressive power; they allow, e.g., for a varying number of entities and relations. This possibility seems important for any suitable account of language and concepts, as we certainly can constantly create new concepts and new relations.

However, I follow a path that is different to trying to come up with the most appropriate mathematical formalization of the sort that PGMs instantiate. While I consider that we deal literally with a network of nodes, I do not endorse the formalization via simple or relational PGMs. In fact, I am sceptical about the very idea that the brain's model can be formalized in the way PGMs, in whatever improved version, suggest. In my rejection of (strict) formalization, as I have pointed out in the discussion of Construction Grammar, I follow Langacker and Goldberg. The reason is that this way of modelling implicitly follows a sort of LOTH paradigm based on discrete symbols that very transparently combine in a fully compositional way. Formalizing in this sense is just putting something in a LOT-like formal language. The very point of this thesis is that we need to replace the LOTH paradigm which often tacitly influences our intuitions about the nature of the compositionality of thought and the workings in higher cognition. This is difficult, because we are so strongly biased towards LOTH and generative (rules + symbols) thinking. This has to do, so I suggested, with the very embodied nature of thought. We cannot help but conceive of thinking as if we manipulate discrete things in a LOT-like manner.

\textsuperscript{122} But see Smith et al. (2020) for a possible approach.
Also implicit in my proposal is the suggestion that nodes need to be interpreted differently. Nodes can also represent, for instance objects and relations. A node is simply a pattern (extracted from lower-level regularities), and objects and relations are patterns. Objects and relations can also be represented as facts. Representing a car is representing the fact that this is a car. Representing a relation X is representing "This is a relation X".

Relatedly, I suggest that first order logic might not be the right benchmark at all for expressive power. For instance, we have seen that Construction Grammar exhibits only partial compositionality, because constructions always contain genuine content that is not derivable from its components (e.g., Langacker, 2008, p.42). PP theorists could simply bite the bullet and endorse that language and thought are not fully compositional in the sense of first order logic. Also, there might be a more radical response. One could hold that expressive power expressed in terms of reference to formal logic (and hence in terms that epitomizes LOTH) simply misses the point that modal theorists make. We need other ways to quantify expressive power, than in terms of discrete formal symbol manipulation. By appealing to abstraction/compression/convolution, we get a completely different paradigm.

Admittedly, this way of thinking is not as entrenched and not as perspicuous as the LOTH-type of formalized thinking, for reasons I have already mentioned.

It would also seem that Williams would prove too much if he were right. The reason is that we can obviously perceive in a richly compositional way. Perception is productive and systematic and includes objects and relations. Our brain can predict all sorts of perceptual scenes, in fact an unbounded number of perceptual scenes. Furthermore, perception is systematic: we can perceive that the chair is right of the table, and we can see that the table is right of the chair. We can perceive that Bob gives Mary a book and we can perceive that Mary gives Bob a book. But if compositionality is a problem for off-line simulation, it must also be a problem for online perception. So, Williams would have an argument to the effect that we cannot explain perception by PP. But then the whole PP paradigm would be undermined, even in the domain of lower-level perception, where PP is less controversial.

In sum, Williams' concern from expressive power is a very interesting challenge and I consider his arguments as valid, given the assumption of simple HBMs made by PP theorists. But rather than undermining the whole PP research programme, Williams
makes a valuable contribution to PP's further development. He points to the necessity of finding a better way to describe and quantify expressive power that provides an alternative to the formal LOTH-based paradigm. I have tried in Chapter 6 to provide such a possible alternative in the form of the Construction Grammar based paradigm of compositionality. Still, my proposal is speculative and conceptual, and it might be interesting to keep exploring other options. For example, an interesting question is what role the temporal order of representations might have in the way the brain realizes compositionality. Namely, compositionality can be implemented as varying temporal order. To grasp that Mary gives Peter a book implies grasping the concepts involved in a certain order. The same concepts but in a different order of activation provide a differently composed situation. In this way, we get the necessary systematicity. If we have nodes that represent patterns that include the temporal order of the activation of other nodes (for instance, recurrent networks can do this work), then it seems we can get rich compositionality. However, this would require a very different formal apparatus then simple HBMs.

Before closing on the issue of compositionality, let me point to a puzzle about expressive power. How can we conceptualize, grasp, and analyse expressive power that is vastly superior to the expressive power of our brains? For instance, how can we conceptualize, grasp, and analyse quantum computers and their expressive power? Solving this puzzle might also point to a solution for the genuine LOTH compositionality of some system that itself does not have this expressive power. We might not have the expressive power of first order logic but can still simulate/represent LOTH-like compositionality, in the same way that we can represent and simulate quantum expressive power. Here, we have a distinction like the use/mention distinction. It seems we can run simulations of higher expressive power (quantum computer) on a system with less expressive power.

11.3. Three objections regarding the unified hierarchical model

Three further challenges are related to another core commitment of PP: the hypothesis of a unified hierarchical model. Those challenges build on the observation that the PP view seems to dissolve a functional modular view and blur the perception/cognition dichotomy given the continuous and homogenous nature of the model topology. I will now deal with each of the three challenges in turn.
11.3.1. Representational reach

Firstly, Williams (2019) doubts that a single hierarchical model has the necessary "representational reach" to explain thought. He asks what property of representations correlate with the level in the hierarchy. The implicit answer provided by the model proposed in this dissertation is that the higher a phenomenon is represented in the hierarchy, the more abstract it is. Williams anticipates this answer and puts forward the following concern:

Whilst I think that this response is correct as far as it goes, there is a significant risk that it simply draws attention to the intuition that fully conceptual representation is of a fundamentally different kind to purely perceptual representation. Of course, this might very well be correct, but it is the very fact that needs to be explained. The worry here is that vague appeal to a unified inferential hierarchy simply masks a tacit reliance on an autonomous domain of conceptual representation. (Williams, 2019, p.13)

My reply is the following. Here, I have endorsed a very inclusive notion of "conceptual representation", precisely because there is no principled qualitative difference between the roles of the prediction units on each level of the hierarchical model. The units differ only in the degree of abstraction of the information they represent (apart from factors like the conscious access, lexicalization, etc.). In this view there is no dichotomy "fully conceptual representation" versus "fully perceptual representation". We might still have an intuition of such a perception/cognition dichotomy, and with this I agree. However, I suspect that it can be fleshed out via the distinctions between online versus offline cognition and conscious versus subconscious cognition, i.e., other factors than the level of abstraction are responsible for it. PP is even more radical because action is represented much like perceptual or conceptual representations. Factors like conscious access, however, would not supply a clear dividing line for the perception/cognition distinction if we accept the existence of at least some subconscious online conceptualizations (e.g., tacit grammar knowledge has this property and looks more conceptual than perceptual).

Obviously, here, I have not provided a full solution to Williams' concern about representational reach. There is, for instance, a need to explain why we do have conscious access to some prediction units and not to others. A complicating factor is
that conscious access might itself be graded, i.e., there is a phenomenology of
differently strong degrees of conscious access. Certain intuitions, like linguistic
intuitions, for instance, might correspond to access that is in between fully conscious
and fully subconscious. Often, we cannot articulate intuitions well, but we certainly
"feel" them to a certain degree. But cognitive phenomenology is a complex field that
exceeds the scope of this dissertation (see also Chapter 12). My response must,
therefore, remain tentative.

11.3.2. Modularity

A second objection by Williams (2019) is that a single unified hierarchy, as posited
by PP, is not adequate as a model of the mind. Rather, the mental model must be
modular because we have very differentiated domain specific intuitive theories, like
folk psychology and folk physics. This requires multiple discrete structures.

However, modularity or structural variety can naturally (and somewhat paradoxically)
be accommodated within a unified generative model of the PP sort. Hierarchical
Bayesian Models can infer, for instance, very different data structures in which
certain domains are represented, like taxonomic structures, or tree structures (see,
e.g., Kemp & Tenenbaum, 2009, for an illustrative example in the context of property
induction). The idea is the following. How could we explain that we have two very
different intuitive theories: folk psychology and folk physics, which appear to require
different "modules"? Given that the model is hierarchical, we can posit concept-units
that correspond to each domain, and which have been inferred as the latent cause of
what is represented as the content of that theory. A domain is hence represented as
a very high-level schematic concept unit that serves as the representation of a
context. Depending on the context prior one part or another of the same large,
interconnected model is switched on, and others are switched off. This means that
some parts of the network have a high probability conditional on that context prior.
Other parts of the network have a low probability and are switched off in the context
represented by the prior. This then looks very modular, but we get this modularity out
of a homogenous, continuous, hierarchical model.

For example, imagine you are in a situation that requires folk physical cognition. In
this case, the organism predicts "I am in a folk physics situation", which implies the
activation of the FOLK-PHYSIC concept unit serving as a context prior that gives
preferred access to representations in this subnetwork. All concept units connected to it have a high conditional probability (across the different levels) and are then easily accessible. Others like FOLK-PSYCHOLOGY are not activated. In other words, concepts most closely associated with it have a low probability of being accessed, i.e., the folk psychology module is switched off. Parts of the same single hierarchy can be switched on and off, and this explains the seemingly modular nature.

A model of higher cognition based on a strong form of encapsulation is implausible anyway for the simple reason that most domains are more or less strongly interconnected. Certain concepts are used across domains. For instance, an exemplar of the concept HUMAN is after all also a physical object and can appear both in folk physics and folk psychology (though a human might not be a prototypical object of folk physics, of course). Furthermore, there is one phenomenon that speaks strongly against strong forms of domain encapsulation, namely metaphor. In a metaphor we leverage the fact that we can easily relate very different domains. For instance, we can speak of persons in terms of celestial bodies (“Juliet is the sun”). Therefore, given the resources of the PP model, the existence of different intuitive theories or domains does not seem a decisive argument against a large single hierarchical model.

In a more technical vein, recently Friston & Buzsáki (2016) have shown that modularity can arise in a hierarchical Bayesian model, namely via “factorization”, which is part of approximate Bayesian inference. Factorization is a process that simplifies the representation of a joint probability distribution over variables, by approximating it by the product of their marginal probabilities ("mean field approximation"). This implies a description in terms of a structure that looks as if it is composed of independent variables which can be seen to correspond to "modules". This corresponds to probabilistic independence: for instance, seeing what an object is does generally not allow for inferring where it is located; the what and where are independent factors. It is more efficient to encode separately where an object is and what it is, rather than all combinations of objects and locations in a complex joint probability distribution. (However, this is merely an approximation, for course: for instance, my computer is normally on my desk, so I can infer with a certain

\[^{123}\text{They consider a factorisation within the mental model between representations of temporal succession (when) and representations of content (what). But this idea can be generalized.}\]
probability the location of my computer from the fact it is my computer). Factorization implies a certain way of approximately "carving out nature":

[...] evolutionary (Bayesian belief) updates have shaped the brain into an efficient (minimum free energy) mean field approximation that we know and study as functional segregation. (Friston & Buzsáki, 2016, p.502)

So, to posit a homogenous hierarchical model is not in contradiction to a certain level of modularity. The hierarchical model is not "homogenous" in its actual functional occurrence. It is homogenous in its basic organizing principles, and that is consistent with the appearance of a certain degree of (functional and physiological) modularity or segregation.

11.3.3. Lack of a coherent account of what is tracked by the hierarchy

A third concern of Williams is that it is, according to him, not clear what the PP model hierarchy tracks. According to Vance (2015), the hierarchy of the PP model "tracks" the distance from the sensory surface and representations of phenomena at increasing spacio-temporal scales. That seems to mirror what other theorists have said. Williams asks:

"Do my thoughts about electrons activate representations at a different position in "the hierarchy" to my thoughts about Paris, the English football team’s defensive strategy, or the hypothesized block universe? If so, by what principle?"

I suspect that this question is based on a sort of use-mention confusion, and it seems to me that Williams is attacking a strawman here. It is certainly completely implausible that what is tracked in the PP model hierarchy is the size of things: "electrons" at low levels and the "universe" at a high level in the hierarchy, for instance.

How the notion of increasing spatiotemporal scale should be interpreted most charitably can be illustrated best with the example of visual perception. The representations on pixel-level are extremely unstable as they suffer constant, fast variations, and each pixel maps a small area - hence on pixel-level the representations are at a spatiotemporal smaller scale. In other words, the receptive field on pixel-level is small and intensely varying. As many possible pixel-level
patterns map onto one edge representations, the potential changes on the edge-level slow down because this level extracts more stable higher-level patterns in the form of a many-to-one mapping with lower-level pattern. The key here is to understand that an abstraction process is taking place along the hierarchy. A higher-level representation has a many-to-one mapping to the next lower level, or in other words, the receptive field increases in higher levels. Finally, face representations—which happen further up in the processing hierarchy—are even more stable because they are more abstract representations in the sense just mentioned. The face of Sally, for instance, is recognized despite many changes of the specific shapes projected on my retina due to varying viewpoints, light conditions, occlusions, and so on. So, the important point is that higher up in the hierarchy more invariants are extracted.

The concept ELECTRON and the concepts of, say, ELEPHANT, PARIS or UNIVERSE, are on (roughly) the same level of abstraction in the model I have proposed. All those are concepts of objects and are lexicalized, consciously accessible representations at higher levels of the hierarchy. The fact that you cannot perceive electrons with the naked eye does not make them qualitatively different concepts from elephants. Electrons are things that can be experienced and manipulated, though more indirectly than elephants, of course. It is not because some exemplar of a concept is very small or big that this concept is represented at a lower or higher level in the hierarchy. Thinking about electrons does not happen via representational vehicles that fluctuate fast and have small receptive fields. In other words, we do not think small and fast about electrons, though we can think that they are small and fast objects.

Now, we could think of oriented edges represented in the brain and have a concept of them. However, the representations by which we think of such edge representations in the brain’s visual pathway are then not the oriented edges represented in lower levels that our visual processing hierarchy uses to build, for instance, a face. In fact, there are many such edge form patterns neurally represented that are not humanly interpretable; they are blurry and only resemble edges. Rather those lower-level edge form representations, when we think about them, are meta-represented as exemplars (things in the world), not the vehicles of
our thoughts about them. This distinction is important and to ignore it is to commit a use–mention fallacy.
Chapter 12. Further challenges and the way forward

In this last section, I address some objections that affect the PP paradigm not only as applied to higher cognition but also more generally. First, I address some concerns of Litwin & Miłkowski (2020) and, briefly, two other important objections that might threaten fundamental posits of PP. I then close by pointing to further desirable work regarding higher cognition within the PP framework.

12.1. Litwin & Miłkowski’s critique of PP

Litwin & Miłkowski (2020) strongly criticize current research and theorizing under the PP paradigm on various grounds. The authors take issue mainly with the unificatory ambitions of PP. They think that the theoretical and conceptual basis is not yet solid enough for a unified theory. But at the same time there is a premature proliferation of all sorts of PP accounts without strong empirical evidence:

As a unifying theory, PP fails to deliver general, simple, homogeneous, and systematic explanations. [...] PP-based models are seldom empirically validated, and they are frequently offered as mere just-so stories. (p.1)

With regards to the unificatory ambition of PP, I have tried to make clear throughout this thesis, that I do not assume PP to be a theory, but a paradigm or research program. In fact, L&M also consider PP to be (“at best”) a "computational framework":

PP is usually assumed to be a unifying theory, but remains a computational framework at best. (p.5)

What it actually might become, given the current diversity of approaches within the PP community, is rather a research program or tradition in the sense of Laudan (1977), encompassing multiple alternatives and mutually exclusive theories. (p.23)

124 In a very recent paper, which I do not discuss here, Miłkowski & Litwin (2022) endorse a similar view as in 2020. They consider PP as a “universal modeling tool with an unrestricted number of degrees of freedom” and stress that PP “should not be understood as a unifying theoretical perspective, but as a computational framework, possibly informing further theory development in cognitive science” (p. 461).
I have tried to be careful not to position PP as a "unificatory theory of cognition" throughout. I have also avoided at this stage—despite its great attractions—specifically endorsing the Free Energy Principle. It seems obvious to me that no theory "can explain everything" in cognition, especially everything on all levels of analysis and description. I cannot even imagine how one could start to model, for instance, the phenomenon of metaphor with Friston's Free Energy formalism.

The authors further observe with concern a proliferation of accounts under the PP banner without immediate and direct empirical support. This concern deserves further discussion, as it could be brought to bear on my approaches to semantic paradox, copredication and metaphor as well. The authors are especially sceptical about the applicability of PP to thought and its pathologies, by reference to Williams' (2019, 2020) critique, which I have discussed in the previous chapter.

[...] the issues with the processing hierarchy we indicated above (see also Williams, 2018a, 2018b, 2018c) could be circumvented by limiting the intended scope of the theory to perception and action, and by excluding cognitive or psychopathological phenomena which remain problematic for PP. (pp.20–21)

Those applications are, so they point out, typically verbal and conceptual, and often (discriminative) empirical evidence is not provided.

[...] the under-determination of fundamentals results in a "horizontal" trajectory of PP development: It expands "to the side," being extrapolated in the form of many re-descriptions to particular psychological and cognitive phenomena prematurely. (p.3)

L&M think that there are more urgent and more fundamental issues that PP need to fix first.

Thus, we urge the defenders of PP to focus on its critical problems instead of offering mere re-descriptions of known phenomena, and to validate their models against possible alternative explanations that stem from different theoretical assumptions. Otherwise, PP will ultimately fail as a unified theory of cognition. (p.1)

We argue that PP currently fails to stand as such a unifying theory, and that its failure is deeply rooted in its current theoretical structure. The interpretation of its mathematical underpinnings turns out to be ambiguous, and the PP hierarchy seems implausible as a general blueprint of a cognitive architecture. (p.3)
The problem with this horizontal expansion seems to be twofold according to L&M. Firstly, many PP accounts are merely verbal redescriptions, and no suitable empirical support is provided. Secondly concepts are not used in a homogenous way across authors, or the use is otherwise defective or incoherent. Let me respond to both issues in turn.

Regarding the first issue, I tend to agree that an account spelled out within the PP framework should be more than a mere redescription in PP vocabulary. But I think that redescriptions are also legitimate and useful work. Note that it is quite remarkable that such a broad scope of phenomena in so many domains can be redescribed in terms of the PP framework. While that does not verify the framework, it does convey some (ironically Bayesian) confirmation on it.

Now the point might be that it is not the redescription itself or that it is a "mere" redescription which is problematic, but the fact that the application of the concepts in the redescription might be otherwise defective. In my analysis, there seem to be at least four ways in which such an application could be defective. Note that those defects could also affect redescriptions that are not "merely" redescriptions.

One such way is that the concepts are interpreted overly flexibly leading to a situation which characterizes non-falsifiable pseudo-sciences like psychoanalysis or astrology. If descriptions can be concocted for incompatible hypotheses, the framework and its concepts can accommodate everything and therefore is not useful at all. This seems not to be a threat for my accounts of cognitive phenomena suggested here (for concepts, metaphor and copredication), because, as I have argued, the PP paradigm supports certain existing approaches more than others, i.e., it is not flexible enough to "accommodate everything".

The second way in which the application of PP could be defective is that its concepts are used in different ways by different authors. This might be damaging only if PP is understood as a monolithic framework, which I am not endorsing. It seems to me that it is not a problem that different interpretations compete, or even that they co-

---

125 That, however, should also be qualified slightly. It might simply be the case that the constraints of the paradigm are not fine-grained enough to distinguish hypotheses A and B regarding the parameter in which they are incompatible. That is bad luck indeed, and not helpful. But there might be many other competing hypotheses that are incompatible on dimensions that the paradigm is able to discriminate between.
exist—as long as this is made clear, and we avoid equivocations across interpretations.

A third way in which the use of the concepts of the PP paradigm for redescriptions could be defective is that the use is in some form internally incoherent. To see the point, we need to discuss specific examples of such defects. The authors point out specific examples, namely equivocating between a computational and a psychological use of terms. This affects central concepts like "precision" (salience, confidence) or "prediction". According to L&M, precision is a computational notion, and has nothing to do with a subjective feeling of confidence. This conflation of homonymous terms is unwarranted as it could lead to a fallacy of equivocation. The computational notion of precision is semantically not equivalent to the psychological notion of confidence, for instance. We would need a "further explication of how the terms refer to each other". Furthermore, precision is identified with a wide range of psychological phenomena: subjective feelings of confidence, trust, salience, or sharpened attention. This "cannot result in informative explanation" (p.7).

L&M are right in demanding a "further explanation". We have here a substantial assumption of identity between computational notions and psychological notions that is often tacitly taken for granted. Notions like "prediction" are computationally used, but they also have "homonymic counterparts in the subjective domain". I do think there is some plausibility to the identification of those computational terms and psychological states. But I also agree that there is space for more clarity about how the computational PP terms relate to all of the traditional (folk-)psychological concepts. As an example, I have appealed to prediction errors as the causes of the feeling of "oddness" or "infelicity" of semantically or syntactically defective sentences. How exactly is this phenomenology working? What sort of errors are consciously accessible and perceived as linguistic intuitions? Those questions
deserve further work. The PP framework has it that representations are in the form of predictions in a technical sense and those are beliefs. Beliefs also include sub-personal beliefs, so there is no principled distinction between personal and sub-personal predictions. This claim is revisionary, but not implausible, and indeed, one should try to work out in more detail how all of the folk-psychological notions can be accounted for within the PP framework. In 12.2.1 will get back to a related problem, pointed out by Sprevak (2021c), that specifically affects the distinction between the folk-psychological notions of belief and desire.

A fourth defect in the application of PP to higher cognition might be that the concepts used are too schematic, leading to very high-level verbal and non-formalized descriptions that cannot be immediately translated into specific hypotheses for empirical verification. This complaint, however, might be unfair because one might not pursue a project in the sense of a scientist carrying out "normal science" (Kuhn, 1962), with the main concern being immediate empirical verification of specific phenomena. Rather the project might be located on a more conceptual, theoretical, and abstract level, e.g., the level of a paradigm. Science progresses on many levels of description and is full of examples of such "higher-level projects" that have been very fruitful and influential without being questioned because they were not accompanied immediately by decisive empirical evidence. For instance, Fodor's LOT (see Chapter 6) and Sperber and Wilson's Relevance Theory (see Chapter 9) are examples of cognitive models couched on a high level of description where the concern for immediate and detailed empirical verification was not the priority. Nevertheless, those have turned out to be very successful and influential frameworks.

What is important in those projects is to explore whether the paradigm is coherent, fruitful, and elegant, and provides new perspective or links with other disciplines. Even though immediate and decisive empirical verification is not my main concern, I have been trying to put forward strongly empirically informed approaches and to point out what further empirical work could look like. Furthermore, note that in my account of concepts (see Chapter 5) I have made a strong emphasis that an account of concepts should be linked to the huge body of accumulated empirical and theoretical concept research.
Even if evidence is provided in some PP account, so L&M further object, often the evidence is merely "compatible" with the account. Compatibility is not enough for confirmation, according to them, following Coltheart (2013). Empirical data should discriminate between alternative accounts. I agree with the authors that, ideally, theories should be supported by discriminatory evidence against the foil of alternatives. But the appeal to compatibility with existing data can still be useful. Firstly, compatibility is a necessary condition for the adequacy of a theory. By showing compatibility some progress is made (in the same way as some progress is made by showing that a phenomenon can be redescribed in a new vocabulary) as one is justified in increasing one's credence in the theory (this is, ironically, based on Bayesian logic, see e.g., Gershman, 2021, Chapter 9). Secondly, note that if PP were compatible with many phenomena, and no alternative framework is compatible with so many phenomena, then this would clearly increase the support for the PP framework. In any case, the accounts of concepts, metaphor and copredication I propose in this thesis involve comparisons with other (incompatible) accounts, and so aim to avoid the objection via this alleged "compatibility fallacy." However, I do suggest that the mere fact that one can apply the PP apparatus to such a diverse range of higher cognitive phenomena, does to some extent speak in favour of the PP paradigm (without, of course, claiming that it confirms or proves the accounts or PP paradigm decisively).

12.2. Some further challenges

There are at least two further challenges that have been brought up in the debate about PP that appear to undermine some of its core commitments that are very relevant especially for higher cognition. One is related to the critique by Litwin & Miłkowski above discussed that core concepts are used inconsistently, and this affects the plausibility of the principle of active inference. The second is related to the role of the precision weighting apparatus as a potentially dubious "magic modulator". Given the centrality of both elements in the PP paradigm, a tentative response to those objections is necessary.
12.2.1. Active inference—the problem of the direction of fit

Sprevak (2021c) points out that the way the PP paradigm integrates action and perception via the concept of active inference as dual aspects might be problematic. Let me use my own example to try to make Sprevak’s point. According to PP, action geared towards the consumption of a piece of cheesecake is triggered if two conditions are met. Firstly, a piece of cheesecake is being predicted to be in the mouth, though there is strong evidence for the agent that there is in fact no cheesecake in the mouth. Secondly, among the two possibilities to minimize this prediction error active inference is chosen, and not perceptual inference. The first leads to the consumption, while the second leads to the update of the belief that there is no cheesecake in the mouth.

The problem here is that the prediction or belief that there is cheesecake in the mouth (which is simultaneous to the evidence that there is no cheesecake in the mouth) cannot be of the same type as the belief that there is cheesecake in the mouth when there is indeed cheesecake in the mouth. The first is a desired state, while the second is a believed state. According to Sprevak, it is not clear how the difference between those two types of predictions that there is cheesecake in the mouth is realized representationally and algorithmically.

Note that this is an objection closely related to Litwin & Milkowksi’s complaint of incoherent uses of PP concepts, especially those that have computational and psychological applications. In the cheesecake example we use for both the desire and the belief the notion "prediction", but they must be qualitatively different sorts of predictions (namely of the belief-type and of the desire-type respectively). Indeed, we seem to have a problem here. However, based on the following observation I think a response is possible, though it needs to be fleshed out in more detail in further work.

Notice that in the starting situation where the agent merely desires but does not have the cheesecake in the mouth, the contradicting predictions (the bottom-up one that there is no cheesecake in the mouth; and the top down one that there is cheesecake in the mouth) are not settled or stabilized predictions. We do not manage to stably predict either of the two states. Neither of the two candidate predictions is gaining because both signals are tuned up, or both are clamped. This situation might
correspond to the desire state. In the belief state the predictions can be settled easily because belief and evidence coincide. Now this requires further elaboration, because why is it that the top-down predictions that there is cheesecake in the mouth is being clamped? Shouldn't the evidence that there is no cheesecake in the mouth win? (In fact, we pretty much know for certain there is no cheesecake in the mouth). One answer is that this clamping happens via the precision weighting apparatus. The precision weighting apparatus provides a prior that also tunes up the top-down signal to trigger action under certain circumstances. But how exactly this is being done, needs to be worked out. Also, one might immediately put forward the objection to which I will turn to in a moment, namely that the appeal to precision weighting is suspect. But the point I want to make here is that one can come up, at least in principle, with a story about why the predictions corresponding to belief and desire seem to be qualitatively different, and how to account for both types.

12.2.2. The "magic modulator"

The last objection I want to address is related to the role of the precision weighting mechanism posited by PP. A concern, which is even expressed by PP proponents (Clark, personal conversation), is that it seems to be a sort of "magic modulator". The precision weighting apparatus could be used to fit any explanation. The purpose of the precision weighting is to orchestrate the selection of relevant information depending on the goal and task, background or higher-level knowledge and context. Often disturbed dynamics of the precision weighting mechanism is appealed to in explanations of psychopathological phenomena. But how does the agent's precision weighting system determine what is important? By clarifying this, by the way, one would solve one of the harder problems of AI, namely the frame problem: how do we determine what is important (see also Chapter 9)?

However, it is precisely the power of hierarchical models that such "meta-knowledge", namely knowledge about what knowledge to apply in each situation, can be bootstrapped and learned as well. It is just higher order knowledge, i.e., priors on even higher levels of abstraction. If we did not have those biases that "correctly" lead to relevance judgements, we would not be around. There are no

---

126 "Important" signals are here understood as those selected for further processing (i.e., considered by the brain deserving to be tuned up).
explicit (propositionally represented) rules programmed into the brain about what is relevant. This is a simple and—admittedly—unsatisfactory answer, but it is the most plausible one. It is unsatisfactory because I have not provided any rules or mechanism but have just appealed to implicit learning and evolutionarily driven emergence. But demanding an explanation in the form of clear rules or functional modules might itself just be a cognitive bias that we are subject to when thinking about cognitive mechanisms. Such a bias might correspond to the LOTparadigm, which the PP framework precisely pretends to substitute (see Chapter 6, and more on that also in the next subsection).

12.3. The way forward

The previous section has already brought out some areas that need further development and clarification. Let me summarize them here and add some further points. I will also highlight certain obstacles that need to be overcome. Apart from the obvious observation that much of PP is still underspecified and more detailed accounts on an algorithmic and implementational level need to be provided by neuroscience (see e.g., Jian & Rao, 2021; Millidge et al., 2021; Sprevak 2021a-d, for a long list of examples), I want to focus on some areas that seem especially relevant for the progress in understanding higher cognition. I also want to reflect on some more fundamental potential obstacles.

Aspects of phenomenology and consciousness in relation to higher cognition ("cognitive phenomenology") remain an interesting area for further research. There happen to be much work underway related to consciousness and phenomenology (see Section 2.1.4.), and it should be explored how the insights could be applied to the phenomenology of higher cognition. Further, and relatedly, the different types of mental states like will, desire, belief, etc. should be fleshed out more carefully in terms of how their phenomenology arises within the PP model. PP seems uniquely positioned to bring new insights regarding cognitive phenomenology, because it is precisely certain perceptual phenomena that are early examples of the explanatory success of PP (like binocular rivalry).

In this dissertation I have also tried to build bridges to other research traditions and disciplines. It seems to speak for a paradigm if it can connect to other fields and
disciplines. For instance, I have tried to connect the PP account of concepts with traditional concept and metaphor research. I have also explored links between PP and Cognitive Grammar. I could only scratch the surface in an exploration of how PP could team up with those disciplines and traditions.

More formalizing and modelling\textsuperscript{127} would be desirable, though I will raise some concerns in a moment. It would be useful to see the emergence of a platform of tools for simulations based on PP principles, and a lively community around it. In fact, there is some work now in machine learning, using PP principles in deep neural network models, like PredNet (Lotter et al., 2015, 2016, 2020). This is a working model for predictive processing of video imaging (e.g., Mikkilineni et al., 2021; Rane et al., 2020; Ofner et al., 2021; Ofner & Stober, 2021; Zakharov et al., 2021). However, PP is still far from being a complete functional "cognitive model", like, for instance, Eliasmith's SPAUN (e.g., Eliasmith, 2013), for which an extensive open access simulation platform is available. Also, PredNet so far focuses on visual processing and not yet on higher cognition more generally.

However, I do see various fundamental obstacles that we might need to overcome before we can reach a deeper understanding of higher cognition within PP or start to pursue PP-based engineering and simulation of genuinely artificial intelligent agents who are able to think flexibly and conceptually. The first obstacle has to do with the potentially mid-boggling complexity involved in getting conceptual thinking off the ground, due to the need to identify the right set and alignment of initial biases/constraints/priors. There might be complex cascades of biases and meta-biases, and inborn biases and learned biases most likely need to be mixed in the right manner. Maybe to get at a human-level performance in many general higher cognitive tasks—beyond those specialist capabilities that deep neural networks already exhibit—we need to evolve many priors in extremely complex ways (much like genes). The question is what priors are hard-wired genetically; and what physiological constraints also serve as priors? We might not be able to make an explicit inventory of them and then program them into a complex raw net of canonical

\textsuperscript{127}It should be noted, however, that formal, mathematical modeling is increasingly being carried out within the Active Inference formulation of PP relying on the Free Energy principle, especially by Karl Friston and colleagues (see also Parr et al., 2022, for an introduction to Active Inference modeling).
circuits. Getting real conceptual thought off the ground might simply be too complex to manage in vitro.

One might want to have a more formal approach to higher cognition within PP. Here, I have avoided any appeal to formalisms and mathematics. But should we not want to apply a formal mathematical apparatus, like Friston's Free Energy model—which comes equipped with sophisticated mathematics—to the aspects of higher cognition? It seems difficult to see how this could be done in a way that is useful for engineering artificial intelligent agents. How could we derive equations about metaphors, for instance? Mathematical descriptions might simply not be the right level of description to understand phenomena like metaphor, semantic paradox or copredication. While I do think that the principles of the PP paradigm scale to higher cognition (and even beyond), I do not think that the quantitative treatment via sets of such equations scale in a way that is relevant for engineering and implementing artificial general intelligence.

The very idea of formalization might also be in fundamental contradiction to the very paradigm that PP represents. Notice that leading proponents of Construction Grammar (most notably Langacker and Goldberg), which I argued is close to PP, shy away from a formalization. This is because often the LOTH paradigm needs to be presupposed for formalization, and this is precisely the paradigm that PP and Construction Grammar aim to replace (see Chapter 6).

There might be some inherent obstacles for the progress in the project to provide a functional model of cognition within PP. Understanding how the brain works with respect to higher cognitive functions involves precisely making its workings transparent by identifying parts and pieces that carry out certain functions. In other words, understanding higher cognition is functional or high-level mechanistic analysis. But according to the PP paradigm, there is no clear distinction between perception/cognition/higher-lower cognition/action. While some degree of modularity or functional segregation is compatible with PP, the PP picture differs radically from more traditional cognitive science. The situation with PP might therefore turn out to be like what we encounter in modern machine learning with deep neural networks (DNNs). The workings of DNNs are not transparent at all, except in the form of some rough high-level approximation. I have already mentioned that certain explanations, like the ones involving the precision weighting apparatus, might not be very
satisfactory. Observed cognitive effects might be phenomena that are just emerging that cannot be explained in perspicuous mechanistic, functional (or rule-like) terms. Notice that there might be a certain parallelism with quantum physics, where we still struggle to understand it (in classical terms), though we know that the quantum mechanics apparatus works impressively well.

Despite those potential obstacles, I aimed to make the case that PP is a paradigm with many merits and that on the right level of description it can be fruitfully applied to provide a different perspective on higher cognitive phenomena. PP is not something entirely new but integrates elegantly many tested and fruitful predecessor ideas that are both revisionary and well-motivated.
Concluding remarks

I have tried to go some way towards arguing that the PP framework holds promise in terms of providing new approaches to higher cognition. Given that the notion of concept is such a central posit in higher cognition, I started out by sketching a model of concepts within PP. I also tried to build bridges to the huge literature on concepts in philosophy, psychology and neuroscience and locate my sketch within the debates about the format of concept (in the sense of the amodal/modal distinction, and in the sense of the types of knowledge represented: prototypes, exemplars, theories, and others). I connected PP and a linguistic theory, Construction Grammar, and argued that PP promises to be a cognitive computational underpinning of it, much like Fodor's LOT underpins generative grammar. With a PP account of concepts, language, and conceptual thought in hand, I tackled some challenging and fascinating open issues in the philosophy of language, linguistics, and psychology: semantic paradox, copredication and metaphor, in the hope of providing new perspectives on those. One strong motivation for applying PP to such diverse phenomena was the following. Science often progresses by an accumulation of theories and accounts developed for specific and narrowly delimited phenomena of interest. This leads to a fragmentation of the landscape. Often it is difficult to see connections between those individual specialized accounts. For instance, there is metaphor research, copredication research, research into semantic paradoxes, research into concepts, and so forth. Those fields exist largely as independent islands with little cross-communication. This specialization and fragmentation might be inevitable because this is just how things turn out to be, or it is the most efficient way to advance. The PP framework, so I hoped to show, has the potential to provide a more unified perspective and a set of concepts and principles that might provide cognitive-mechanistic accounts for a large range of phenomena of higher cognition.
References


Keil, F. (2021). Why are concrete to abstract models of conceptual change so appealing but so wrong? 74th Annual Meeting of the Society for Philosophy and Psychology.


