universität wien

DISSERTATION | DOCTORAL THESIS

Titel | Title

The Modeling Mind:
A Philosophy for the Sciences of the Mind

verfasst von | submitted by

Raphael Gustavo Aybar Valdivia MSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktor der Philosophie (Dr.phil.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 792 425
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Cognitive Science/Kognitionswissenschaft
Betreut von | Supervisor:	Ronald Sladky Bakk. MSc PhD Univ.-Prof. Tarja Knuuttila MSc M.Soc.Sc PhD Dipl.-Phys. Dr. Andrea Loettgers

This page intentionally left blank

To my parents

Acknowledgments

Writing this thesis has been a learning process for me—both in doing philosophy and navigating academia. For this, I am deeply indebted to my primary supervisor, Professor Tarja Knuuttila, who encouraged me to find my own voice. Her creative insights were instrumental in shaping my concepts of modeling frameworks, target construction, and ontological agnosticism. I am also grateful for her professional support, particularly in helping me secure funding and providing thoughtful advice and networking opportunities. I feel I have grown not only academically but also personally, thanks to her guidance.

I began writing this text several months before the COVID-19 pandemic was declared. During that period, I am grateful to Dr. Natalia Carrillo and Dr. Rami Koskinen, from whom I learned how to work within the philosophy of science. As the pandemic progressed, research conditions changed dramatically. I worked mostly from home, and interaction with peers diminished. I had nearly completed a first draft by the end of the pandemic—a somewhat obscure version that required significant effort to shape into what it is today. I can only hope the reader finds a clear articulation of ideas in this work.

In preparing this manuscript, I worked closely with my second supervisor, Dr. Andrea Loettgers, who read and commented on each chapter multiple times. She was incredibly supportive and generous with her time. In very kind ways, she consistently pushed me to become a more rigorous thinker. My interactions with my third supervisor, Dr. Ronald Sladky, were of a different but equally valuable nature. He introduced me to Predictive Processing and empirical work within this framework and helped refine my critiques of it. I am thankful for his keen interest in my work and his kindness.

While correcting this manuscript—and after my funding ended—I transitioned to a non-academic position, which I now fully enjoy, though it slowed my progress. However, as Tarja once told me, “Philosophy is a slow business." I was fortunate to pursue philosophy at a pace that allowed my ideas to develop organically.

Thank you to all my friends who supported me in various ways throughout this journey, especially Maximilian Noichl and Moritz Kriegleder. I am also grateful to Evelyn Fischer for proofreading the manuscript. Finally, I thank my parents, Juana Valdivia and Gustavo Aybar, my brothers, Juan Aybar and Ivan Aybar, and my sister-in-law, Giuliana Paredes.

This research would not have been possible without the support of the University of Vienna, particularly the Uni:docs Fellowship Programme for Doctoral Candidates, the Vienna Doctoral School of Philosophy, and the Department of Philosophy.

Abstract

This thesis explores two notions of modeling within the cognitive sciences from a philosophy of science perspective. The first concerns scientific modeling as an epistemically relevant practice—how the construction and manipulation of models contribute to knowledge production in the cognitive sciences. The second is the idea of mental modeling: the notion that minds model the world and themselves in order to perceive and act, as proposed by Predictive Processing. In this framework, cognitive behavior is conceptualized as inferential, probabilistic, and model-based.

My research question is: **What is the epistemic gain of modeling cognition under these assumptions?** To address this, I develop three philosophical concepts: *modeling framework*, *target construction*, and *ontological agnosticism*. I characterize Predictive Processing as a modeling framework—a constellation of conventionally linked epistemic resources, such as theories, assumptions, methods, and concepts, used to construct scientific models. Research under this framework affords plural epistemic gains, depending on how scientists construct and manipulate models.

The thesis focuses on the conceptual dimension of modeling, specifically the construction of target systems. I argue that scientific modeling is not primarily about representing reality but about creating instruments for investigating it. Predictive Processing, in particular, frames cognitive problems as probabilistic inferential tasks, transforming complex cognitive phenomena into heuristically transformed problems solvable through computational, mathematical, or other methods.

I argue that epistemic gains stem from the specific ways in which epistemic resources are drawn together in research practices. This includes the creation of both target systems and scientific models that support testable hypotheses, accurate predictions, and the development of new epistemic tools. Based on this account, I defend ontological agnosticism with respect to mental modeling. In particular, I argue that realist interpretations of Predictive Processing overlook its ontological neutrality as a modeling framework and its role in framing diverse cognitive phenomena in similar ways.

Zusammenfassung

In dieser Arbeit werden zwei Verständnisse von Modellierung in den Kognitionswissenschaften aus einer wissenschaftsphilosophischen Perspektive untersucht. Der erste betrifft wissenschaftliches Modellieren verstanden als erkenntnistheoretisch relevante Praxis - wie die Konstruktion und Manipulation von Modellen zur Wissensproduktion in den Kognitionswissenschaften beitragen. Das zweite Verständnis ist gekennzeichnet durch die Idee der mentalen Modellierung: die Vorstellung, dass der Verstand die Welt und sich selbst modelliert, um wahrzunehmen und zu handeln, wie das Predictive Processing suggeriert. In diesem Rahmen wird kognitives Verhalten als schlussfolgernd, probabilistisch und modellbasiert konzeptualisiert.

Meine Forschungsfrage lautet: **Was ist der erkenntnistheoretische Gewinn der Modellierung von Kognition unter diesen Annahmen?** Um diese Frage zu beantworten, entwickle ich drei philosophische Konzepte: *Modellierungsrahmen*, *Zielkonstruktion* und *ontologischer Agnostizismus*. Ich charakterisiere Predictive Processing als einen Modellierungsrahmen - eine Konstellation von konventionell verknüpften epistemischen Ressourcen wie Theorien, Annahmen, Methoden und Konzepten, die zur Konstruktion wissenschaftlicher Modelle verwendet werden. Forschung innerhalb dieses Rahmens bietet eine Vielzahl von epistemischen Vorteilen, je nachdem, wie Wissenschaftler:innen Modelle konstruieren und manipulieren.

Diese Arbeit konzentriert sich auf die konzeptuelle Dimension der Modellierung, insbesondere auf die Konstruktion von Zielsystemen. Ich behaupte, dass es bei der wissenschaftlichen Modellierung nicht in erster Linie um die Darstellung der Realität geht, sondern um die Erschaffung von Werkzeugen zu ihrer Erforschung. Insbesondere das Predictive Processing stellt kognitive Probleme als probabilistische Schlussfolgerungsaufgaben dar und transformiert komplexe kognitive Phänomene in heuristisch bearbeitbare Probleme, die durch computationelle, mathematische oder andere Methoden lösbar sind.

Ich vertrete die These, dass der epistemische Gewinn aus der spezifischen Art und Weise resultiert, in der epistemische Ressourcen in der Forschungspraxis zusammengeführt werden. Dazu gehört die Schaffung von Zielsystemen und wissenschaftlichen Modellen, die prüfbare Hypothesen, genaue Vorhersagen und die Entwicklung neuer epistemischer Werkzeuge unterstützen. Auf der Grundlage dieser Darstellung verteidige ich den ontologischen Agnostizismus in Bezug auf die mentale Modellierung. Insbesondere argumentiere ich, dass realistische Interpretationen des Predictive Processing dessen

ontologische Neutralität als Modellierungsrahmen und dessen Bedeutung bei der Rahmung verschiedener kognitiver Phänomene auf ähnliche Weise übersehen.

Table of Contents

List of Figures ***10***

List of Abbreviations ***11***

Introduction ..... ***- 13 -***

Overview of the Topic and Problems ..... **- 13 -**

Goals and Scope ..... **- 19 -**

Methods ..... **- 23 -**

Structure of the dissertation ..... **- 23 -**

Chapter 1 - A Philosophy of Scientific Modeling in the Cognitive Sciences ..... ***- 28 -***

1.1. Scientific Concepts in the Cognitive Sciences ..... **- 29 -**

1.2. Scientific Models in the Cognitive Sciences ..... **- 34 -**

1.3. Scientific Models as Representational Devices ..... **- 39 -**

1.4. Target Construction ..... **- 42 -**

1.5. Conclusions ..... **- 48 -**

Chapter 2 - Scientific Models in Predictive Processing ..... ***- 50 -***

2.1. Predictive Processing as a Scientific Modeling Framework ..... **- 51 -**

2.2. Modeling Assumptions in Predictive Processing ..... **- 54 -**

Bottom-up and Top-down Information Processing ..... - 54 -

A Functionalist Assumption in Cognition and Brain Research ..... - 56 -

Generative Models and the PP View of the Brain ..... - 57 -

2.3. The Predictive Processing Model of Perception ..... **- 61 -**

Target Systems: Probabilistic Inferential Tasks ..... - 62 -

Predictive Processing Models versus Phenomenological Models of Perception ..... - 64 -

2.4. Generative Models ..... **- 69 -**

What are Generative Models? ..... - 70 -

Prediction Errors and Sensory Signals ..... - 72 -

Estimates, Precision Estimates and Hierarchies ..... - 73 -

Generative Models as Epistemic Artifacts ..... - 75 -

Chapter 3 - Model Construction in Predictive Processing ..... ***- 79 -***

3.1. Model Construction ..... - 79 -
Model Construction in the Philosophy of Science ..... - 79 -
Model Construction in the Cognitive Sciences ..... - 81 -

3.2. The Action Observation Network Model ..... - 84 -
Biomotion Perception and the Action Perception System ..... - 84 -
The Action Observation Network ..... - 87 -

3.3. The Agent-Environment Coupled System Model ..... - 90 -
Targetless Modeling ..... - 90 -
The Agent-Environment Coupled System Model of Niche Construction ..... - 92 -
The Construction of a Scientific Model ..... - 95 -

3.4. The Epistemic Gain of Modeling the Mind as a Predictive System ..... - 98 -
The Epistemic Gain of Target-Directed Modeling ..... - 98 -
The Epistemic Gain of Targetless Modeling ..... - 100 -

Chapter 4 - Ontological Agnosticism ..... - 105 -

4.1. Mental Representations in the Cognitive Sciences ..... - 106 -
Internal Representations ..... - 110 -
External Representations ..... - 112 -
Embodied, Enactive, Extended and Embedded Cognition ..... - 114 -

4.2. Predictive Processing vis-a-vis Mental Representationalism ..... - 118 -
Objections to Conservative Predictive Processing ..... - 121 -

4.3. 4E Cognition and Ontological Agnosticism ..... - 125 -
Predictive Processing vis-a-vis 4E Cognition ..... - 125 -
Generative models as Extended Epistemic Artifacts ..... - 128 -
Ontological Agnosticism ..... - 131 -

Chapter 5 - A History of the Predictive Mind ..... - 135 -

5.1. A History of Perceptual Inference ..... - 136 -
Helmholtz' Theory of Perception ..... - 138 -
Inference as a Scientific Analogy ..... - 140 -

5.2. Cybernetic and Predictive Processing ..... - 144 -
A New Science of Systems ..... - 144 -
Ashby's Systems Theory of Cognition ..... - 149 -
Error as a Concept of Information Theory ..... - 151 -
Adaptive and Stable Systems ..... - 154 -
Models in Cybernetics and Predictive Processing ..... - 157 -
Why should we be Ontological Agnostics? ..... - 162 -

Chapter 6 - Transdisciplinarity in Predictive Processing ..... - 165 -

6.1. Transdisciplinarity in the Cognitive Sciences ..... - 166 -
Predictive Processing as a Modeling Framework ..... - 166 -
The Received View of Interdisciplinarity ..... - 170 -

Transdisciplinary Knowledge Transfers and Model Templates ..... - 171 -

6.2. Epistemic Resources in Predictive Processing ..... - 174 -

Conceptualizations of Cognition ..... - 174 -

Bayes' Rule as a Model Template ..... - 176 -

The Free Energy Principle as a Model Template ..... - 179 -

A Critique to the Metaphysics of the Predictive Mind ..... - 184 -

Conclusions ..... - 190 -

Bibliography ..... - 197 -

List of Figures

Figure 1: The Predictive Processing modeling framework 52

Figure 2: A visual example of top-down modulation of perception 57

Figure 3: A hierarchical generative model 72

Figure 4: Anatomical connectivity of the nodes of the Action Observation Network 88

Figure 5: A graphical representation of the AECS model 93

Figure 6: An agent-environment coupled system under the Free Energy Principle 95

Figure 7: The Homeostat 158

Figure 8: A regulator and a controlled system. 160

Figure 9: Bayes' rule 177

List of Abbreviations

4E	Enactive, Embodied, Extended, and Embedded Cognition
AECS	Agent-Environment Coupled System
ANN	Artificial Neural Network
AON	Action Observation Network
APS	Action Perception System
BMD	Biomotion Detection
CPP	Conservative Predictive Processing
DCM	Dynamic Causal Modeling
FEP	Free Energy Principle
PP	Predictive Processing
pSTS	Posterior Superior Temporal Sulcus
ToM	Theories of mind
vPMC	Ventral Premotor Cortex

Introduction

Overview of the Topic and Problems

This thesis develops an understanding of the concept of model employed in the cognitive sciences, with a particular focus on Predictive Processing, a prominent theory in this field. The examination of this theory is instrumental in articulating an understanding of the ways scientists deliver knowledge about cognition through modeling. Furthermore, it allows for the assessment of the validity of substantive philosophical ideas about cognition and the brain that are based on it.

The thesis put forth is as follows: the target systems of scientific models of cognition, the things they are models of, are constructed within the construction of the models themselves. This argument implies that scientific models of cognition do not represent cognitive processes, mechanisms, or entities as existing in the world independently of the scientific theorizing about them. To phrase it differently, cognition becomes an object of scientific inquiry because scientific modeling practices entail a conceptualization of target cognitive systems, specifically one that characterizes cognitive phenomena as solvable tasks. The epistemic gain derived from modeling cognition in this manner is related to the control, simulation, or prediction of the behavior of target systems; the generation of testable scientific hypotheses; and the development of new epistemic tools for investigating cognition. In light of these considerations, I make the case that Predictive Processing does not advance any metaphysical view of the mind.

Predictive Processing (hereafter PP) argues that cognition operates through predictive modeling and that the "(...) brain is a sophisticated hypothesis-testing mechanism." (Hohwy 2012, 1). Cognitive beings model their environment through action, transforming it into a predictable niche. Through perception, they discover regularities in their environment and use this information to form mental models that allow them to anticipate future environmental events. These models are constantly improved because predictions are not always accurate, as environments are dynamic. Cognitive agents improve their models by taking into account prediction error, which is the discrepancy between what is predicted and what is experienced.

When taken as a serious description of cognition and the brain, PP is quite controversial. One might ask whether it is possible to demonstrate that specific brain regions are responsible for the realization of predictive routines. In recent years, scientists have searched for empirical evidence to support PP (see Kleckner et al. 2017; Walsh 2020; Hodson et al. 2024). I approach PP through a very different lens. I do not seek to demonstrate that PP provides a realistic view of cognition, or the opposite, but rather to understand why it makes sense in actual scientific practices to conceptualize the brain as a predictive system, i.e., to determine the nature of the scientific contribution that can be made by this approach, as well its limitations.

To be more precise, it is not my intention to show that PP provides an accurate description of our mental mechanisms and mental life. I believe that accepting the realism of the current scientific view of the brain as a predictive machine is analogous to accepting the earlier scientific views of the brain as a kind of hydraulic pump, or the idea of the brain as an analog computer, which continue to have influence today. Although scientists may believe that the mind is a predictive machine, their private beliefs are not necessarily informative about the knowledge produced by modeling cognition as probabilistic inference. I try to find out what role this view of cognition plays in actual scientific practices.

In recent years, philosophers have engaged extensively with the conception of cognition put forth by PP. Realists about PP believe that the brain is literally a predictive machine, that predictive mental (generative) models are somehow materially realized in this organ, and that there is empirical evidence that certain brain regions are involved in predictive routines. Similar realist positions are common in the cognitive science literature. Many cognitive scientists believe that there are forms of natural computation and that brain activity is an instance of them. These claims are contested by antirealists. Instrumentalists maintain that these are fruitful analogies, metaphors, or fictions that facilitate understanding of reality, but they should not be taken as serious descriptions of reality.

This dissertation engages with the above discussions and defends an **ontological agnosticism**. In this context, ontological agnosticism is the philosophical position that in scientific practices, the concepts used to investigate reality are ontologically neutral. For example, a formal concept in logic or mathematics is an ontologically neutral concept. It does not refer to or assert anything substantive about real entities, mechanisms, or processes. According to this view, our concept of a model in the sciences of the mind does

not have to refer to any entity or structure in the world to have an epistemic function in scientific practices.[1]

My ontological agnosticism aligns with Toon's mental fictionalism (2023) in that it rejects both mental eliminativism (see Ramsey 2024) and realism (see Demeter 2009). In Toon's view, the ordinary discourse about the mind can be regarded as a useful metaphor. I suggest that a comparable process occurs in the scientific theorizing about cognition, namely, that the concepts employed to describe cognition in science are ultimately metaphorical. The utility of the metaphors in question is contingent upon the scientific achievements that result from their use. It is thus possible to be ontologically agnostic about the reality of predictive models and yet to investigate cognition under this assumption, insofar as doing so may yield epistemic gains.

Although scientific concepts are similar to formal concepts in logic and mathematics in that they are ontologically neutral, an important distinction exists between the two. Scientific concepts are not formal. The formation of scientific concepts depends on a number of factors, including the underlying assumptions on which they are based (for example, to model cognitive phenomena, scientists must assume whether cognition is realized in the brain, in the brain-body, or in the brain-body-environment), the available knowledge of scientists, the historical contexts in which they arise, and the material cultures. The point I am trying to make is that each generation of scientific communities has its own set of intuitive beliefs about the natural and social systems of interest that lead to the generation of diverse conceptualizations of target systems. These beliefs are closely intertwined with the available scientific analogies, metaphors, and technologies used in the investigation of scientific problems. This suggests that the choice of scientific concepts and instruments for investigation is determined by the material and socio-cultural realities of scientists.

According to my ontological agnosticism, the epistemic function of scientific concepts and other epistemic resources has nothing to do with capturing the true properties of the objects of inquiry. Rather, their utility can be traced to the epistemic work they perform in specific research practices. Consequently, an understanding of their epistemic function can only be achieved by conducting case studies of scientific experimentation, simulation, or modeling. Within these cases, concepts of cognition serve as workable descriptions that do not allow for ontological interpretations of the target systems under investigation beyond the

[1] A substantive claim in this context is an ontological claim about cognition. In other words, it is a statement about the nature of cognition itself, or about the underlying structures or physical processes that are necessary for its realization. The cognitive science literature contains many examples of such claims. These include the ideas that cognition is a form of computation, that it is representational, that it is non-representational, that it is information processing, that it is embodied, enacted, embedded, and extended.

boundaries of a particular scientific practice. However, within the context of specific scientific practices, they shape the target systems in various ways.

Applied to PP, my agnosticism boils down to this: ideas about cognition as model-based, predictive, or inferential in nature are merely helpful conceptualizations that render target cognitive systems amenable to scientific inquiry. However, their use in scientific practices does not necessarily prove that they represent cognitive entities, mechanisms, or processes in nature. Alternatively, models might be viewed as instruments employed by scientists, acquiring meaning within the specific contexts of their application. I believe this modest understanding has the potential to illuminate both the functions they serve and the contexts in which they are ill-suited.

In scientific practices, conceptualizations have the basic function of framing the target systems in ways that facilitate their investigation. This function has been referred to as "mind-framing" by Hasok Chang (2022, 71-72):

> Mind-framing is about our conceptualization of entities, and that is a wholly different issue from mind-control. Whenever we designate an entity by a concept, the entity is framed by the concept, which resides in our minds (or our languages). My view is that real entities are mind-framed even though most of them are not mind-controlled. Real entities are mind-framed simply because all entities are mind-framed. All that we can ever think or talk about are mind-framed entities.

Mind-framing, as Chang suggests, is different from mind control simply because the use of concepts to make sense of target systems does not lead directly to the control of such targets. For example, I can conceptualize the social world as consisting of the family, civil society, and the state, and argue that there is a dialectic that explains their dynamics. However, simply by making such a conceptualization, I do not gain control, predictability, or any other epistemic benefit over any of these targets. For this, scientists rely on mathematical, computational, and other epistemic resources that operate on workable characterizations of targets that are the results of mind-framing activities.

Mind-framing is distinctive in cognitive science because scientists study the mind through the lens of epistemic resources such as scientific concepts, models, and technological means. These epistemic resources are themselves cognitive, or at least extensions of scientific cognition. Consequently, scientists' use of these epistemic resources shapes the way they reason about and imagine systems of interest, as they often project these resources' properties onto the systems they study. I argue that in PP, scientific conceptualizations of target systems—such as mental structures as generative models, perception as probabilistic inference, or cognition as prediction-error minimization—

become intertwined with the properties of the scientific models employed to investigate them.

Research in PP is heavily influenced by epistemic tools from Bayesianism, a statistical approach that is useful for addressing problems of probabilistic inference. In statistics, generative models represent the joint probability distribution of two variables (x, y), that is, they estimate the probability of y given x. A simple example is that a generative model can estimate the probability of the next letter in a sequence of letters, such as "W-O-N-D-E-R-F-U". In this example, the model can determine that L is the most likely letter to follow. To achieve optimal predictive accuracy, generative models are trained by analyzing large data sets. As I mentioned above, research based on PP attributes the inferential or predictive properties of these models and their components to cognitive problems as their target systems.

In the construction of scientific models of cognition, the target systems being modeled are framed in ways that depend on the tools used in their construction, which sometimes include mathematical and computational tools. If this is the case, then views of cognition as predictive modeling and perception as probabilistic inference are determined by the material culture of science. Moreover, they have emerged from scientific practices rather than from philosophical debates. Within these practices, statistical and other concepts, analogies, and tools, including those mentioned above, have played a central role in creating the language in which cognition becomes a probabilistic and inferential task.

From the perspective I advocate, it makes little sense to assume that scientific models are representations of mental phenomena, entities, processes, or mechanisms. If they are not: What is the epistemic value of exploring cognition with them? To understand this question, I argue that it is necessary to distinguish between two meanings of the concept of a model in PP. Only then can we properly assess its contribution to our understanding of cognition by distinguishing what belongs to the science and what belongs to the philosophy of PP. I now introduce these concepts of a model.

A **scientific model** is a type of instrument used by researchers in various scientific fields to address theoretical and practical problems. This type of instrument has been the subject of extensive research in contemporary philosophy of science, which can be explained by the current prominence of model-based science. Scientific modeling is arguably the default mode in which theoretical science is currently practiced (Gelfert 2011), and the cognitive sciences are no exception. Following Knuuttila (2011, 2017a), I understand scientific models as epistemic artifacts, that is, human-made instruments that scientists construct and manipulate to produce knowledge. In Chapter 1, I articulate my understanding of scientific models as epistemic artifacts in the cognitive sciences.

The second concept of a model belongs to the cognitive sciences. It denotes putative mental entities or structures known as **mental models**. Sussman (1975, 277) defines mental entities as "whatever is ascribed to a person when a mental predicate is applied to him [her, them] [...] It may be a mode of consciousness, a set of behavioral dispositions, a brain state, or what have you." Anger, fear, love, and sensations are examples of mental entities by this definition. Sussman thinks of mental entities as theoretical entities because they are used to make sense of or explain some behavior. As I understand it, cognitive scientists posit similar theoretical entities when they talk about mental models or conceptualize cognition as predictive modeling. That is, they postulate that mental models exist to make sense of various cognitive phenomena, such as observable behaviors (e.g. tasks involving visual perception, memory, or attention) or neural activity.

A clear distinction and proper understanding of these different concepts of models is necessary in PP, as one of its main concepts is that of a generative model. This concept is discussed throughout this dissertation, particularly in Chapter 2. For the time being, it is sufficient to state that my distinction between the two notions of models reflects the ambiguity of the concept in the PP literature. Generative models are scientific models that scientists employ to investigate problems by framing them as tasks of probabilistic inference. They have been used across several domains of science, including PP, where they are used to model perception, attention, action, and many other phenomena. The same term is used to refer to hypothetical mental models that, not coincidentally, have similar properties to the scientific models used to study target cognitive systems.

To clarify this issue, it is necessary to distinguish between the two types of generative models in PP. The first refers to the type of scientific model used to study the alleged mental structures responsible for cognition and behavior. The second refers to the mental structures themselves. To avoid confusion, I will clarify whether I am referring to the scientific model or the mental structure whenever I use the term "generative model."

Having distinguished the two concepts of a model in PP, I now introduce the interpretation of generative models as mental structures, entities, or mechanisms. Generative cognitive systems function as efficient perceivers and actors because they model both the world and themselves. In perceiving the world, they construct a model of their environment, which they use to anticipate future events. These models are generative in that they produce new predictions and continuously refine them in response to environmental changes.

Through their actions, generative cognitive systems transform the world into a predictable niche. Before acting, they forecast the consequences of their actions in specific situations. Without going into detail, generative mental models can be understood as cognitive structures that process information by correcting the errors of their self-generated

predictions. Specifically, a generative model anticipates future environmental and internal states based on probability distributions and refines its predictions by learning from its own prediction errors. In the PP literature, generative models have been described as realized in the brain, in the brain and body, or as extended structures in the environment.

PP is not the only approach using generative models in science. On a global scale, predictive, generative, and machine learning models are changing the way science is practiced. However, the current prominence of generative models has not yet attracted the attention of philosophers. The notion of generativity was already used by Husserl in another context, namely a theory of intersubjectivity (see Steinbock 1995). But so far it has not been examined as property of a class of scientific models, nor have the epistemological implications of using these models in science been explored (two exceptions in this regard are Tee 2020 and Andrews 2021). I expect that this dissertation will help fill this gap in the literature.

In summary, this dissertation provides an understanding of the two notions of model at play in PP. It argues that the idea of models in the mind is rooted in scientific practices in which the properties of models are projected into reality. In other words, mental models are a kind of by-product of the mind-framing activities of scientists. Mental models are metaphysical constructs that arise from misinterpretations of scientific knowledge production. As I will show, the use of generative models to investigate cognitive problems leads to a projection of the properties of these models onto cognitive phenomena. However, this process need not lead to metaphysical views, and in fact conceptualizing cognition as a generative process is necessary for scientific practices in PP. I refer to the process of creating an object of inquiry in scientific modeling practices as **target construction**. I elaborate on this concept in the next parts of this introduction and in Chapter 1.

Goals and Scope

The objective of this dissertation is to elucidate how modeling cognition as probabilistic inference, as is done in PP, broadens our understanding of the brain and mind. In order to achieve this objective, three arguments will be developed.

The first argument is that PP is a **modeling framework**. The second argument is that the target cognitive systems are created in the construction of scientific models. I call this **target construction**. The last argument is my **ontological agnosticism** about the reality of mental models. My account of modeling frameworks is presented in Chapters 3 and 6, of target

construction and scientific modeling in Chapters 1 and 2, and of ontological agnosticism in Chapters 4 and 5.

A significant part of my work on this dissertation has been to determine the best terminology to convey my ideas. This is a particularly difficult task in disciplines such as the cognitive sciences, where terminology is not consistently defined or used. For example, in PP there is no consensus on how to refer to the general approach to cognition as predictive modeling. Over the years, attempts have been made to distinguish between PP, the Bayesian Brain, Predictive Coding, Active Inference, and the Free Energy Principle (see Anderson and Chemero 2013; Gallagher and Allen 2018; Piekarski 2021; Williams 2018; Zahavi 2018). Discrepancies and inconsistencies in terminology are to some extent unavoidable. For example, Hohwy (2013) notes that the same terms are used to describe phenomenological experience, neural activity, and computational models. Another example is Friston's concept of free energy, which is sometimes used to refer to an informational quantity and sometimes to a thermodynamic one (Kim 2023). I have also previously referred to PP as a theory in this introduction, despite my reservations about calling it a theory. One of the strengths of referring to it a modeling framework, as I explain in Chapter 6, is that it eliminates the need to introduce an additional *ad hoc* designation for what each of the above names should refer to.

My choice of the concept of target construction is a response to the excessive attention that the philosophy of science modeling literature has paid to the alleged capacity of scientific models to represent target systems, as if target systems were pre-given, transparent entities ready to be captured by models. Very often, modelers lack a clear understanding of what they are studying, and their understanding of targets emerges in the process of constructing a model. What are target systems, if not our best understanding of them? If there is no clear distinction between the understanding of targets and the targets themselves, then it is reasonable to conclude that targets are constructed in the process of constructing models. By articulating the concept of target construction, I argue that model construction also involves constructing a research object. Whether this research object corresponds to a cognitive phenomenon in the cognitive sciences, I remain agnostic.

Finally, ontological agnosticism is my response to the philosophical debates about the reality of mental models that follow from my examination of the science of PP. In the context of modeling practices, concepts are not designators of natural entities; rather, they are operational descriptions that acquire their roles when used to address scientific questions and problems. Since target systems are framed by these concepts, but the concepts are not designators of natural cognitive mechanisms or processes, neither are the resulting target systems representations of cognitive entities.

Having presented the rationale for my choice of concepts, I will now elaborate on the first. In cognitive neuroscience, PP is often referred to as a theory of how the brain processes information. In other cases, PP is presented as a **framework**. It is the latter that I am interested in. A framework can be defined as a constellation of epistemic resources, such as concepts, rules, methods, and assumptions, that are relevant to scientific practices. This concept is useful because it considers both the conceptual and material resources used in research practices.

I refer to PP as a **modeling framework** because the main strategy researchers use to address questions and problems in the cognitive sciences is scientific modeling. In my understanding, a modeling framework is a set of related epistemic resources that provide scientists with a *modus operandi* for modeling target systems. These resources may be conceptual, such as concepts and assumptions, or methodological, such as computational and mathematical tools. When used to model cognition, they conceptualize it in a way that scientists can solve using the epistemic resources at their disposal. The heuristic strategy of conceptualizing cognition as predictive modeling involves transforming cognition into behavior that can be analyzed, manipulated, and controlled using the framework's epistemic resources.

One aspect of scientific modeling that the thesis pays special attention to is the fact that target systems modeled within the scope of a given modeling framework are conceptualized in a uniform manner. This is at least the case for PP models of cognition. Target cognitive systems, including memory, attention, action, perception, and brain activity, are all conceptualized as predictive tasks within this framework. An elaboration of the idea that target cognitive systems in PP are cognitive tasks can be found in Chapters 1 and 2.

Given that modeling frameworks provide a means for conceptualizing different target systems in a uniform manner, their conceptualizations should not be understood as abstractions that isolate the fundamental or core properties of natural phenomena. From this perspective, scientific models of cognition are not instruments that represent, isolate, or map the mechanisms or processes in nature that are responsible for observable cognitive behavior. Instead, they frame cognition as tasks that can be solved using the available epistemic resources of the respective framework. This is how I think **target construction** occurs in the cognitive sciences. My view is that modeling cognition is not about accurately representing cognitive entities or structures in reality. Instead, it is about constructing simpler and solvable cognitive tasks.

A case exemplifying target construction that is extensively discussed in the dissertation is that of scientific models of perception in PP. These models conceptualize perception as the task of probabilistically inferring the unobservable causes of sensations. This

characterization is dependent on the epistemic resources involved in the construction of the scientific models in question. In this context, Bayesian statistical concepts and tools are particularly suited to deal with tasks of probabilistic inference. The argument is that scientific models of perception in PP can address issues related to perception because they conceptualize it as an inferential task, the kind of task that the statistical tools in question are well-suited to address.

A last objective of this dissertation is to outline the implications of my examination of scientific modeling in PP for ongoing philosophical debates about it. In my view, the most important implication is **ontological agnosticism**. Having already introduced this notion, I will now explain why it follows from my account of scientific modeling in PP. Ontological agnosticism is my response to a specific debate in the philosophical literature on PP. For the sake of brevity, I will give only a brief overview of the two main contenders, namely mental representationalism and 4E (enactive, embodied, extended, and embedded) cognition. For a more detailed characterization, I refer the reader to the work of Allen and Friston (2018) and Chapter 4. The debate between the two contenders concerns the nature and reality of mental generative models.

Mental representationalism is the hypothesis that cognition is information processing over semantic structures. Internalism adds that it is a brain-bound process. Proponents of PP committed to this hypothesis believe that generative mental models in PP are internal mental representations. 4E cognition, on the other hand, is a research program in cognitive science that argues that cognition is not brain-bounded and involves the interplay between a cognitive agent and its environment. Those who are committed to this view and PP argue that generative mental models are cognitive structures that are realized in the agent-environment coupling.

My position, which I articulate in Chapter 4, is that generative models in science are ontologically neutral epistemic resources, not mental entities or structures. And neither do these epistemic tools serve the purpose of mapping mental models realized in the brain or the agent-environment coupling. Arguing for the opposite, that is, for the reality of generative models, happens when scientists or philosophers conflate the tools used to investigate reality with reality itself. An examination of scientific practices that employ these models reveals that there is no inherent need for scientific practices to posit the existence or nonexistence of mental models in the first place.

Before concluding this section, I would like to emphasize that the science of PP is not yet mature, although a substantial amount of research in cognitive science is being conducted under its umbrella. Moreover, generative models are becoming the default scientific models for studying cognition, with applications to almost every research topic in the field. However,

the mere fact that generative models are becoming the dominant scientific instruments for investigating cognition does not necessarily give PP the status of an established theory of cognition. It is not clear what the research under its umbrella accomplishes. Moreover, its terminology is inconsistent. By developing a philosophy of science account of PP and articulating the epistemological implications of modeling within this framework, my contribution will be to clarify the kind of epistemic work that PP does for scientists. This work has implications for philosophical debates about it, because we cannot know whether PP favors a substantive conception of cognition if it is unclear what kind of knowledge about cognition is gained from research practices based on it.

Methods

My research methods are eclectic because they are tailored to the specific goals of each chapter. For this reason, I will present them as I outline the structure of the dissertation. The resources I use to study PP include concepts, views, and methods from the philosophy of science, such as case studies. I engage with the literature on scientific modeling and propose the concepts of modeling framework and target construction, as explained earlier, to make sense of the research practices of modeling cognition as probabilistic inference. I also draw on literature from theoretical debates and transdisciplinarity in the cognitive sciences to develop my account of modeling frameworks. Finally, I engage in philosophical debates about PP and the history of some scientific concepts to articulate my ontological agnosticism. I will now briefly outline what I do in each chapter.

DeepL Write and ChatGPT were used in the dissertation to stylistically improve individual passages limited to linguistic editing.

Structure of the dissertation

Chapter 1 presents my perspective on scientific models and target construction in the cognitive sciences. Chapter 2 introduces the reader to PP from this perspective. Chapter 3 addresses the question of why it is epistemically fruitful to model cognition in PP by examining two cases of scientific modeling practices. Chapter 4 introduces the philosophical debates about cognition that follow from PP and elaborates on my ontological agnosticism. Chapter 5 provides further support for my ontological agnosticism by

examining the history of the main concepts of PP. Finally, Chapter 6 further articulates my view of PP as a modeling framework by taking a closer look at the transdisciplinary epistemic resources used by researchers working under its umbrella. Each chapter is described in detail below.

Chapter 1 presents my philosophical perspective on scientific modeling in cognitive science. First, I examine the role of conceptualizations of cognition in the discipline. To do so, I distinguish between the philosophical and scientific uses of concepts. I present concepts in cognitive science as operational descriptions relevant to research practices that frame cognitive behaviors as solvable tasks. Then, I challenge the prevailing view that cognitive science is primarily concerned with explaining complex cognitive phenomena. Instead, I argue that cognitive science is a model-based discipline with epistemic aims beyond mere explanation. Finally, the chapter discusses how models and target systems are co-created in modeling practices.

Chapter 2 is a philosophical introduction to PP. The main goal is to explain what generative models are, when they are understood as epistemic artifacts with conceptual and material dimensions. The chapter focuses mainly on their conceptual properties. Their material dimensions are examined in chapters 3 and 6. The chapter first examines the assumptions about cognition and the brain that operate in PP, namely a view of information processing in the brain as both a top-down and a bottom-up phenomenon, a functionalist approach to the brain and cognition, and a view of the brain as a predictive machine.

The chapter then takes the scientific model of perception as an example of a scientific model in PP. It shows, by comparison with a phenomenological model of perception, how the scientific model is reductive in the sense that it constructs a target cognitive phenomenon as a task. Perception in PP is described as the task of a system that predicts the unobservable causes of its sensory inputs using a hierarchical generative model. This model is hierarchical because it is composed of different layers that send information to each other in a bottom-up and top-down manner. Upper layers make predictions about the information in lower layers, while lower layers correct these predictions. A perceptual task occurs when the higher-level information accurately predicts what is happening at the lower levels. When this task is not accomplished, the system corrects its top-down predictions by minimizing error, i.e., it computes the mismatch between predictions and evidence to make future predictions more consistent with the evidence. PP systems also form precision estimates that determine a learning rate for prediction error and the weight of priors. The last part of the chapter explains the main concepts of PP, namely predictions, prediction error, precision estimates, prior, and how they together provide a generic model that can be used to study any target cognitive system.

Chapter 3 addresses the question of the epistemic gain from modeling cognition as a probabilistic inference. It does so in three steps, the first of which is an articulation of my view of model construction as a scientific strategy for addressing questions and open problems in the cognitive sciences. The chapter then presents two case studies of model construction in PP and then reflects on how the epistemic gains from modeling cognition as a predictive process are diverse and contingent upon specific research practices.

The first case study is the Action Observation Network, a model of the communication of brain areas involved in action perception. I show how the construction of this model enables scientists to formulate testable hypotheses and specific scientific predictions about the behavior of a target system. The second case study concerns the Agent-Environment Coupled System, a model of niche construction. The construction of this model does not increase our knowledge of a natural phenomenon. Rather, the epistemic gain from constructing this model has to do with theory development, broadly understood as the creation of new scientific models or the application of existing ones to new target systems.

Chapter 4 develops my agnostic position on the philosophical discussions of the predictive mind. The first part examines the contenders in these discussions by reviewing the concepts of representationalism, internalism, distributed cognition, embodied, embedded, extended, and embedded cognition. The second part examines the so-called conservative PP, which is a representationalist, internalist, and realist interpretation of mental models. I raise several objections to this interpretation and show that it erroneously attributes the properties of scientific models of cognition to mental structures. I show that the same scientific models can characterize cognitive systems that are neither representational nor internal. The fact that the same model can be used to characterize extended and distributed cognitive systems undermines the internalist, representationalist, and realist assumptions of conservative PP. The overall lesson of this chapter is that PP tolerates, but does not necessarily imply, multiple philosophical interpretations of mental modeling.

Chapter 5 provides further support for my ontological agnosticism. It does so by developing a historical perspective on the concepts of inference, prediction, error, and model used in PP. The chapter shows that these concepts originated in scientific approaches that used them as analogies for different purposes. Helmholtz was the first to introduce the notion of inference into the science of perception, using it as an analogy to bridge bottom-up physiological and top-down psychological theories of perception. Moreover, he explicitly stated that the analogy should not be taken as a literal description of perception as inference.

After examining the notion of inference, the chapter explores the historical influence of cybernetics in PP for two reasons. One is to examine what is understood as a model in the

cognitive sciences. The second is to show that cybernetic concepts, such as information processing, behavior, or systems, and analogies, such as the mind and computers, which are fundamental to the cognitive sciences, belong to a new kind of science that does not aim to gain knowledge of reality by representing it or discovering the fundamental laws of cognition, but instead aims to gain control and predictability of observable behavior. This historical perspective supports my ontological agnosticism. Among cyberneticians, Ashby has been particularly influential for PP. He developed an anti-realist approach to the notions of model and error currently used in PP. These historical considerations support my view of cognitive science concepts as operational descriptions rather than representations of natural cognitive processes, entities, or mechanisms.

Chapter 6 provides further support for my view of PP as a modeling framework, i.e., a constellation of conventionally related epistemic resources such as assumptions, concepts, modeling techniques, and methods relevant to model construction. In the first part of the chapter, I critically examine traditional views of cognitive science as an interdisciplinary field. The need for interdisciplinarity in this field is often associated with the strategy of explaining cognitive phenomena through multilevel explanations that account for their complexity. I criticize this narrative by exploring the dynamics of knowledge transfer in this discipline. I suggest that in many cases interdisciplinary scientists are not interested in developing multilevel explanations by aggregating partial disciplinary explanations.

Against this conception, I argue that cognitive science is a transdisciplinary field of research, and that what unites the different disciplines involved in the study of cognition are indeed modeling frameworks that gather epistemic resources to investigate questions and problems. The remainder of the chapter is devoted to exploring the epistemic resources used in the PP modeling framework, which include model templates such as Bayes' rule and the Free Energy Principle. These templates provide conceptualizations associated with mathematical and computational resources that provide generic ways to construct target systems. They are transdisciplinary tools used to study systems in fields as diverse as biology, psychology, and neuroscience. When applied to questions and problems in these fields, they add mathematical and informational rather than disciplinary constraints.

Taken together, the chapters develop a new perspective on scientific modeling in PP that is informative for the epistemology of the cognitive sciences and for ontological discussions of mental modeling in the philosophy of mind. By focusing on actual research practices and on the properties and affordances of epistemic resources in PP, they clarify the epistemic work done by modelers in this framework. If concepts such as prediction, inference, model, and the like do not refer to actual things, but rather to the scientific means of understanding, predicting, controlling, intervening, or simulating them, then philosophers should not use

them to justify metaphysical claims about the mind. Finally, it is possible to be agnostic about mental modeling realism and mental representationalism while recognizing that these are to some extent irrelevant issues for the science of PP. An overemphasis on them can also be problematic in that it can be detrimental to the exploration of what the cognitive sciences are currently accomplishing, which requires an examination of actual scientific practices.

Chapter 1 - A Philosophy of Scientific Modeling in the Cognitive Sciences

This chapter articulates my philosophy of science approach to modeling in the cognitive sciences. Section 1.1 begins by examining the role that scientific concepts play in this field of research. To this end, I compare scientific and philosophical concepts of cognition. Whereas philosophical concepts are supposed to concern the nature of cognition, scientific concepts of cognition are characterized as workable descriptions that function as epistemic instruments within research practices. These instruments, it is argued, have the primary function of articulating an understanding of the target systems investigated in these practices.

Section 1.2 presents an alternative perspective to a dominant narrative in the cognitive sciences, namely that scientists are primarily concerned with explaining complex cognitive phenomena. Instead, I propose that the construction and manipulation of models are the primary scientific strategies employed by researchers in this field. The section finally discusses the role of models in the construction of target cognitive systems.

Sections 1.3 and 1.4 present my perspective on scientific models in cognitive science. Section 1.3 examines the assumption that scientific models represent phenomena. Specifically, I challenge the notion that scientific models depict the underlying processes or mechanisms of cognitive behavior. Instead, I argue that scientific models should be regarded as representational instruments, with the capacity to render tangible diverse conceptualizations of cognition.

Section 1.4 examines the construction of targets and the conceptual and methodological dimensions of models. The argument is put forth that target systems acquire the properties of the models employed in their investigation, specifically those pertaining to their conceptual dimension. This is illustrated by target systems in PP, which represent perception as a form of probabilistic inference. The present argument is developed in discussion with the literature of the philosophy of science, with reference to the fictionalist and artifactualist views of models. I conclude by proposing that an understanding of target systems in cognitive science requires conceptualizing cognitive behaviors as tasks that some systems perform.

1.1. Scientific Concepts in the Cognitive Sciences

Cognition has multiple meanings in science. It is arguably an umbrella term that can refer to a wide range of phenomena, behaviors, and processes in different types of beings and entities that are research objects in the cognitive sciences. The term can refer to the behavior of an animal constructing its niche, human social behavior, memory systems of different species, and even the cognition of a scientist. Cognitive scientists study human and animal cognition and, not without controversy, the cognitive behavior is attributed to artificially intelligent systems and organisms without nervous systems. Living systems are said to be cognitive or sentient because they perceive changes in their internal states and in their niches. In cognitive science, this usually amounts to perception. Furthermore, they are said to be cognitive to the extent that they respond adaptively to such changes, which amounts to action. They rely on a variety of cognitive capacities such as learning, language, memory, tool use, and social behavior to perceive and act.

The concept of cognition not invented by cognitive scientists. Before cognition became amenable to modern scientific investigation, it was a well-established philosophical topic. The advent of the scientific study of the mind and cognition did not put an end to philosophical debates about it. Philosophers of mind have proposed a variety of accounts concerning the nature of the mind, its relationship with the body and brain, and whether it is material or immaterial. These accounts are known as (philosophical) ontologies of the mind. They are inquiries into "the most fundamental metaphysical categories to which the mind may belong" (Jacquette 2009). Philosophers have proposed many different answers to these questions, some of which are mutually exclusive. These include monism, dualism, eliminativism and reductionism, as well as a variety of other positions that can be seen as subspecies or variations of these views. The reader is directed to the introduction of Jacquette (2009) for a definition of each of them. For the purposes of this introduction, however, it is preferable to focus on the shared strategies employed by these approaches.

One approach that philosophers have taken in formulating accounts of the mind is to consider the insights provided by the sciences regarding the mind and to develop accounts that align with these insights. This is exemplified by the early functionalist theories, which posit that the mind can be understood as an informational system analogous to a computer program, and that the embodiment of cognition is to some extent irrelevant. These ideas first emerged on the philosophical scene in the 1960s, just a few years after Turing and Ashby drew the first parallels between the brain and the computer. As Cobb (2021, 4) notes, these scientists soon recognized that the analogy between the brain and computers was not entirely accurate. They observed that "even the simplest animal brain is not a computer like

anything we have built, nor one we can yet envisage." Nevertheless, the inaccuracy of the analogy has not kept scientists from using it. As Cobb acknowledges, "the brain is not a computer, but it is more like a computer than it is like a clock, and by thinking about the parallels between a computer and a brain we can gain insight into what is going on inside both our heads and those of animals" (2021, 4).

Philosophers have also employed methods such as reflective analysis and thought experiments,[2] on the assumption that rigorous investigation can facilitate the identification of the defining characteristics of the mind. One strategy is to identify the necessary and sufficient conditions for something to be treated as cognitive or mental. An example is provided by Pernu (2017), who identifies intentionality, consciousness, free will, teleology, and normativity among these conditions, while acknowledging that cognitive agents need not possess all of them simultaneously.

Philosophers and scientists sometimes regard their approaches to the mind as radically different or even unrelated. They distinguish philosophical first-person descriptions, reflective analysis, and analytical distinctions from scientific efforts to provide accounts of the mind informed by empirical evidence that deal with observable or measurable entities. However, this neglects the fact that despite their methodological differences, philosophers and scientists can cooperate, and their work can intersect.[3] In particular, both groups engage in conceptual work when studying cognition. It is worth pointing out, though this may seem obvious, that conceptualizations are neither neutral nor empty vehicles for an objective or transparent representation of reality.

Not only do philosophers and scientists frequently neglect to consider the historical development of the concepts they employ in their studies of cognition, but they also frequently fail to take into account the diverse contexts in which these concepts are used. This is despite the fact that cognition, behavior, and similar concepts are not univocal. For example, the concept of inference is used in logic, probability theory, and psychology. In cognitive science, the statistical and psychological concepts of inference and reasoning are

[2] Rachel Cooper (2005) defines a thought experiment as a "judgment about what would be the case if the particular state of affairs described in some imaginary scenario were actual."

[3] For example, Cooper (2005) suggests that thought experiments in science and philosophy can be treated together. "Moreover, because of its necessarily nonempirical nature, work involving thought experiments is particularly likely to fall on the border between philosophy and science. Articles on the E.P.R. experiment, Schrödinger's cat, or bilking experiments (thought experiments showing that causal paradoxes would emerge if one could go back in time and kill one's father) are as likely to be found in the Physical Review as in the Philosophical Review. Newcombe's paradox is discussed equally by economists and philosophers. Psychologists and philosophers alike worry about the Turing Test and Searle's Chinese Room. It is hard to distinguish science from philosophy and even harder to distinguish philosophical from scientific thought experiments" (Cooper 2005).

often used interchangeably. For instance, Chater, Tenenbaum, and Yuille (2006) point out that "probability theory has always had a dual aspect, serving both as a normative theory of 'correct' reasoning about chance events and as a descriptive theory of how people reason about uncertainty-for example, as an analysis of the mental processes of an 'intelligent' juror."

Since the distinction between philosophical and scientific concepts is not the focus of this chapter, it will not be explored in detail here. My aim of introducing this distinction was to challenge the idea that philosophical and scientific concepts of cognition are essentially equivalent. Philosophical concepts are relevant to ontologies of the mind, while scientific concepts, as we shall see, play different functions. Despite their methodological differences, however, they share to some extent a concern with the nature of the relationship between the concepts involved and the mental entities they refer to. I am agnostic about whether the concepts of the mind map, represent, or designate mental entities and kinds. My view is that the sciences of the mind study target systems, which are theoretical entities created in research practices determined by the relevant scientific concepts.

Consider the scientific concept of intelligence as an illustrative example. In *A History of Intelligence and "Intellectual Disability"*, Goodey (2011, 2) posits that the concept of intelligence as it is understood in modern cultures is not a natural kind, but rather a historically contingent form of human self-representation and social reciprocity. Goodey does not simply assert that the concepts of intelligence and "intellectual disability" are modern cultural products. Rather, he contends that the concept of intelligence as a means of distinguishing human species membership is a modern phenomenon. A scientific conceptualization of a given phenomenon need not remain confined to the domain from which it originates; rather, a concept originating in science can transcend that domain and influence how a phenomenon is understood in the broader social context. Goodey explains how the scientific concept of intelligence plays an essential role in modern cultures as follows:

> Intelligence stands at the core of modern lives. It marks us out from the rest of nature. It is crucial to our sense of self and an instant yardstick for sizing up others. Psychologists measure it, biologists search for its DNA, women demand it of sperm donors; learned professors from Harvard to Heidelberg foresee our descendants turning into transhuman, bodiless intelligences able to migrate as software to other planets. If these are the dreams of intelligence, the nightmare is its absence. This means being denied family, friends and ordinary relationships; doctors give us treatment without our consent and withhold it when we need it; social workers stop us having sex, sterilize us or take away our children; psychiatrists lock us up without

> right of appeal; police officers frame us; courts acquit the parents who kill us; and politicians fund geneticists to make sure people like us never turn up again. Both dream and nightmare are so vivid it seems they must be based on some hard scientific reality (...) (2011, 1).

Goodey suggests that the activities of psychologists in measuring and classifying forms of intelligence transcend the realm of scientific practices to become influential in certain social interactions. Without a classification of an individual as "intellectually disabled," a medical professional would not be permitted to treat that individual without consent in certain circumstances. One method for analyzing the function of scientific concepts is to examine the social practices they enable. Another avenue of inquiry is to investigate whether and how the scientific understanding of intelligence is shaped by the concepts and tools available to scientists. This is the approach that I am going to take. An example can be found in Ashby (1991, originally published in 1958), specifically in the idea that the measurement tools used to study intelligence determine what is understood by intelligence in science. As Ashby (1991) puts it:

> «Intelligence» today is defined by the method used for its measurement; if the tests used are examined they will be found to be all of the types: from a set of possibilities, indicate one of the appropriate few. Thus all measure intelligence by the power of appropriate selection (of the right answers from the wrong) (413).

The concept of intelligence as the power of appropriate selection is related to the scientific practice of measuring intelligence with tests. But it fails to capture other facets of intelligence. For example, can a test measure a person's creativity? Under the constraints imposed by this practice, the more correct answers a person gets, the more intelligent that person is. My point is that the characterization of this cognitive faculty in scientific practices depends on the epistemic resources available for its investigation.

In the example above, if an individual's IQ score is significantly above the mean, this may indicate that the individual belongs to a particular category, such as "very superior" as defined by the IQ classification system. The category in question is primarily a label for a construct called intelligence, the appropriate context for which is the domain of specific scientific practices. Intelligence in scientific practices is thus conceptualized in such a way that it can be measured in specific patterns of individuals. The conceptualization has the epistemic role of transforming the cognitive phenomenon (intelligence "in the wild") into a research object (intelligence as the power of appropriate selection), thereby constructing it a measurable target system.

Just as the concept of intelligence as the power of appropriate selection did not exist before extensive research on it was done in modern psychology, all scientific concepts

> (...) were once created and formed in a specific context of intense scientific research, and from there found their way into common language. At their site of their origin they were open, tentative, insatiable, and flexible, while later they appear solidified, stable, either as 'natural' categories or in some cases just as 'facts' (Steinle 2009, 200).

According to Steinle, scientific concepts have the potential to transcend their disciplinary origins and become common ideas in society. In the process, they also become reified as real things. The concept of intelligence as the power of appropriate selection, following the example above, was not intended to be a complete but operational characterization of intelligence. Nor was it intended to be a socially acceptable characterization of intelligence. It was put forward as a means of facilitating the quantitative study of intelligence by heuristically transforming it into a simpler, measurable **target system**. Target systems are not phenomena, nor representations of phenomena; rather, they are entities constructed within scientific practices.

One questionable narrative is the idea that a scientific conceptualization has historically prevailed among rivals because it captures the necessary and sufficient conditions of a phenomenon. The historical dominance of the concept of intelligence in modern societies is not due to its exhaustiveness. The concept of intelligence as the power of appropriate selection is insufficient to account for all forms of intelligence. Its success has to do with its reproduction in scientific and other social practices. Yet, it is unsuitable for measuring social intelligence, for example, the intelligence of scientists attempting to create standardized protocols for measuring intelligence or the seemingly unreflective behavior of animals in symbiotic relationships, such as the remora fish and shark. This is because the tools that enable the application of such a concept (e.g., IQ tests) are inappropriate for studying their intelligence. Therefore, it is implausible that a single scientific concept of intelligence can adequately account for all instances of intelligence.

The mere fact that a concept is not exhaustive does not justify replacing it with a more comprehensive concept. There are other reasons to pursue conceptual change besides exhaustiveness. Indeed, conceptualizations may be useful precisely because of their specificity and non-exhaustiveness. In the context of scientific inquiry, concepts facilitate the discovery of new phenomena or the acquisition of new insights into previously known phenomena when used as operational descriptions. In a paper on scientific definitions of life, Knuuttila and Loettgers (2017b) argue that scientific definitions play roles beyond those of classifying devices (i.e., describing the necessary and sufficient conditions for a

phenomenon). For instance, definitions of life can form the basis of diagnostic tools that help scientists search for biomarkers. In my view, the function of concepts and definitions is to delineate and frame target systems in ways that allow for scientific experimentation and modeling.

The cognitive sciences rely on operational descriptions of cognition to study phenomena. For example, perception cannot be studied unless a scientist makes some assumptions about what it consists of. In PP, perception is modeled as a phenomenon analogous to probabilistic inference. Scientific modeling in this context amounts to projecting the properties of probabilistic inference onto a target cognitive system. In doing so, it renders the target system epistemically accessible to scientists using tools well suited for dealing with probabilistic inference, such as Bayesian statistics. Of course, perception in nature is not inferential, but such a characterization, unlike that of perception as a complex cognitive behavior, is simple (non-exhaustive) and allows scientists to study targets with particular tools. That is, the conceptualization does not specify what the phenomenon is, but how it is understood within a scientific practice.

In summary, conceptualizations of cognition need not be exhaustive. They are epistemic resources that serve the function of characterizing the target systems. In scientific modeling, scientists combine conceptualizations with other epistemic resources, such as mathematical and computational structures, to construct tangible artifacts for investigating the targets. For example, scientific models of perception as probabilistic inference require statistical tools to investigate their targets. These artifacts are used to study target systems whose description is contingent upon the epistemic resources used to construct a model of them. Consider, for example, Ashby's suggestion, introduced earlier in this section, that the available measurement tools determine the scientific understanding of intelligence as the power of appropriate selection.

1.2. Scientific Models in the Cognitive Sciences

Cognitive scientists have different views on the concepts they use to study cognition. Those who consider information processing a scientific analogy and an instrument for knowledge production are instrumentalists. Realists, on the other hand, consider information processing to be a natural phenomenon and scientific concepts to be designators of mental entities, processes, or mechanisms.

In my view, the prevalence of a common narrative in the field partly explains why many cognitive scientists are realists. According to this narrative, cognition is a complex, multifaceted phenomenon, often referred to as "multilevel." Due to its complexity, scientific disciplines can only provide partial explanations. Complete explanations require interdisciplinary efforts. I suspect this narrative is responsible for the view of concepts as designators of mental kinds because it assumes that scientific disciplines using the same terms ("cognition," "perception," "action," etc.) are dealing with the same entities instead of considering them transdisciplinary phenomena, as discussed in Chapter 6. But why should we assume that two or more disciplines are dealing with the same phenomenon just because they use the same terms? If we cannot prove this, then we cannot assume that their partial explanations will converge into a complete one.

Another implication of the above narrative is the assumption that cognitive science is interdisciplinary[4] because it seeks to provide complete explanations of complex, multilevel cognitive phenomena. More specifically, cognitive science is seen as a discipline concerned with bridging disciplinary knowledge to overcome what is known in the field as the 'integration challenge'. According to Bermúdez (2014, xxviii), "this is the challenge of developing a unified framework that makes explicit the relations between the different disciplines on which cognitive science draws and the different levels of organization that it

[4] For many cognitive scientists, interdisciplinarity is one of the hallmarks of their discipline (Bergmann et al. 2017; Dale, Dietrich, and Chemero 2009; see also Cooper 2019). Cognitive scientists acknowledge the one-sidedness of disciplinary approaches to cognition and take for granted that interdisciplinarity is necessary to develop unified and complete explanations. According to this narrative, interdisciplinary research integrates findings from different disciplines at different levels (Thagard 2005). Marr's (2010) levels of analysis complement this view of interdisciplinarity by specifying what disciplinary accounts of cognition should explain. Complete accounts describe cognitive behavior at the computational, algorithmic, and implementation levels. For example, a complete account of perception integrates a) a computational description of perception as the computational task of transforming sensory inputs into perceived objects or scenes (artificial intelligence); b) a description of the algorithms or rules that a perceptual system follows to accomplish the above task (cognitive psychology); and c) an explanation of how the brain physically implements the above process (neuroscience). I think that several problems with this view of interdisciplinarity render it incomplete. First, it assumes that the complex nature of cognition determines the interdisciplinarity of the science studying it. In many cases, however, scientists are simply not interested in complete explanations of cognition, and instead work with simplistic characterizations of cognitive phenomena that can be simulated by computational systems. According to Marr, complete explanations are reductive because they treat cognition as computation. But models of cognition can be non-computational and still contain descriptions. As for the completeness assumption, many explanations in cognitive science are partial, even if they follow Marr's levels. My point is that the outcomes of modeling practices need not be partial representations of complex and multilevel cognitive phenomena, but of target systems performing simpler cognitive tasks.

studies." Overcoming the integration challenge amounts to using partial disciplinary findings as building blocks for developing complete explanations that account for the complexity and multilevel nature of cognition. Interdisciplinarity is thus understood as a scientific strategy for producing complete explanations by integrating partial disciplinary explanations into exhaustive ones. An example of this narrative is the following: cognitive psychology is said to explain perception at the behavioral level, and neuroscience is said to explain the underlying mechanisms responsible for such behavior. Together, they can somehow provide an explanation of this behavior at multiple levels.[5]

I am skeptical of this narrative for several reasons, one of which is that it seems to neglect the dynamics of knowledge production and the intended purposes of research activities. For the most part, cognitive science research is not intended to provide multilevel explanations. Scientific modeling practices in the field may serve other purposes. For example, they may explore the intuitions that scientists have about cognitive phenomena by modeling them into tangible representational media (McClelland 2009). While exploration can be seen as a step toward producing multilevel explanations, it does not necessarily lead to this outcome. The pursuit of completeness is not the only possible epistemic purpose of exploration.

In some cases, interdisciplinary collaboration in the field is unrelated to the production of multilevel explanations. I think a more accurate picture of knowledge transfer in the field can be obtained by looking at how cognitive scientists reuse and repurpose tools, methods, and techniques that have disciplinary origins in new domains. For example, mathematical and computational models serve the purpose of model building without necessarily being constrained by the assumptions of the disciplines in which they were originally used. These are transdisciplinary rather than interdisciplinary tools, since they do not help to provide multilevel explanations, but rather to explore transdisciplinary problems such as cognition, perception, or action. In Chapter 6, I address the transdisciplinary nature of the cognitive sciences.

From this perspective, analogies such as that between cognition and computation can be seen as means for framing transdisciplinary scientific problems. Just as cognition can be

[5] One assumption underlying the narrative of interdisciplinarity as a means of providing multilevel explanations has been hegemonic. This is the assumption that cognition is information processing or a form of computation about representational structures (Bermudez 2014). However, it should not be considered the central theoretical assumption of cognitive science (e.g., as in Thagard 2005) because it no longer plays this role in contemporary cognitive science. Embodied and dynamic systems approaches, two contemporary frameworks in the field, explicitly reject it (Bruin, Newen, and Gallagher 2018; Chemero 2001; 2009). In principle, a science of mind need not model cognition as a computational process. Instead, it can model it as the behavior of a dynamical system interacting with the environment.

seen as information processing, so can many other phenomena that are studied in the sciences for theoretical or practical purposes. In my view, identifying the role of conceptualizations of cognition such as that of information processing can be done by examining their place and function in knowledge production dynamics. An examination of the role of concepts in research practices reveals that, unlike philosophical concepts, scientific concepts are rarely intended to be complete or exhaustive.

Scientific conceptualizations serve the basic function of delineating target systems. Often, phenomena such as perception, decision making, creativity, tool use, empathy, or learning are characterized by reference to various scientific analogies that transform these poorly understood phenomena into more familiar or simpler ones (Bailer-Jones 2009). These analogies also allow for the use of mathematical or computational tools in the study of cognition. As will be shown in Chapter 3, the above strategies lead to the production of scientific knowledge because they make it possible to gain control over target systems by predicting their future behavior or simulating it in computational models. They also allow scientists to formulate testable hypotheses.

As Bailer-Jones (2009) suggests, analogies play a crucial role in the construction of scientific models. This seems to me to be particularly true of scientific models in the cognitive sciences, where views of cognition diverge because of the different analogies used to define it. In some cases, modeling is about constructing artificial systems that simulate behavior, and this is said to be instructive about some natural behavior. Although one might argue that in such cases the study would yield knowledge not about cognition but about the behavior of an artificial system, things are not so simple. For example, a modeler may construct a model of perception to explore her assumptions or ideas about what perception consists of, since their implications may not be fully understood unless they materialized in an external medium that makes them tangible. At least in this sense, constructing a model that simulates perception can be informative by showing which assumptions can be materialized and which cannot.

According to Gelfert (2011, 521), "scientific models, along with computer simulations, are routinely deployed across the spectrum of the natural sciences (and, increasingly, the social sciences as well). In some disciplines, scientific modeling has become the default mode of how to approach scientific questions." Gelfert's description of scientific modeling as a default strategy captures the status of modeling in the cognitive sciences particularly well. I will argue why I think this is the case in Chapter 3. For now, I will assume that scientific modeling is the most common strategy in cognitive research practices. I follow Knuuttila's (2011; 2017; 2021) understanding of scientific modeling as the practice of constructing and

manipulating artifacts with material and conceptual properties that provide epistemic access to various target systems of interest.

The construction of scientific models in the cognitive sciences often involves the use of epistemic resources from various disciplines, including anthropology, computer science, linguistics, neuroscience, psychology, biology, and philosophy. As noted above, this is not necessarily aimed at providing either complete or multilevel explanations. Scientific models can be constructed for a variety of different epistemic purposes, and their success in achieving these purposes depends not primarily on their ability to accurately represent something in the world, but also on factors such as the expertise of the scientists, the intended purposes, skills, and available knowledge. But why do cognitive scientists construct models in the first place? McClelland answers this question as follows:

> [Scientific] models allow the exploration of the implications of ideas that cannot be fully explored by thought alone. As such, they are vehicles for scientific discovery, in much the same way as experiments on human (or other) participants. But the discoveries take a particular form: A system with a particular set of specified properties has another set of properties that arise from those in the specified set as consequences. From observations of this type, we then attempt to draw implications for the nature of human cognition (2009, 16).

Exploring the intuitions, assumptions, and ideas that scientists have about what constitutes cognitive behavior is just one of the myriad possibilities that scientific modeling offers researchers. The point is not only to test whether they are correct. It is also about finding their limits. Unlike philosophical strategies such as thought experiments or conceptual analysis, it has the advantage of embodying ideas in tangible and manipulable epistemic artifacts. Importantly, before scientific models can yield epistemic gains, they must work, so they must first be constructed.

Knuuttila (2011) suggests that the assumption that scientific models are representations of phenomena conceals the multiple ways in which models provide epistemic access to the problems of interest. According to the artifactual approach to modeling (Knuuttila 2017; 2021), scientific models are epistemic artifacts constructed and manipulated by scientists to address open theoretical and practical issues. This approach challenges the view of scientific models as representations in the philosophy of science by treating them as entities that are partially independent of theory and reality (Knuuttila 2011). Scientific models do not have a pre-established representational relationship to their targets, and in many cases, scientists construct them without yet a clear understanding of what those targets are or as a means of exploring them (Knuuttila 2005b).

Knuuttila (2021) argues that scientific models are purposefully constructed as systems of dependencies. This means that models have parts and an organization—that is, patterned interactions of their parts. Models work when the built-in dependencies produce a desired behavior. Models are said to be about target systems. Accordingly, they work when they exhibit behaviors resembling those of their target systems. Importantly, scientists must specify the similarity between a model and a target system (Giere 2010). For instance, scientists may build a deep learning model of decision-making behavior and argue that it does not provide insight into how humans actually make decisions (Fintz, Osadchy, and Hertz 2022). Or they may claim the opposite. In that case, they would need to justify why the model is informative about this behavior in humans. Similarly, PP scientific models exhibit inferential behavior that is sometimes considered analogous to perceptual behavior. However, the claim that PP scientific models capture the mechanisms underlying perceptual behavior requires proper justification.

1.3. Scientific Models as Representational Devices

What makes the construction and manipulation of scientific models an epistemically valuable practice in the cognitive sciences? Do scientific models have a distinctive property that makes them more relevant to the study of cognition than theories, observations, or experiments? Of course, there are no easy answers to these questions. One approach to addressing these questions is to examine how philosophers of science have understood scientific models and determine whether these views are informative for understanding modeling practices in this field.

The vast majority of philosophers of science consider scientific models to be representational devices (for an overview of the concept of scientific representation, see Sánchez-Dorado 2024), that is, human-made objects that have the function of representing, depicting, or mapping some properties, processes, or structure of target systems, be they real, possible, or even purely theoretical entities. De Oliveira (2018) summarizes the understanding of models as representations in the philosophy of science as follows:

> But how is it possible to learn through modeling? It is easy to see how mathematical equations, computer simulations, and robotic agents could help advance our understanding of mathematics, computing, and robotics, respectively. How is it that building and operating models enables scientists to learn something not only about the models themselves, but also about the real-world phenomena scientists are ultimately interested in? In short, why is modeling epistemically valuable? The

> answer seems obvious and intuitive: models can give us knowledge of their targets because they represent those targets. That is, models are related to real-world target phenomena in such a way that they can stand in for those targets in empirical research and that the outcomes of modeling generate knowledge of the phenomena being modeled. It is because they are representations of certain phenomena that mathematical equations, computer simulations and concrete models are viable indirect routes to understanding and explaining those phenomena (...).
>
> (...) Philosophers generally agree that modeling is a legitimate means to knowledge, and they also generally agree that the intuitive answer above is right, i.e., modeling is epistemically valuable because models stand in a special relation to their targets (namely, one of representation).

Importantly, Sanches de Oliveira (2018) seems to conflate the representational function of models with the ontological status of models as representations. Philosophers of science often distinguish between the two (see Godfrey-Smith 2006 and Salis 2021). I do not intend to situate my account of modeling in PP within philosophical discussions of scientific representation nor to support or refute the notion that models in cognitive science represent cognitive phenomena. I agree with de Oliveira's (2018) statement that "representationalism in the philosophy of science is a dead end," and my account of the epistemic value of scientific models in cognitive science is intended to be independent of the notion of scientific representation and the debates surrounding it. I acknowledge that models can perform representational functions and are, in a minimal sense, representations because they are **models of** something or **represent a target system as** something else. In PP, cognition is represented as probabilistic inference. However, this representational practice does not say much about the epistemic value of modeling. Indeed, one could argue that models denote things other than themselves without assuming that they mirror reality.

There are alternative views of scientific modeling in the philosophical literature. Philosophical accounts that focus on model construction (or model building) and manipulation do not always assume that models always have a representational function. Of course, they do not deny that in some cases models can be representations of targets or enable researchers to make inferences about real things in the world (Suárez 2010). However, it seems to me that questions about the epistemic function of modeling can be addressed independently of the discussion of models as representations. In what follows, I use some ideas from the philosophy of scientific modeling and scientific representation insofar as they are helpful in articulating an account of scientific modeling in the cognitive sciences and PP.

Philosophers of science committed to the view of models as representations are careful enough to distinguish between the types of representation that models allow. One that is particularly important for my account is Weisberg's distinction between the types of representation afforded by concrete, mathematical, and computational models:

> Roughly speaking, concrete models are physical objects whose physical properties can potentially stand in representational relationship with real-world phenomena. Mathematical models are abstract structures whose properties can potentially stand in relation to mathematical representations of phenomena. Computational models are sets of procedures that can potentially stand in relation to a computational description of the behavior of a system (Weisberg 2013, 7).

Weisberg is careful enough to emphasize that scientific models can, but do not have to, represent phenomena, pointing out that they can also be representations of mathematical and computational descriptions of phenomena. Models can be non-representational, which means that representing phenomena is not an inherent capacity of models, but rather an achievement of modeling practices (Teller 2001; Knuuttila 2005a, 43).

Although models are not always representations of phenomena, as concrete things they are instantiated in tangible representational formats and media. Models can be, for example, mathematical, computational, diagrammatic, linguistic, and so on. They can be instantiated by various apparatuses, often called representational vehicles. To avoid misunderstandings, whenever I speak of this material dimension of models, I will use the term representational medium ("representational media" in the plural). Importantly, although models can be representational vehicles in this above-mentioned sense, they need not be direct representations of phenomena. As Weisberg points out, mathematical and computational models are representations of mathematical representations of phenomena and computational descriptions, respectively.

Mathematical and computational models are among the most widely used epistemic tools for investigating cognition. For example, agent-based models are computational vehicles that can be used to simulate the behavior of cognitive agents in social groups, such as collective decision making (see Lewandowsky et al. 2019).[6] Interestingly, the concept of an agent in agent-based modeling is not a substantive philosophical or metaphysical one.

[6] In addition to providing information about cognitive behavior and tasks, computational and mathematical models also serve the purposes of theory and model development. Not only can they help scientists explore their intuitions about cognitive behavior, but they can also allow them to compare competing scientific models of cognition (see also Guest and Martin 2021) or serve as cognitive aids. To this end, modelers must specify all of their assumptions in order to run computer simulations (see also Farrell and Lewandowsky 2018, 21).

Agents in modeling practices are understood in a technical sense as rational agents, and their rationality is measured according to a performance measure (see Russell and Norvig, 2022, Chapter 2). By constructing and manipulating these models, ideas about how social agents behave under certain conditions can be explored, and even scientific predictions about collective behavior in real-world scenarios can be made. If taken as representational, these models are representations of computational or mathematical descriptions of phenomena, in other words, their targets are not things in the world but idealized systems. I will explore the creation of target systems in modeling practices in the next section.

1.4. Target Construction

This section articulates my understanding of target construction in the cognitive sciences. To this end, I first introduce the distinction between model descriptions and model systems, and the view of models as epistemic artifacts with conceptual and material dimensions in the philosophy of science. Against the distinction between model descriptions and model systems in the philosophy of science, I argue that scientific models represent target systems, but that these targets are constructed in modeling practices.

In the philosophy of science literature, scientific models are typically viewed as representations of target systems. Within this literature, especially within the "fiction views of models", a distinction is made between direct and indirect views of how target representation is possible:

i. Direct Representation: Models directly represent target systems.
ii. Indirect Representation: Models are composed of model descriptions and model systems, and only the latter represent target systems.

For the fiction views, scientific models are either concrete imaginary systems or rely on some imaginary systems (model systems) to perform a representational function. The representational function of models is understood as the capacity of an imaginary system, either concrete or abstract, to stand in for some properties of a target system. Simply put, a scientific model requires a thing to be imagined, which represents a real entity, process, or mechanism in the world. There are several versions of the fiction view, and I do not intend here to discuss which is the most convincing. The fiction views are relevant here because they provide a plausible account of the components of scientific models, and because I articulate my account of target construction in discussion with them.

A common denominator of the fiction views is the idea that **model descriptions** are a key component of scientific models. Frigg (2022) introduces model descriptions below:

> Models are usually presented to us by way of descriptions, which we earlier called "model descriptions". These descriptions should be understood as props in games of make-believe. This squares with the practice of modelling where model-descriptions often begin with "consider," "assume," or "imagine," which make it explicit that the descriptions to follow are not intended to be descriptions of real-world objects but should be understood as a prescription to imagine particular situations (414).

According to the fiction views, part of the conceptual work of modeling has to do with specifying the properties of a tangible system that will serve as a surrogate for studying a target system. In the philosophy of science, this is known as a model description. Model descriptions specify that certain systems (or classes of systems) have certain properties. Importantly, these descriptions do not exist in scientists' minds but rather "[...] in some representational medium (mathematical formalism, words, pictures)" (Godfrey-Smith 2006, 733). Representational media are used to construct a tangible representational vehicle—a model—which is distinct from an abstract theoretical entity, such as the concepts discussed earlier.

Although models have a tangible dimension, fiction views do not consider the representational vehicle to be an essential part of the model and distinguishes between the imaginary model system, the model description, and the representational vehicle that instantiates the description. For them, only the imaginary model system is the actual scientific model.

Moreover, model descriptions require that certain systems be imagined having certain properties. This system is the model system, not the target system that the model represents. For example, PP models of perception, which will be the subject of the next chapter, prescribe imagining a perception as an inferential and probabilistic task. According to the fiction views, a model system based on this description would correspond to an imaginary system existing in a mathematical or computational medium suitable for solving probabilistic inference tasks (e.g., a Bayesian model). And this imaginary model system, in turn, is a representation of perception in nature.

Contrary to the fiction views, it seems to me that the representational medium and vehicle are core components of scientific models. A model is a description of a system that, of course, is fictional because it is a human creation. However, since the model is not separate from the medium that represents it, the medium is also a core component of the model.

Let me consider the case of PP models. The model system in this case corresponds to an imaginary system that perceives by making predictions about the unobservable causes of its internal and environmental states and improves its predictions by learning from prediction error signals. Despite its fictional character, such a model system is not simply the product of an act of imagination, if by this is meant an act that is not constrained by the representational means used in constructing the model, the available knowledge about perception, and the concepts and assumptions of a theory. According to the fiction view, the model system would be a fictional system that represents perception in nature. The model vehicle, in turn, would require that perception be imagined as an inferential and probabilistic task, and the model system would correspond to an imaginary perceptual system that makes probabilistic inferences.

Knuuttila (2017) questions the strict separation that the fiction views establish between model descriptions and model systems, acknowledging that for these views, what properly constitutes a scientific model is only the model system. For Knuuttila, model descriptions are irreducible parts of models, and imaginary model systems cannot be separated from the representational medium in which they are instantiated. Salis (2021) goes a step further, arguing that model systems are not constituents of scientific models. Instead, she proposes that scientific models are composed of model descriptions and model content, using the analogy of literary fiction to explain both concepts. According to this analogy, model descriptions are like the "props" of a make-believe game, while contents are the "truths" inferred from these props.

In theater and film, the term "prop" refers to the objects that actors use in their performances, such as tables, chairs, or glasses. Model descriptions can be seen as props in make-believe games. They can be either linguistic symbols or mathematical and other forms of representational media that constrain how particular model systems are to be imagined. According to Salis (2021), model descriptions constrain content in much the same way that props constrain an actor's performance. For Salis, it is not the imagined model system that represents reality, but rather the content of the model. This content is derived from the initial assumptions of the model descriptions and the truths inferred from them. In her view, model descriptions impose constraints on the content of the model, not on an imaginary model system.

I agree with Salis that invoking an imaginary system to explain the representational function of scientific models is troublesome. Nevertheless, I remain unconvinced that her concept of content is a superior alternative for accounting for this function. I suggest the alternative concept of **target construction**. This concept points out that whatever content the model description is constraining, it is a content attributed to the target system of the model, and

not of the model itself. In other words, rather than having content, a scientific model allows to attribute content to a target.

In my view, Salis (2021) confuses the content of models with that attributed to a target system. The notion of model description is problematic because it assumes that the conceptual dimension of modeling involves imposing constraints on a model. In contrast, the notion of target construction suggests that these constraints are imposed on the target system. Thus, the conceptual dimension of modeling constrains the way scientists imagine, reason about, and conceptualize target systems. The notion of model content is unwarranted because the content of a model is simply what is attributed to a target system. When constructing models, scientists are not primarily interested in the models themselves (unless the model is the target). Their primary goal is to gain knowledge about or control over target systems. For this reason, I prefer to shift the discussion from model description to target construction.

When scientists talk about modeling cognitive behaviors, processes, or mechanisms, the target systems should be understood as entities whose content emerges during the construction of a scientific model rather than as representations of natural cognitive phenomena. Of course, models have content as well, but the knowledge they provide is about the target systems themselves, as their content is projected onto them. More precisely, target systems are theoretical entities that take on distinctive forms in modeling practices. They are constrained by the conceptual properties of models and the mathematical and computational properties of representational vehicles. Scientific models in cognitive science are based on conceptualizations such as perception as inference and cognition as information processing. The target systems of these models are inferential and information-processing tasks, respectively. I will use the term "target construction" to refer to the process of projecting the properties of models onto target systems created in modeling practices. I will now examine target construction in more detail.

A prominent example of target construction in the cognitive sciences are cognitive behaviors. These are conceptualizations of target systems as simple and observable tasks created in research practices that may in some respects resemble cognitive behavior in nature. Examples of cognitive tasks are the activities of clustering items based on their category membership given to participants in an experiment, any conceptualization of decision making used to model this behavior in an artificial agent, or the target construction of perception as the task of inferring the unobservable causes of sensory signals in PP. These are simple characterizations that make cognition epistemically accessible via various representational vehicles, which are "the material means with which [scientific]

representations are produced and in which they are embodied (such as ink in paper, a digital computer, biological substrata, and so forth)" (Knuuttila 2011).

Some authors consider model descriptions to be the entire scientific models. According to Bailer-Jones, "a [scientific] model [is] an interpretative description of a phenomenon that facilitates access to that phenomenon" (2009, 1). This definition is somewhat incomplete because it ignores that scientific models are tangible entities, not just linguistic structures. Nevertheless, the definition highlights what seems to me to be a central feature of scientific models of cognition, namely that they embed an interpretative description of their targets. In my view, however, this is not a description of the model but of the target system itself.

As discussed earlier, scientific concepts and assumptions are necessary for the study of cognition. Within modeling practices, they are used to characterize cognitive behavior as something else, such as a cognitive task, an inferential task, an information processing task, and so on. These are examples of interpretive descriptions, because in all these cases cognition is analogized to something else. This practice does not produce a description of a model, but a description of the target system, and it is a prerequisite for gaining understanding of or control over a target system.[7]

Target cognitive behaviors, i.e., cognitive tasks constructed in the construction of scientific models, embody these interpretative descriptions. Scientific modeling practices that study these targets do not provide direct knowledge of real cognitive phenomena, but only of these targets. Compared to cognition in nature, targets cognitive systems are idealized, isolated, or simplified characterizations of cognition. For example, consider how an experimental study of visual perception describes this cognitive phenomenon as a visual search task:

> In a visual search task, subjects search for a target item among a number of distracting items. If the time required to complete the search is roughly independent of the number of distractors, the search is said to be parallel; if search time increases linearly with the number of distractors, the search is said to proceed serially (Treisman and Gelade 1980 quoted in Sireteanu and Rettenbach 2000).

The above task description was part of an experimental practice that conceptualized visual perception as a visual search task. Understanding how this task is performed in an experimental setting does not directly explain how, for example, human visual perception

[7] There are conventional and historical reasons that influence the scientific decision to choose one interpretive description over another. As discussed in Chapter 2, modeling frameworks in the cognitive sciences, such as Cognitivism or Connectionism, provide scientists with analogies of minds as computers, machines, or information processing devices, on the basis of which these interpretative descriptions are formulated (see Gigerenzer and Goldstein 1996; Fernandez-Duque and Johnson 1999).

works. Similarly, the target systems of scientific models of cognition are cognitive tasks, i.e., idealized behaviors that certain systems are expected to perform. Thus, task specification is a prerequisite for modeling cognition. If the models exhibit a desired behavior, they are considered cognitive (even if only metaphorically). In addition to task specification, the other two steps in constructing cognitive models are: specifying the parts of a model and describing how the task results from the interaction of these parts.

Consider the well-known Atkinson-Shiffrin model of human memory. In 1968, Atkinson and Shiffrin published a paper presenting a computational model of the human memory system and the processes that regulate it. The parts of this model are a sensory register and a short-term and long-term memory stores. The task of the model is to assign sensory inputs in the sensory register to the stores. To do this, the model follows algorithms in the form of control processes that determine when to allocate resources to one of the two stores. As I will now explain, this model exemplifies some features of models of cognition that assume that cognition is a form of computation, namely

i. Scientific models of cognition are models of cognitive tasks, i.e. observable behaviors emerging from the interactions of parts of a system.
ii. Scientific models of cognition specify algorithms ruling these interactions. In computational modeling, algorithms are step-by-step instructions that guide a system to resolve a problem or perform its task.

Importantly, the characterization of this model's target system—a memory task—is constrained by the target description, which is a computational characterization of the phenomenon. Consequently, the memory task does not directly inform us about how human memory works. This is because modelers transform human memory heuristically into a computational phenomenon. The Atkinson-Shiffrin model deliberately excludes the neural realization, subjective dimensions, and extended nature of human memory (see O'Regan & Nöe, 2004), representing memory as a computational task.

Philosophers sometimes assume that the scientific modeling strategy of omitting irrelevant features of a phenomenon is useful because it allows one to focus on its essential aspects (see Strevens 2011). They assume that this decomposition strategy can be complemented by recomposition to simplify a complex system into its critical components and interactions (Weisberg 2007). Rice (2018) criticizes this approach for failing to consider the pervasive influence of simplifications, distortions, omissions, and idealizations in creating target systems. Rice notes that these elements are epistemically valuable because they isolate critical system components and create holistic distortions of phenomena. In contrast, my account of target construction suggests that models construct, rather than distort, a scientific understanding of reality.

I think the target systems of scientific models of cognition should be understood as entities that are constructed during the modeling process. They are not merely representations of the computational mechanisms or processes that are responsible for cognitive behavior. Furthermore, scientists are not capable of discovering these systems because computational models can represent the natural computations underlying cognition. The conceptualization of cognition as computation or information processing facilitates the construction of a target system. Target systems are not phenomena that models represent. Rather, they are research objects ready for investigation, manipulation, or intervention with the appropriate epistemic resources.

1.5. Conclusions

The goal of this chapter has been to articulate my philosophical understanding of scientific modeling and target construction in the cognitive sciences. The next chapter will introduce the reader to PP from this perspective. This philosophical understanding differs from common views in the cognitive science literature about the kind of knowledge that scientific models of cognition provide. In particular, I challenge the idea that scientific models of cognition are vehicles that represent cognitive entities, mechanisms, or processes in nature. This idea underlies realistic views of computation in the field, which assume that computational models of cognition are able to identify natural forms of computation that are relevant to cognition by virtue of some assumed correspondence between computations in the brain and those performed by the model.

My approach to target systems might be regarded as anti-realist, skeptical, constructivist or instrumentalist, given that I assert that the cognitive sciences are not concerned with natural cognitive phenomena but rather target systems created through scientific modeling practices. However, it is not my intention to suggest that cognitive science is not concerned with cognitive phenomena. Rather, my aim is to provide a clarification regarding the kind of entity that scientific models are about. An examination of scientific modeling in this field may facilitate the development of a critical approach, one that is attentive to the role and limitations of the assumptions guiding this discipline. This approach would also question problematic assumptions, such as the notion that computation is a natural phenomenon. This chapter has not addressed the issues of anti-realism and realism in the cognitive sciences directly. These topics will be discussed in Chapter 4. At this point, I have only presented an argument against the idea that scientific models of cognition provide epistemic access to phenomena by representing them.

I have put forth an alternative perspective, namely that scientific models are epistemic artifacts constructed on the basis of various conceptualizations and assumptions that scientists embed in scientific models. For the purposes of this discussion, a scientific model is understood as a tangible artifact with conceptual and material dimensions. The products of modeling practices are not objective representations of mental processes or behaviors. This is not simply because cognitive processes are not transparent phenomena waiting for some neutral observer to acquire objective knowledge about them. The crucial point is that in constructing a scientific model, a target system—that is, a cognitive task—is created, and this target is constrained by the conceptualizations, assumptions, and representational media and vehicles employed by scientists. My account of target construction will be further justified in Chapter 6.

Ultimately, target systems may be considered by-products of modeling practices that largely depend on the conceptual dimension of modeling. In the philosophy of science, this conceptual dimension is at times referred to as "model description." In this chapter, I have employed the concepts of target construction and target description to argue that the simple and idealized systems are, in fact, target systems that are co-constructed in the construction of a model, rather than mere fictional entities. These epistemic entities can be manipulated and controlled within the modeling practices. In the context of the cognitive sciences, as discussed in further detail in the following chapters, target systems are cognitive tasks constructed using operational descriptions that render cognitive phenomena into solvable problems in modeling practices. Chapter 6 explains how PP and other cognitive science frameworks provide established ways of framing cognitive tasks and standardize the understanding of cognition in this field.

Chapter 2 - Scientific Models in Predictive Processing

This chapter introduces Predictive Processing from a philosophy of science perspective. I argue that PP is a scientific modeling framework composed of various epistemic resources. I focus on its modeling assumptions and scientific concepts and explain the notion of a generative model; the type of scientific model used in this framework.

Section 2.1 puts forth the argument that PP is a modeling framework, that is, a constellation of conceptual and methodological epistemic resources employed by scientists to construct models of cognition. In the preceding chapter, the notion of scientific models of cognition as epistemic artifacts was introduced. These artifacts, defined as human-created cultural objects, facilitate the production of diverse forms of theoretical and practical knowledge. The primary focus of this chapter is on the conceptual dimensions of generative models. The methodological dimension of these models is examined in Chapters 3 and 6.

Section 2.2 is an examination of the assumptions and scientific analogies about the brain and cognition that inform the conceptualization of target systems in generative modeling. These include a view of information processing in the brain as both a top-down and bottom-up phenomenon, a functionalist approach to the brain and cognition, and a view of the brain as a prediction-driven system.

Section 2.3 examines the target systems of generative models, taking the example of scientific models of perception as probabilistic and inferential tasks. A comparison is then presented between PP and phenomenological models of perception. Phenomenological approaches emphasize the experiential aspects of this phenomenon, conceptualizing perception as an act. In contrast, PP conceptualizes perception as an observable task that arises from the interactions of a system's components. The objective of this comparison is to ascertain the extent to which the PP model is deliberately reductionistic and to clarify why this reductionism is a crucial strategy in the scientific investigation of cognition.

Section 2.4 examines the components of generative models, including predictions, prediction errors, estimates, priors, and precision estimates. The function of a generative system is to minimize the discrepancy between its self-generated predictions and the evidence with which they are compared. This is accomplished through the interaction of components within a hierarchical system comprising layers that communicate in both

bottom-up and top-down manners. The upper layers generate predictions about the lower layers, with the lowest level representing sensory signals. Conversely, the lower layers transmit information utilized to correct the predictions generated by the upper layers. A cognitive task is executed effectively when the upper layers accurately predict the information transmitted by the lower levels. If this is not the case, the system attempts to reduce prediction error by generating new predictions. PP systems generate precision estimates to establish a learning rate that weights prediction errors, i.e., determines how much they will alter prior information.

2.1. Predictive Processing as a Scientific Modeling Framework

Cognitive scientists draw upon epistemic resources from established research traditions to construct models. These resources include assumptions, analogies, methods, and concepts. For example, the Atkinson-Shiffrin model of human memory assumes that memory is an information-processing system. Driven by this analogy, the modelers employed the concept of storage to define the short-term and long-term memory stores, which are two components of the model. Other informational concepts, such as input and output, informed the understanding of environmental information and memory as a cognitive task. In constructing the model, which was based on informational concepts, the target system of memory was also constructed by the modelers as an information processing task, namely the transformation of environmental information into short- or long-term memories.

The Atkinson-Shiffrin model serves as an exemplar of a model constructed using the epistemic resources of the cognitivist framework. Target systems of cognitivist models are information processing tasks, that is, as the regular observable outcomes of interactions occurring between the components of an information processing system. In the Atkinson-Shiffrin model, this task is the allocation of information into short- or long-term memory stores. The characterization of a target system, such as memory, can vary significantly depending on the epistemic resources of the framework in question and empirical knowledge about the target. For example, research conducted within the PP framework conceptualizes memory using other concepts, such as episodic and semantic memory (referring to remembering and knowing, respectively), or without isolating memory from perception and attention (Fountas et al. 2022).

Depending on the chosen scientific framework, memory can mean different things. But what exactly is a framework? Why does it matter for understanding the target system of a model?

PP is often defined as a framework, yet the concept has not been discussed in the literature. To grasp the idea of a framework, one must first recognize that there are numerous frameworks in the cognitive sciences. Connectionism, for example, utilizes artificial neural networks as models of cognition. Connectionism is not just a theory of cognition because it comprises other epistemic resources. Frameworks cluster sets of conventionally related resources for investigating theoretical and practical problems. In other words, they comprise all the methodological and conceptual tools used by researchers in a given field working within the same epistemic community.

When using frameworks to investigate target systems, the target systems tend to be conceptualized uniformly. In the above example, memory is defined as an information-processing task within a cognitivist framework that relies on concepts such as information processing. Any other phenomenon within this framework would also be modeled as an information-processing task. A similar situation occurs in research within the PP framework, where cognitive phenomena such as perception, attention, and mental disorders are conceptualized as inferential or predictive tasks—that is, solvable problems. I argue that this conceptualization of cognitive phenomena as solvable tasks is linked to the epistemic resources of a framework. Ultimately, scientists model cognitive phenomena as solvable tasks because they have the epistemic tools to solve them.

Given that the predominant research strategy in cognitive science is scientific modeling, it is appropriate to refer to frameworks within this discipline as "modeling frameworks." The following figure sketches my understanding of PP as a modeling framework:

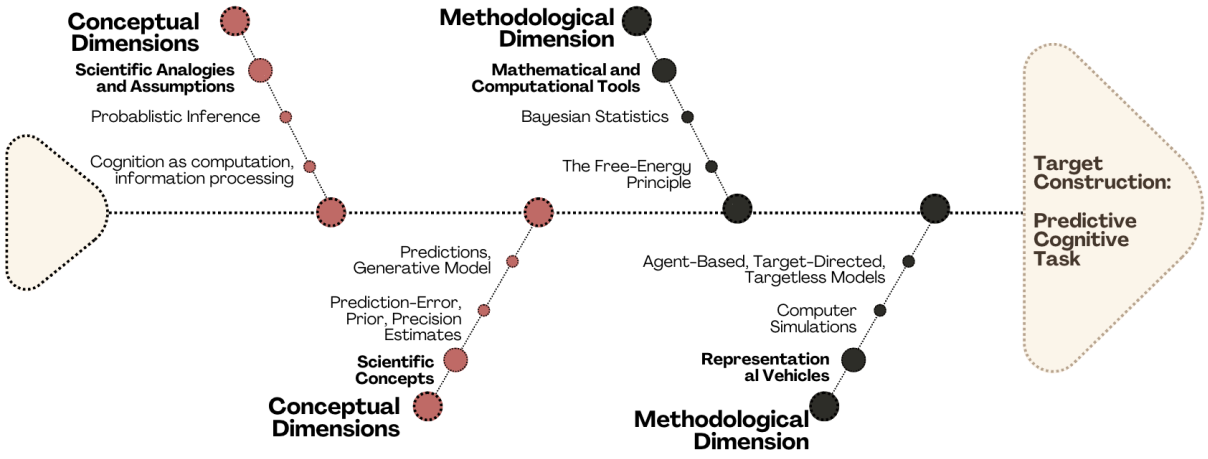


Figure 1: The Predictive Processing Modeling Framework

Modeling frameworks are constellations of epistemic resources used to model target systems. Unlike ontologies, they do not specify the properties and relations of entities of a given kind. Figure 1 shows that the PP framework includes epistemic resources from various fields, including information theory, statistics, computational theory, and artificial intelligence. It is possible to cluster these resources as conceptual and methodological. The latter include for instance mathematical and computational tools. This chapter focuses exclusively on their conceptual dimensions.

The conceptual dimension of a modeling framework comprises scientific concepts, assumptions, and analogies. In the Atkinson-Shiffrin model, for example, the analogy of information processing establishes that memory will be modeled as an information processing system. As we will see, concepts determine the parts of the system, while assumptions specify how they interact with each other. In the same model, concepts determine the parts of the system. Assumptions, such as the idea that cognition is similar to probabilistic inference or that cognition is both bottom-up and top-down, differentiate the PP model from other models, such as connectionist or cognitivist ones.

The concept of a modeling framework helps differentiate research traditions within the cognitive sciences. These traditions are no longer considered theories but rather constellations of conventionally related epistemic resources for model construction. Frameworks provide general strategies for modeling, which explains why many target systems are constructed similarly by modelers working within the same framework. The foundational cognitive science research tradition, Cognitivism, conceptualizes cognition as a representational capacity: the ability to use semantic structures to represent and act on the world. Analogously, Connectionism conceptualizes cognition as parallel distributed information processing. This framework uses artificial neural networks (ANN) as representation vehicles; ANN are epistemic instruments capable of performing this form of information processing. Connectionist models have been used to study various target systems, such as brain activity, decision-making processes, and social interactions. In all these cases, target systems are modeled as parallel distributed information processing tasks.

Despite their relatively brief history, the cognitive sciences have seen the development of many modeling frameworks. In addition to the two mentioned above, there are at least two more: 4E Cognition and Dynamic Systems Theory. I expect to convince the reader that PP is also a modeling framework because it clusters distinctive epistemic resources for scientific modeling. The following is an introduction to PP as a modeling framework.

2.2. Modeling Assumptions in Predictive Processing

Bottom-up and Top-down Information Processing

A main assumption of PP is that information processing in the brain is both top-down and bottom-up. This assumption is contrary to a long-held philosophical assumption, namely that "the mind is a *tabula rasa* or blank sheet until experiences in the form of sensation and reflection provide the basic materials — simple ideas — out of which most of our more complex knowledge is constructed" (Uzgalis 2020). One might argue that scientific approaches that conceptualize perception as a purely bottom-up phenomenon are based on this assumption. For them, perception is a reactive process triggered by sensory stimuli. More precisely, perception consists of transforming sensory stimuli into mental representations of environmental objects.

> Most traditional theories of perception are passive and input-dominated, in the sense that they give prominence to the (bottom-up or feed-forward) flow of information from the sensory periphery and assume that perception, e.g., object perception, is achieved by progressively combining object features of increasing complexity at increasingly higher levels of the visual hierarchy (Pezzulo et al. 2017, 2-3).

In **bottom-up approaches**, perception is characterized as the passive accumulation of sensory stimuli. However, even within these approaches, perception also has an active dimension, as the mind processes sensory input based on its own internal association rules. The *tabula rasa* assumption only implies that environmental information is passively acquired; it does not imply that cognition is passive. PP rejects the *tabula rasa* assumption, as it posits that cognitive agents actively seek and interpret environmental information—or sensory signals—rather than passively receiving it.

Bottom-up approaches often rely on the brain-computer analogy. For them, the brain is an organ that executes discrete instructions step by step, producing observable outputs (see Brette 2022). These outputs are cognitive tasks. They are the observable results of information processing routines within the brain triggered by sensory inputs. In sum, bottom-up approaches depict the brain as a stimulus-driven apparatus that operates exclusively in response to inputs (Cao 2020).

In contrast, **top-down approaches** posit that the experiences, expectations, and specific circumstances of a cognitive system modulate sensory information processing. One example is the "cognitive penetrability of perception" (Cermeño-Aínsa 2020; Vance and Stokes 2017). This phenomenon involves the modulation of sensory signals by mental

states. Consider, for instance, how fear often influences the perception of auditory stimuli and movements. Macpherson illustrates the concept as follows:

> I defined the cognitive penetration of experience as follows: in any case of perception, hold fixed what it is that is perceived (the objects properties and relations seen, heard, touched, and so on), the perceiving conditions (the level of light, shadow, mistiness, for example), the state of the sensory organ (perfect human vision, shortsighted human vision, for example) and the location of one's focus of attention. With those conditions fixed, if it is possible for two subjects (or one subject at different times) to have different perceptual experiences due to the differing content of the states of their cognitive systems, and moreover, there is a semantic or intelligible link between the content of the cognitive states and the content of the perceptual experience, then perceptual experiences are cognitively penetrable. States of the cognitive system include beliefs, judgments and desires, and should likely also be taken to include the concepts that we possess (Macpherson 2017, 7).

Unlike the above example, PP conceptualizes cognition as a top-down process that does not require conscious or high-level mental states. In PP, top-down modulation occurs when perceiving involves actively sensing or sampling the environment. This modulation can occur at the neural level without the subject's awareness.

> An example that can be useful to illustrate this position relates to visual tracking (adapted from Friston et al. 2012; Seth 2014). How are we able to follow the trajectory of an object in the sky, for instance, of a bird, if we have no clue where it is heading? According to Predictive Processing, our capacity to follow a visual cue can be explained by taking the brain to predict, based on previous knowledge, the future position of the bird in the sky. Processing coming from a higher level of the neural hierarchy advances a hypothesis about stimuli received at a lower level. Going back to our example, we could say that based on past information the brain advances a prediction about the position of the bird. The prediction is met by sensory information coming from a lower level of the hierarchy. The brain processes an error signal that encodes the difference between the prediction advanced and sensory input, providing thus feedback about the accuracy of the prediction. The process allows the system to advance better predictions over time (Clavel Vázquez 2020, 664-665)

To recap, top-down approaches assume that the brain is an information-processing system and posit that perception occurs as a result of prior knowledge influencing the processing of sensory signals. PP is a mixed approach that draws on both top-down and bottom-up perspectives. In other words, the brain's processing of information is not unidirectional or deterministic. Rather, it takes on different forms depending on the evolution of the coupling

between the agent and the environment. For example, Sato et al. (2020) demonstrated that elite competitive swimmers have superior sensorimotor abilities in an aquatic environment compared to novices. This phenomenon is associated with cortical reorganization in response to long-term, task-related behaviors.

One of the fundamental assumptions used in the PP framework is that the brain operates as an information-processing system and that information is processed in both top-down and bottom-up ways. Another related assumption is the functionalist view of cognition, which will be discussed below.

A Functionalist Assumption in Cognition and Brain Research

One of the assumptions of PP is that cognition is a task or is best understood as such. I call this the **functionalist assumption**. This assumption informs how brain activity and cognition are modeled within this framework. Under this assumption, cognitive phenomena are modeled as the behaviors of systems. More precisely, they are treated as observable tasks resulting from the patterned interactions of system components. A functionalist approach is not concerned with the material and physical realization of these processes but with the information-processing routines underlying them.

PP's functionalist assumption concerns the conceptualization of cognitive target systems as tasks of probabilistic inference realized by systems with interacting parts. In PP, these systems are known as generative models. These models comprise predictions, prediction errors, prior knowledge, precision estimates, and evidence. In this model, an observable cognitive behavior occurs when the cognitive system produces an accurate prediction. To accomplish this, the system engages in a probabilistic routine. In other words, in light of bottom-up sensory signals, it produces top-down predictions of their most probable cause based on some prior knowledge. The system then compares its predictions with the sensory signal and computes the difference as prediction error. This prediction error then becomes the system's input, which is used to update its prior. The function of a PP system is to minimize prediction error. The following is an example of how this model characterizes cognitive behavior:

Figure 2: A visual example of top-down modulation of perception

Reading left to right; the predictive brain meets the raw sensory stimulations from the central inscription with a strong expectation of the numeral 13. Reading top to bottom, the raw sensory stimulations are met with a different prediction – the letter B. In Bayesian terms, in the context of reading the 12, the 13 hypothesis makes the raw visual data most probable, while in the context of reading the A, the B hypothesis makes the raw visual data most probable (Clark 2015b, 6, reproduced with permission).

The perception of Figure 2 illustrates how prior knowledge, and perception influences sensory information processing. PP describes perception as the task of inferring the causes of sensory signals. Perception is the behavior of a system (the brain) that processes information (top-down anticipations and bottom-up sensory signals) and produces an output (a perceived scene or object). To do so, the brain uses prior knowledge from past experiences to generate top-down anticipations that serve to modulate bottom-up sensory information processing.

Since predictions are based on prior knowledge, a cognitive system must first gather information about their environment. The prior knowledge enables the system to associate sensory signals with their possible causes. Consider Figure 2 once more. Imagine someone who is familiar with numbers but has no experience with letters. This person would only perceive numbers. In this case, a lack of prior knowledge about letters renders the perceptual system blind to them.

Generative Models and the PP View of the Brain

The concept of a generative model applies to both mental and scientific models. I first briefly discuss the scientific model and then address its interpretation as a mental model, a

common assumption in PP framework research. The scientific model describes a system that performs probabilistic inferences due to the patterned interaction of its components. I will explain these components in detail later in the chapter.

In a generative model, prior knowledge is used to generate hypotheses about probable causes of sensory evidence. The most probable hypothesis is then used to make a prediction. This prediction is then compared to the sensory evidence, and a prediction error is computed. This error is then used to update the weights of the hypotheses. The system repeatedly performs this operation until the prediction error becomes insignificant. Referring to Figure 2 again can help explain this model. Assume a cognitive system has equal prior knowledge of numbers and letters; that is, both are equally probable hypotheses of sensory signals. In this case, the selection of a prior is aleatory. Because of this, with prolonged exposure, the cognitive system's perception will alternate between a letter and a number. This alternation occurs because the perception of a letter triggers a mismatch with the prior knowledge of numbers, and vice versa. Since the hypotheses are equally probable, the brain will continually calculate discrepancies between its priors and sensations and trigger error signals in any perception while alternating between the two hypotheses (Clark 2015b).

In my interpretation, generative models are functional scientific models. They describe generic systems that perform tasks thanks to the interactions of their components. The functional structure of generative models is projected onto target systems. Because of this, cognition in PP is modeled as a probabilistic or inferential task. This view contrasts with the majority of philosophical interpretations of PP, which view generative models as mental models. Consider the following:

> The core idea [of Predictive Processing] is that the brain is not a passive receiver of sensory signals but that it predicts sensory inputs (Wiese 2018, 191).
>
> To deal rapidly and fluently with an uncertain and noisy world, brains like ours have become masters of prediction –surfing the waves of noisy and ambiguous sensory stimulation by, in effect, trying to stay just ahead of them (Clark 2016, xiv).
>
> The Brain is a sophisticated hypothesis-testing mechanism, which is constantly involved in minimizing the error of its predictions of the sensory input it receives from the world (Hohwy 2013, 1).
>
> [...] Brains do not sit back and receive information from the world, form truth evaluable representations of it, and only then work out and implement action plans. Instead brains, tirelessly and proactively, are forever trying to look ahead in order to ensure that we have an adequate practical grip on the world in the here and now.

> Focused primarily on action and intervention, their basic work is to make the best possible predictions about what the world is throwing at us. The job of brains is to aid the organisms they inhabit, in ways that are sensitive to the regularities of the situations organisms inhabit (Hutto 2018, 2446).

The counterpart to viewing generative models as mental models is the idea that PP is an explanatory theory. If the brain is a probabilistic system, then probabilistic scientific models account for its functioning. However, I argue that generative models are merely epistemic resources and that their components, such as predictions, prediction error, priors, and precision estimates, do not correspond to any cognitive mechanisms in the brain.

The functional assumption under consideration is arguably situated at the computational level of description, as defined by cognitive scientists (Marr 2010). A computational description specifies what cognitive systems do. In PP, for instance, perception is defined as the process of probabilistically inferring the unobservable causes of sensations. This description is functional because it describes a task and the system components responsible for it. Functional descriptions are complemented by algorithmic descriptions, which specify the rules systems follow to perform tasks. Together, these descriptions define a cognitive system and its components, as well as the patterned interactions that lead to observable outputs. One issue in the PP literature is the indiscriminate use of the term "generative model" to describe four distinct theoretical entities:

i. A computational model.
ii. An algorithmic structure.
iii. The entire scientific model.
iv. A putative mental model.

Generative models are sometimes regarded as the algorithmic structure that specifies the rules governing the interactions of system components in functional models. Strictly speaking, PP imports epistemic resources from other frameworks for this purpose. Bayesian equations are the most popular tool for this purpose, as they are well-suited to problems involving uncertainty. However, they are not the generative models of PP, which are epistemic artifacts composed of computational and algorithmic structures.

There are many interpretations of the notion of a generative model, and mine is just one of them. I consider generative models to be epistemic artifacts rather than mental structures, which distinguishes my interpretation from others'. The extensive use of these artifacts may have led scientists to believe that brains are statistical machines. For example, Friston's interpretation of the brain is significantly influenced by statistical modeling terminology. According to Friston (2019, 1), the brain is best understood as a constructive organ that

formulates explanations or hypotheses about unobservable causes of sensations and evaluates these hypotheses against sensory evidence. Friston et al. (2014, 148) further describe the brain as a phantastic organ:

> this perspective shifts from the Brain as a passive filter of sensations (or an elaborate stimulus-response link) towards a view of the Brain as a statistical organ that generates hypotheses or fantasies tested against sensory evidence. In short, the Brain is now considered a phantastic organ (from Greek *phantastikos*, the ability to create mental images).

The use of high-level epistemological terms, such as "inference," "explanation," or "hypothesis testing," may be appropriate when describing complex forms of cognition, such as scientific cognition or reasoning. However, this terminology may also be a source of misinterpretations when describing elementary forms of cognition such as perception.[8] Furthermore, it may not be a suitable approach for describing virtually any cognitive phenomenon, as suggested by proponents of PP.

Nevertheless, I believe that the use of these concepts in scientific discourse is acceptable, on the condition that it is made clear that they are not intended to be descriptions of the physical processes of cognitive processes in the brain. In other words, I argue that is not that the brain performs probabilistic inferences, but rather that the analogy of probabilistic inference is an instrument in science for investigating cognition and the brain. Following this analogy, the brain is not only as the locus of high-level or abstract reasoning, but a controller or regulator of embodied activity (Clark 1997). As the locus of motor control, the brain commands the rest of the body and the regulatory processes necessary for survival, including homeostasis (Bears, Connors, and Paradiso 2007). Similarly, Fuchs has put forth the argument that the brain serves as a mediator between the cognitive being and the external world, acting as the organ responsible for understanding the world, others, and oneself. In his words,

> The brain is the mediator making the world accessible to us, and the transformer connecting our perceptions and movements. But in isolation, the brain would be just a dead organ. It is only animated in connection with our senses, nerves, and muscles, with the internal organs, our skin, our environment, and in relation to other human beings (2018, xxvii)

[8] For this reason, Zahavi (2018) characterizes PP as an "intellectualist theory of the mind." The concepts of inference and hypothesis testing in PP particularly support this interpretation. However, even among proponents of PP, some challenge these concepts (Bruineberg, Kiverstein, and Rietveld, 2018). As Orlandi (2016, 2018) observes, the analogy of inference can result in an inaccurate or overly intellectualized characterization of perceptual acts.

Fuchs puts forth an embodied view of the brain that differs from PP in that it avoids using high-level epistemological concepts to describe it. However, when it is recognized that these concepts are scientific analogies rather than designators of physical structures, PP can become a model for any cognitive phenomenon. According to both views, perception and action are the two primary cognitive processes orchestrated by the brain. Perception involves anticipating future environmental and internal states, and action involves regulating the body and environment through predictive control. These processes are intertwined: Actions result in desired perceptual states, and perception enables agents to discern how actions transform environmental and internal states.

According to PP, the brain engages in top-down control by making predictions about the body and environment. One example of this control is the selective regulation of bottom-up signals in sensory attenuation, which regulates bodily states (Windt, Harkness, and Lenggenhager 2014). It is important to note that predictions are not mental representations. Although they may entail semantic content, they are unconscious motor sampling routines that orchestrate active sensing of the environment. As exploratory routines, they are involved in controlling environmental and internal states rather than representing them (Schroeder et al. 2010; Brooks and Cullen 2019). As Seth and Friston (2016, 2) observe, predictive control is "geared towards control or regulation of physiological variables rather than accurate representation of some extracranial state-of-affairs."

In this section, I outlined the assumptions that underlie the construction of PP scientific models. These assumptions include a particular interpretation of how the brain processes information, as well as a functionalist characterization of the brain and cognition. In PP, the brain is conceptualized as an organ that controls and regulates the body and indirectly controls the environment. More specifically, the brain is engaged in a constant process of sampling environmental and bodily states to obtain crucial information for perception and action.

2.3. The Predictive Processing Model of Perception

This section examines the PP scientific model of perception to illustrate how this framework conceptualizes cognitive phenomena. Initially, the section addresses how probabilistic inference serves as an analogy in PP to construct target cognitive systems, acting as a generic conceptual blueprint for modeling any cognitive phenomenon. This conceptualization employs a reductionist strategy that breaks down complex cognitive phenomena into simpler probabilistic inference tasks that can be solved using existing

scientific methods. To gain insight into the reductionism of the PP model, it is compared to a phenomenological model of perception. Since the phenomenological model is holistic, it is an appropriate point of comparison with the PP model. This comparison shows that, despite its reductionist nature, the PP model is a valuable epistemic resource in perception research.

Target Systems: Probabilistic Inferential Tasks

In PP, scientific models of perception rely on concepts to define systems and their interactions. This conceptualization uses an analogy to articulate a scientific *representation* of perception *as* the task of probabilistically inferring the unobservable causes of sensations. The target system is the task that the model is supposed to be informative about. The constructed target system is an idealized behavior because it is purposefully distorted. In the PP literature, perception and action are sometimes referred to as perceptual and active inference, respectively. The above idealization consists of projecting the properties probabilistic inferences onto these behaviors.

The probabilistic inference analogy of cognition implies that perception does not provide direct access to the environment. Information is gathered only through active sensing. In the analogy, the causes of sensations are unobservable and must be inferred. Cognitive systems use their current sensations and prior knowledge to make inferences. These inferences are not always accurate because environments are dynamic and because action transforms them.

PP heuristically transforms perception, action, and other cognitive phenomena into problems of probabilistic inference. Concepts such as "prior knowledge," "uncertainty," and "dynamics" belong to the field of information theory. In particular, the term "uncertainty" comes from artificial intelligence, where it is a property of information derived from environments whose states are not fully observable or behave non-deterministically (Russell and Norvig 2022, 403). Cognitive systems operating in such environments must identify patterns to make predictions and gather prior knowledge. However, as these environments are dynamic, prior knowledge must be updated to account for environmental fluctuations.

Cognitive systems engage in probabilistic inference in perception and action to gain knowledge of uncertain and dynamic environments. These systems can perceive when they accurately predict specific environmental or internal patterns that elicit their sensations. Action is defined as intentionally transforming environmental or internal patterns into

predicted, expected states. Both perception and action are possible because cognitive systems engage in predictive routines.

According to Clark, the concept of **prediction** in PP refers to

> [...] the kind of automatically deployed, deeply probabilistic, non-conscious guessing that occurs as part of the complex neural processing routines that underpin perception and action. Predictions, in this latter sense, are something brains do to enable embodied, environmentally situated agents to carry out various tasks (2016, 2).

For Clark, predictions are not conscious guesses, but rather perceptual anticipations and commands for action. If this were not the case, perceivers would be aware of the process of selecting the most probable hypothesis during perception and action. In PP, **sensations** serve as evidence for cognitive systems to compare their predictions about hidden environmental or internal causes that produce them. A generative model produces a set of hypotheses based on prior knowledge and estimates their probability before the evidence is presented. Then, a prediction is generated based on the hypothesis that is most consistent with prior knowledge about the environment and signals. Agents can perceive the hidden causes of sensory signals when the signals are accurately explained. This occurs when there is no discrepancy between the expected signals and the sensory evidence. If there is a discrepancy, the generative model computes it as a prediction error and changes or refines its hypothesis to produce a new, more accurate prediction. Thus, prediction error informs the agent whether to trust the formulated hypotheses.

A generative model's inputs are **prediction error** signals that indicate the discrepancy between predicted and actual evidence. Rather than processing sensory input and generating a prediction, these systems generate predictions and then process prediction errors as new information. If the prediction error signals are relatively low, the system does not alter its predictions. Conversely, if the signals are high, the predictive system updates or changes its predictions. The function of a generative model is to minimize prediction error signals. In doing so, the model learns to make accurate predictions by accumulating knowledge of the environment and its internal states.

In PP, the process of perception is continuous rather than occurring in discrete stages. This is due not only to the fact that generative models calculate discrepancies between observed and predicted outcomes and generate predictions until discrepancies are reduced, but also because environments and sensory signals change over time. To make accurate predictions, a perceptual system must be able to handle environmental variations and track patterns in dynamic environments. Consequently, generative models continually accumulate new

information and form new hypotheses when existing ones are no longer useful for predicting environmental and internal states.

Cognitive systems increase their predictive power in two distinct ways. One approach is to gather new information through perception. The other one involves intervening in the environment to make it predictable. Perception informs action by providing information about how the movements of cognitive agents modify environmental states (Dewhurst 2019). To understand how perception and action reduce the discrepancy between predictions and evidence, consider the example of an individual expecting to see an apple, as discussed by Parr, Pezzulo, and Friston (2022, 25–26):

> She generates a top-down visual prediction (e.g., about seeing something red and not jumping). This visual prediction is compared with a sensation (e.g., something jumping)—and this comparison results in a discrepancy. The person can resolve this discrepancy in two ways. First, she could change her mind about what she is seeing (i.e., a frog) to fit the world, hence resolving the discrepancy. This corresponds to perception. Second, she could foveate the nearest apple tree and see something that looks very much like an apple. This also resolves the initial discrepancy, but in a different way. This entails changing the world—including her direction of gaze and subsequent sensations — to fit what is in her mind, not changing her mind to fit the world. This is the other direction of fit. This is action.

Predictive Processing Models versus Phenomenological Models of Perception

In the following, I compare the PP model of perception with an alternative model developed in phenomenology. This comparison helps to articulate the distinctive scientific strategy involved in conceptualizing the phenomenon as probabilistic inference. The phenomenological model considers factors that are usually excluded from the inference analogy. Does this imply that one model is inherently superior to the other? I would argue that this is not necessarily the case, as both models serve different purposes. In any case, the comparison illustrates what the reductionist heuristic of the PP model involves.

As discussed above, perception in the PP framework is conceptualized as task of probabilistic inference. Hohwy describes this conceptualization as follows:

> States of affairs in the world have effects on the brain– objects and processes in the world are the causes of sensory inputs. The problem of perception is the problem of using the effects –that is, the sensory data that is all the brain has access to–to figure out the causes. It is then a problem of causal inference for the brain, analogous in

> many respects to our everyday reasoning about cause and effect, and to scientific methods of causal inference (Hohwy 2013, 13).

For Hohwy, this conceptualization is not an ontological thesis about perception. Rather, it is a useful scientific analogy for investigating perception. According to him, it is "a basic and useful formulation of the problem of perception" (2013, 13). This analogy is useful because it allows for the application of statistical tools, particularly those derived from Bayesianism to the study of perception. PP's conceptualization of perception differs markedly from our typical experience of it. The idea of perception as a probabilistic inference task does not align with common sense. People do not experience perceptual acts as probabilistic tasks. Although proponents of PP might argue that predictions are unconscious, PP fails to account for the experiential dimension of perception. As a scientific characterization, it is applicable only in scientific contexts. Within these contexts, target systems—that is, probabilistic tasks—are created by projecting the properties of probabilistic inference onto perceptual phenomena. Outside these contexts, perception is not a process of probabilistic inference.

As mentioned earlier in this chapter, one of the fundamental assumptions about perception in PP is that it is not a passive or reactive process. Rather than being merely passive associations of sensory data, perceptual inferences are interpretative acts. In the words of Friston,

> [...] there is a distinction between percepts, which are the products of recognizing the causes of sensory input and sensation per se. Recognition (i.e., inferring causes from sensation) is the inverse of generating sensory data from their causes. It follows that recognition rests on models, learned through experience, of how sensations are caused (2005, 815).

Friston's approach to perception is partially consistent with the phenomenological account presented later in this section because, for both, perception is more than an aggregate of sensations. Perception is a cognitive act because it involves the intelligent apprehension of environmental or internal objects. In PP, perception is an active process because top-down predictions are partially responsible for producing a cognitive system's inputs. Perceptual inferences are temporal processes through which perceived objects are recognized. This recognition occurs as a result of the complex interplay between perceptual anticipation and sensory evidence.

Perception is a transdisciplinary phenomenon that has been the subject of numerous investigations in various disciplines, including philosophy, psychology, neuroscience, artificial intelligence, robotics, and aesthetics. These disciplines investigate distinct facets of perception. It is therefore reasonable to assume that they can, at times, complement

each other. Among these approaches, phenomenology is argued to be non-reductionist, at least with respect to the experiential dimension involved in the phenomenon.

Embree (2010) distinguishes two orientations of phenomenological research. One is philosophical, which Husserl called "transcendental." The other is a subject-oriented approach exemplified by phenomenologies of illness, perception, social cognition, attention, or imagination. The philosophical approach focuses on the structures of experience. The subject-oriented approach shares its area of study with the natural sciences. For instance, while cognitive scientists study social cognition, phenomenologists study intersubjectivity as a transcendental structure of human subjectivity and social cognition.

According to Embree (2010), both orientations entail reflective analysis of experience and have the potential to clarify the nature of the phenomenon under consideration. Similarly, Zahavi (2021) argues that phenomenological inquiry does not expand our empirical knowledge of the subject matter under investigation. Nevertheless, phenomenological research can contribute to scientific advancement by elucidating the nature and foundations of knowledge. Thus, phenomenological research can be advantageous for cognitive scientists investigating topics such as perception, attention, and imagination. This is because phenomenological research reveals aspects of phenomena that are overlooked when they are approached through reductionist heuristics.

According to Embree (2010), phenomenological analyses are reflective, descriptive, and culturally appreciative. These analyses are reflective due to their detailed observation of subjective experiences, including beliefs, values, and desires. They are descriptive in that they provide information about individuals' experiences during intentional acts, including their concepts, emotions, feelings, and interactions with real and fictional entities. Finally, they are culturally appreciative because they explicitly acknowledge that the natural and sociocultural worlds ground these experiences. According to Embree (2010, 2), a typical phenomenological procedure is to begin "from something that is somewhat familiar, reflectively analyzing how it is encountered, and then expressing a richer comprehension of the thing in question."

Husserl's definition of perception is illustrative of a holistic, rather than reductionistic, perspective to this phenomenon. This contrasts with the view of perception as inference. In an appendix to the *Logical Investigations* (2001, 335), Husserl identifies and differentiates four distinct meanings associated with the concept of perception.

i. External perception: "the perception of things, their qualities and relationships, their changes and interactions."

ii. Perception of the self: "the perception that each can have of his own ego and its properties, its states and activities."
iii. Sensuous perception: "what we perceive by the eye and the ear, by smell and taste, in brief, through the organs of sense."
iv. Internal perception: "[it] concerns mainly such 'spiritual' experiences as thinking, feeling and willing, but also everything that we locate, like these, in the interior of our bodies, do not connect with our outward organs."

The four dimensions of perceptual acts are overlooked when perception is viewed through the reductionistic analogy of inference. Perception is intentional, directed toward either physical or mental entities, which can be either internal or external. External and sensuous perceptions are directed toward external entities, while self and internal perceptions are directed toward internal ones. The intentionality of perception makes it an interpretive act. Perception is not merely the aggregation of sensations, but rather a complex phenomenon influenced by contextual factors and experience. As Gallagher and Zahavi (2012, 7) observe, "the perception of my car as my car suggests that perception is informed by previous experience." This interpretive dimension renders it a cognitive act.

Along with contextual and experiential factors, phenomenologists take into account the extendedness and embeddedness of perception. This viewpoint is not necessarily rejected by PP; however, the inference analogy and the model itself are unable to account for these aspects. Gallagher and Zahavi (2012, 8), like many other phenomenologists, propose the hypothesis that much of the semantic work involved in perception is offloaded to the environment. As they state, "the phenomenologist would say that perceptual experience is embedded in contexts that are pragmatic, social, and cultural, and that much of the semantic work (the formation of perceptual meaning) is facilitated by the objects, arrangements, and events that I encounter." Gallagher and Zahavi (2012, 8) illustrate this point with the following example: "Rather than saying that I represent the car as driveable, it might be better to say that – given the design of the car, the shape of my body and its action possibilities, and the state of the environment – the car is driveable and I perceive it as such."

Both the PP and phenomenological models of perception have one common assumption: previous experiences influence perception. However, while the PP provides a generic conceptualization that can be used to model any cognitive phenomenon, phenomenologists emphasize the importance of distinguishing the qualitative aspects of cognitive phenomena. For instance, they differentiate perception from memory or imagination because perception is intentionally oriented toward present objects rather than past or imagined ones. One qualitative aspect of perception is its phenomenal unity. This means perception is a single act: a person cannot perceive and imagine a red car simultaneously.

Furthermore, the object of perception is a phenomenological unit. A red car in the perceptual horizon appears as a distinct object from the surrounding environment (Wiese 2018).

A further aspect of perception is that it is extended in time. It can be argued that the PP and phenomenological models are consistent with one another in this regard. Visual perception enables the observation of objects from partial perceptual appearances. As time progresses, the various perceptual appearances merge into a unified perception. Husserl posits that perception is not merely an apprehension of the object in its present phase. This act is extended in time, encompassing both retentions (the recent past phase of the object) and protentions (anticipations of what will happen with the object). As Gallagher and Zahavi (2014, 85) posit, "perceptual presence is not punctual, it is a field in which now, no-longer-now and not-yet-now are given in a horizonal gestalt. This is what is required if perception of succession and duration is to be possible."

Without delving into the specifics, a protention can be defined as the "implicit and unreflective anticipation of what is about to happen as experience progresses" (Gallagher and Zahavi 2014, 85). It can be argued that this concept is analogous to that of predictions in PP, as both are forms of non-conscious anticipation of experience. If this is the case, then even if the PP model is arguably reductionistic, it does capture certain aspects of perception that are also identified by other models. Both the PP model and phenomenology posit that unreflective anticipations (referred to as predictions and protentions, respectively) play a key role in perception.

Phenomenological and scientific approaches to cognition can be distinguished by their respective research aims. As Gallagher and Zahavi (2012, 9) observe, phenomenologists are not concerned with the "brains behind this experience." In other words, the aim of phenomenological research, or at least a subset thereof, does not entail augmenting our empirical comprehension of conscious experience. In their own words, the aim is not "to develop a naturalistic explanation of consciousness nor to uncover its biological genesis, neurological basis, psychological motivation, or the like" (2012, 9).

Unlike phenomenologists, cognitive scientists seek to gain empirical knowledge about the mind. Modeling and experimentation are two primary research strategies for investigating perceptual phenomena at the behavioral and neural levels. Although conceptual work is necessary for modeling, it is insufficient for increasing empirical knowledge about the mind. Nevertheless, the deliberately reductionist analogy of inference can be useful for modeling brain activity. As discussed in the following chapter, scientific models of cognition in PP provide a means of making cortical activity empirically accessible. To gain such access, they

heuristically transform perception into a simpler, more accessible target system: a task of probabilistic inference.

Phenomenologists and philosophers of mind differentiate their methodologies from those of scientists, asserting that they are not concerned with the underlying mechanisms of cognitive behavior (see Dreyfus 1971). According to Craver (2006, 373–374), mechanistic explanations are

> [...] supposed to do more than merely predict how the target mechanism will behave under a variety of conditions. They are, in addition, supposed to account for all aspects of the phenomenon to be explained by describing how the component entities and activities are organized together such that the phenomenon occurs.

Many cognitive scientists aim to identify the mechanisms underlying cognition. However, it is controversial that the cognitive sciences are primarily concerned with mechanistic explanations. Often, scientists are not interested in the physical realization of cognition, focusing instead on practical rather than theoretical problems. Even within empirically oriented research, scientists often fail to provide a mechanistic explanation. Computational models allow scientists to simulate, explore, and predict the behavior of target systems but are not models of the physical realization of cognition. In light of this, authors such as Rusanen and Lappi (2016) have argued for a more liberal and pluralistic view of explanation in cognitive science that considers non-mechanistic models explanatory.

The PP and the phenomenological models are not simply different because one is mechanistic, and the other is not. In my view, the differentiating factor is the reductionist heuristics involved in the inference analogy. However, this analogy can be epistemically advantageous. The next chapter presents a case study showing how the reductionist strategy renders cognitive phenomena empirically accessible.

2.4. Generative Models

In this section, I explain the concept of a generative model in PP. First, I define generativity and describe the architecture of a generative model. Next, I examine some of their components in further details: sensory signals, prediction errors, predictions, estimates, and precision estimates. I conclude by reflecting on why they are epistemic artifacts.

What are Generative Models?

The concept of generativity is one of the key concepts for understanding present-day science, technology, and society. Currently, there is a rapidly growing proliferation of generative or prediction-driven AI, such as ChatGPT and DeepL, in the social world (see Crawford 2021). Generative AI is a class of computer systems designed to create content, including text, images, diagrams, audio, and code. These systems are trained on large data sets and use algorithms to identify patterns in the sets and generate mock data.

To understand the concept of generativity, consider an unsupervised artificial neural network model that is tasked with identifying a number in an input image. This model receives input images of hand-drawn numbers from one to nine and classifies them. A non-generative model, such as a discriminative model, would break down the images to identify basic patterns that could be used to classify the inputs. Once trained, the discriminative model can classify input images as instances of distributions of patterns corresponding to numbers. Generative models also use patterns, but for a purpose other than classification. These models use patterns to generate mock images of the numbers 1–9 that discriminative models would use as inputs. Generative models refine their mock inputs by comparing them to the inputs. They use the data to refine and improve the model rather than processing the entire dataset each time (Clark 2016, 27).

In science, generative models perform a function similar to that of generative AI in everyday life. Based on prior knowledge, these models generate new data predictions and refine them through a process called error minimization. However, there are important differences between generative AI and generative models. Users of generative AI only care that the outputs are optimal; the process by which a model produces them is irrelevant. In contrast, the process of constructing a generative model to make claims about cognitive systems is relevant. As I have argued in this chapter, such models involve making assumptions about cognitive systems. These assumptions matter because the knowledge that generative models provide about cognition is determined by them.

Realists have argued that scientific generative models are representations of mental models (Gładziejewski 2016; Wiese 2017). Artifactualism, the view I advocate, suggests that the concept of a generative model originates from probability theory and machine learning. In these fields, generative models are "statistical models capable of generating new datasets based on the probability distribution, which is then estimated, thereby generating a distribution similar to the original one" (Ewuzie et al. 2022). Following this view, the properties of these scientific models are projected onto cognitive phenomena.

In my view, the assumption that scientific generative models represent mental models conflates the properties and structure of an epistemic artifact with those of a phenomenon.

For example, Badcock et al. (2019, 1322) argue that "the brain embodies a hierarchical generative model: its physical (internal) states encode a hierarchy of hypotheses about the world that reflects a probabilistic mapping from causes in the environment to observed consequences (e.g., sensory data)." By suggesting that the brain embodies a model, Badcock et al. project a high-level statistical concept onto the target system without carefully considering the epistemological implications of such projection.

The notion of target construction, introduced in the previous chapter, explains the epistemic role of the projection above. It suggests that, in order to study cognition, one must first construct a target system through conceptualizations. The properties of a generative model are projected onto cognitive phenomena to create a target that can be investigated using the model. Under this interpretation, concepts such as process, mechanism, or structure, when applied to cognition, refer not to mental structures, but to the scientific means used to investigate heuristically transformed cognitive problems.

As epistemic artifacts, generative models have conceptual dimensions that are instantiated in a representational medium. Their conceptual dimensions specify a generic system and its behavior as a function of the interactions between its constituent parts. The representational medium is a computational architecture governed by an algorithm. In PP, cognition is conceptualized as a probabilistic inference task. The function of a generative model is to minimize prediction error. To this end, it generates predictions based on prior knowledge, tests them against evidence, computes prediction error, and updates its prior knowledge. The process is iterated until error minimization is non-significant.

Generative models are hierarchical, or multilevel, architectures. Each level contains a set of priors that are used to generate hypotheses about the causes of bottom-up signals at lower levels. Information flows from lower to higher levels in a bottom-up fashion and from higher to lower levels in a top-down fashion. For a generative model to function properly, error signals must remain low at all the levels. Wiese (2017, 717) presents the following diagrammatic representation of a hierarchical generative model:

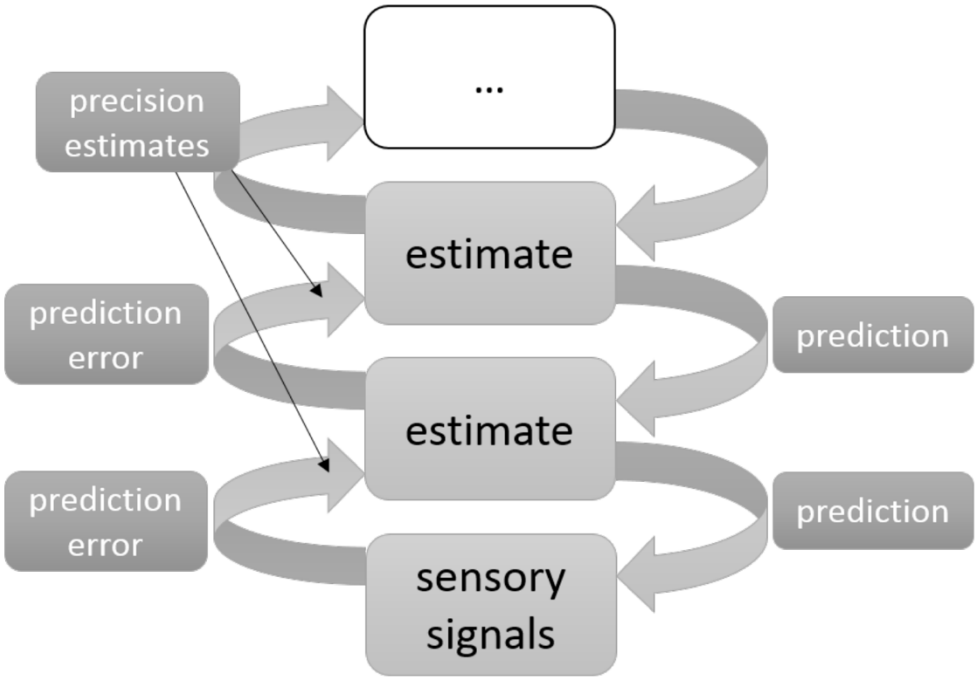


Figure 3: A hierarchical generative model

A generative model estimates which hypothesis of a set of hypotheses is more likely to explain sensory signals, thereby producing a prior probability distribution. The generative model generates a top-down prediction based on the hypothesis that has been identified as the most probable explanation for the sensory signals and compares this prediction with the bottom-up sensory signals. The discrepancy, or prediction error, is then transmitted back to the generative model, which modifies its predictions in light of the newly acquired evidence. Precision estimates update the degree of confidence in predictions based on a posterior probability distribution, which represents the probability of the hypothesis after testing it against the evidence (reproduced with permission).

Prediction Errors and Sensory Signals

In computer modeling, the inputs of a system are the data or information gathered from an environment (Kitcher 1988). In network models, inputs are transmitted from an input node through layers to an output node. In PP, sensory signals, called "observations" or "evidence," are drawn from unobservable environmental or internal states that cause them. Unlike other computer models, prediction errors, rather than these signals, are the inputs of a generative model. Prediction errors are computed from the discrepancy between predictions and evidence. If a prediction does not accurately predict the evidence, the system computes

error signals. These error signals then influence the weights of the hypotheses. As can be seen, error signals are critical to the functioning of a generative model because they enable interaction between its components. In particular, the transmission of error signals gives

> [...] a hint as to the quality of our estimate of the hidden cause. We can update this estimate, compute a new prediction, again compare it with current sensory signals, and thereby (hopefully) minimize the prediction error. Ideally, it does not hurt if our first estimate of the hidden cause is really poor, as by constantly computing predictions and prediction errors and by updating our estimate accordingly, we can become more and more confident that we have found a good representation of the hidden cause (Wiese and Metzinger 2017, 5).

Parr, Pezzulo, and Friston (2022, 4) note that the term "environment" in PP is intentionally generic, encompassing the physical world, the body, and the social environment. The environment has states that are independent of the knowledge of any perceptual system observing it. This knowledge allows the perceptual system to make probabilistic inferences about the hidden causes of observations. Observations are the raw material from which cognitive systems gather environmental information via error-driven learning. Perceptual systems improve their knowledge of the environment by testing their predictions against observations. This process allows them to refine their subjective knowledge and increase or decrease their confidence in it. It is important to note that PP models do not process all potential information derived from observations. Rather, prediction error only informs about what is new or unexpected in the environment.

Estimates, Precision Estimates and Hierarchies

To generate a prediction, a model makes probabilistic inferences about the likelihood of hypotheses, particularly which hypothesis best explains the evidence. These estimates depend on the prior information. First, the estimates are used to assign a probability to each hypothesis. This is known as the **prior probability distribution**. After performing this calculation, the system selects the hypothesis with the highest weight. Then, the prediction is contrasted with the evidence. The distribution determines the degree of confidence in the selected hypothesis, known as **precision weighting**. In the context of generative models, low-weight hypotheses result in low-confidence predictions. In such cases, the generative model is more likely to learn from error signals.

As mentioned above, generative models require the accumulation of information in priors learned from prediction error signals. Generative models are multilevel or hierarchical systems (see Figure 3), which have different sets of priors at each of their levels (Wiese and

Metzinger 2017). Priors at higher levels provide information for making predictions about error signals at lower levels. At the lowest level of the hierarchy, there are no priors, but there are predictions and evidence. The system uses prior information to formulate hypotheses about the lower-level signals. If a prediction fails, resulting in the generation of an error signal, that signal is propagated up the hierarchy. In response, the generative model adjusts or modifies its current estimates. If the model is unable to reduce the prediction error at a given level, the error signals propagate back up the hierarchy.

Final components of generative models are **precision estimates**. The distinction between estimates and precision estimates is technical, and in some cases no distinction is made. Before evidence is processed, estimates assign weights to priors or determine the degree to which a system should trust them. After prediction error is fed back to the system, precision estimates correct the previous estimates. The statistical terms used in this context are prior and posterior probability distributions, respectively. They are used to determine the confidence of a generative model in its priors, predictions, and sensory evidence. In other words, they assess the reliability of the prediction in explaining a given observation. As Wiese (2017, 717) notes, they "can give us a hint as to what extent the error was to be expected, and how strongly it should influence updates of our estimate of [causes]." In essence, they serve as a form of "weighting of prediction error" (Wiese and Metzinger 2017, 9).

When a PP system estimates low precision in priors, it has greater flexibility to modify them and learn from error signals. In other words, low confidence in priors increases the depth of processing of a stimulus. Similarly, Parr et al. (2018, 1) observe that "prior beliefs held with low confidence are rapidly updated to posterior beliefs determined by sensory data," whereas "confident prior beliefs induce an increase in estimated precision when consistent with sensory evidence, but a decrease when they conflict." Yon and Frith (2021, 1027) define precision-weighting as it follows:

> But how much should we rely on our predictions and how much on evidence? This combination problem can also be optimised by precision-weighting: giving relatively more weight to bottom-up evidence or top-down predictions when one is relatively more precise. One upshot of this idea is that we should depend more strongly on our prior beliefs when the sensory world is most ambiguous.

Precision estimates are produced after the generative model computes prediction error. These estimates determine the rate at which the generative model should learn new information and the weight it gives to error signals. In PP, this process functions as a posterior probability distribution (or likelihood), which is an estimate of the probability of a hypothesis after some evidence is presented. By calculating a posterior probability of the estimate after computing the prediction error, the generative model produces a new estimate of the

accuracy of the prediction in explaining the evidence (how much the system should trust prediction error).

Finally, a generative model has a hierarchical architecture. For our purposes, this means that the components and interactions of the top level are replicated at various other levels. Each level contains priors, makes predictions, and computes prediction errors. According to Clark (2016), priors located at lower levels of the hierarchy are the most accurate for explaining incoming sensory signals. In contrast, hyperpriors—priors located at higher levels of the hierarchy—are more abstract and general and are directed at basic environmental or internal structures. Due to their generality, they are the least informative for explaining current sensory signals. Clark suggests that generative models are hierarchical because they track hierarchies in internal and environmental structures from their most concrete to their most abstract properties. Hyperpriors differ qualitatively from other priors because they impose general constraints with low informational value with respect to observations.

Swanson (2016) has considered Kantian transcendental forms of space and time as hyperpriors. According to Kant, our sensibility a priori structures sensory information into these forms. In PP, these forms would be hyperpriors that impose fundamental constraints on the processing of sensory information. My account of generative models as scientific artifacts undermines these interpretations by questioning the correspondence between the hierarchical structure of a scientific model and the hierarchical structure of a mental structure, simply on the grounds that there is no argument for assuming a hierarchical structure in the latter.

Generative Models as Epistemic Artifacts

In this chapter, I have argued that generative models are epistemic artifacts used to investigate cognitive problems once they have been heuristically framed as tasks of probabilistic inference. I also described the components of generative models and their interactions. The purpose of this analysis was to provide a detailed description of the model used in PP to characterize various cognitive tasks. I paid particular attention to PP models of perception, showing how this behavior is conceptualized as an inference task realized by a system that minimizes prediction error. A perceptual task is considered successful if the generative model can maintain low error signals at all levels. In other words, maintaining low error signals is equivalent to generating accurate predictions. In such a case, the generative model is said to be performing a perceptual task. If this is not the case, the error signals are fed back to the model, which responds by weighting the hypotheses before generating a new

prediction. This process is iterative and only ends when the error signals at all system levels are no longer significant.

I have argued that generative models in PP are primarily scientific models and that it is an open question whether they target mental models or structures. I will revisit this issue in Chapter 4. For now, I have merely suggested that those who assume generative models map mental models seem to conflate the properties projected onto a target system with the properties of a phenomenon. This conflation goes hand in hand with the assumption that scientific generative models are representations of cognitive and brain processes.

If generative models are epistemic artifacts, how do they provide an understanding of cognitive phenomena? One answer is that they enable the formulation of mechanistic explanations (Gładziejewski 2019; Badcock, Friston, and Ramstead 2019). However, this possibility is open to question, as generative models are neutral with respect to the material realization of the tasks they model. They specify information processing routines, not mental mechanisms. However, if mechanisms are understood as theoretical resources, then there is no problem in saying that generative models represent mechanisms.

An example of the use of generative models in non-mechanistic ways is the view of the visual system as a hierarchical system in which higher levels influence information processing at lower levels. According to neuroscientists, sensory information is distributed from the retina to areas responsible for detecting specific features. Specifically, the dorsal stream detects information related to motor function, while the ventral stream detects information related to objects and faces (Bears, Connors, and Paradiso 2007). In this context, generative models could serve as scientific models of communication between different cortical areas (Urgen and Saygin 2020). It is also important to note that other scientific models could serve the same purpose. For example, parallel processing models have been used to study the function of specialized areas (Bears, Connors, and Paradiso,2007). However, generative models are not mechanistic even in these cases, as they are merely functional models.

In Chapter 4, I further support my view that generative models are not scientific representations of mental models or structures. For now, I just want to explain why I consider them to be ontology-neutral epistemic artifacts and functional models that are not concerned with the material realization of cognition. I agree with Cao's (2020, 535) perspective on generative models. She argues that

> the models by themselves are not committed to any particular physical neural architecture, or particular patterns of firing by particular cell populations, characterized physiologically. Rather, [...] they are committed to claims about

> information-processing – claims about functional units and what they encode, compute, and signal to each other.

Cao suggests that generative models can only be used to make empirical claims when additional assumptions and knowledge are involved in constructing and manipulating them. In that case, scientists and philosophers must be careful not to confuse the properties of a generic target system with those of natural or physical systems. Generative models consist of several elements, such as predictions, priors, error signals, estimates, and precision estimates, as well as their interactions. Again, these components do not correspond to analogous components of physical systems.

Finally, I will briefly discuss Tee's account of generative models as epistemic artifacts. According to Tee (2020, 2), rather than representing, generative models "new models on the basis of existing ones." Examples of new models include new data sets, text, or images generated from models created using large data sets. Generative models can produce homogeneous or heterogeneous new models. Models that generate homogeneous models reproduce existing patterns, while models that generate heterogeneous models produce variations of those patterns.

In Tee's analysis, generativity is a property of models that abstract irrelevant features and add idealizations to their target system, thereby creating a new model. In Tee's words,

> What counts as a generative model is determined by the scientist's use of the model. [...] The process of abstraction and idealization which confers the generative capacity on a generative model is a two-staged process that is not a characteristic of non-generative models: (1) the omission of the irrelevant feature of a target system is carried out in the direction of producing, rather than representing, a new model on the basis of the existing model; (2) new idealized model elements which are relevant to the generated model are added. These new model elements are relevant to the objective of the model construction and transform a generative model into a new model (2020, 17).

According to Tee, the generativity of scientific models is related to the processes of abstraction and idealization. These processes occur during the construction of scientific models and cannot be assumed to be inherent properties of entities beyond the modeling practice. This is particularly important in the context of PP models since the cognitive systems being modeled do not make predictions due to a capacity for abstraction and idealization.

However, it is possible that abstraction and idealization are projected onto a target system because there is no constraint on how target systems must be conceptualized. In the PP

literature, for instance, cognitive agents have been conceptualized as generative systems capable of modeling themselves and their environment. Similarly, Clark argues that generative models serve to conceptualize cognition at several levels, including the sociocultural level:

> [...] Humans do not merely sample some natural environment. We also structure that environment by building material artifacts (from homes to highways), creating cultural practices and institutions, and trading in all manner of symbolic and notational props, aids, and scaffoldings. Some of our practices and institutions are also designed to train us to sample our human-built environment more effectively – examples would include sports practice, training in the use of specific tools and software, learning to speedread, and many, many more. Finally, some of our technological infrastructure is now self-altering in ways that are designed to reduce the load on the predictive agent, learning from our past behaviors and searches so as to serve up the right options at the right time. In all these ways, and at all these interacting scales of space and time, we build and selectively sample the very worlds that —in iterated bouts of statistically-sensitive interaction— install the generative models that we bring to bear upon them (2015b, 20).

Although the cognitive sciences' ultimate goal may be to find a model that can explain cognition at different levels — from embodiment to complex forms of social cognition — generative models only provide a generic conceptualization of cognition. They are not mechanistic models that explain cognition. Their generativity is not a property of reality. Since they are used to investigate targets constructed in the modeling process, it is impossible to empirically prove that these targets map the structure of cognitive phenomena. I further justify this ontological agnosticism in Chapter 4.

In my view, the capacity of generative models to provide knowledge about the brain and cognition depends on how they are constructed and manipulated, not on their ability to represent brain and cognitive structures. However, if generative models do not represent cognitive phenomena, how can they provide scientific knowledge? This will be the topic of the next chapter.

Chapter 3 - Model Construction in Predictive Processing

This chapter examines the epistemic gains associated with modeling cognition as probabilistic inference using generative models. Since my argument is that these gains are related to how scientists construct and manipulate models, I first explain what it means to construct a scientific model of cognition. Then, I present two case studies of model construction and manipulation in Predictive Processing. The first case study is the Action Observation Network (Urgen and Saygin 2020), a model targeting the neural system responsible for action perception. This model enables scientists to formulate testable hypotheses and predictions. The second case study concerns the agent-environment coupled system model (Bruineberg et al. 2018), an agent-based model of niche construction. It is argued that this model is useful for theory development, particularly in creating new modeling techniques and applying existing ones to new domains.

The chapter consists of four sections. Section 3.1 introduces the philosophy of science perspective on model construction. Section 3.2 presents the Saygin Lab's research on action perception and the construction of the AON model. Section 3.3 discusses targetless modeling and the construction of the AECS model. Section 3.4 addresses the problem of the epistemic gain of modeling in the PP framework.

3.1. Model Construction

Model Construction in the Philosophy of Science

The topic of scientific modeling has gained significant traction within the field of philosophy of science. This is evidenced by the considerable amount of philosophical literature published on the topic in recent years, including works by Bailer-Jones (2009), Frigg (2010, 2022), Gelfert (2016), Godfrey-Smith (2006), Morgan and Morrison (1999), and Weisberg (2013). As this work does not concern itself with the philosophy of scientific modeling *per se*, but rather with a specific scientific model employed in the cognitive sciences, an exhaustive account of the philosophy of scientific modeling will not be provided here. The

focus will be on the idea that the construction and manipulation of models constitute two of the primary strategies employed by scientists to gain knowledge.

A central topic of debate in philosophy of science is how scientific models provide knowledge of their target systems. The prevailing view is that they provide knowledge of their targets by representing them. Knuuttila (2011) refers to this view as the representationalist approach to modeling. According to this approach, models deliver knowledge of their targets because they are either similar to them (Giere 2010) or because they are structural representations of them (French and Ladyman 1999). Some scholars have expressed skepticism about this approach, arguing that it is too limiting in understanding the diverse epistemic functions of models. They argue that focusing too much on representational capacities may lead to an oversimplification of the diverse roles of models in scientific inquiry (Knuuttila 2005b; 2011).

Knuuttila (2017, 2021) developed an account of models as epistemic artifacts that acknowledges models' representational functions. The novel aspect of the artifactual approach is its contention that models do not provide epistemic access to reality by virtue of an inherent capacity to represent it. Models provide epistemic access because scientists construct and manipulate them to address open questions, explore unknown phenomena, and solve practical issues. Models are flexible artifacts because scientists can construct, reuse, and repurpose them in different ways depending on the specific problem, available tools, and the modeler's expertise.

According to Knuuttila (2011, 2021), constructing a model involves creating a tangible system of dependencies, or an artifact whose internal functioning allows it to be used for epistemic purposes, such as answering specific scientific problems. These purposes may include answering questions about the natural world, life in general, and cognition. A model's epistemic possibilities do not depend on its ability to represent phenomena; rather, they depend on whether its internal functioning allows scientists to manipulate it to address research problems.

There are numerous methods for constructing scientific models that draw from a wide array of epistemic resources. These resources can include computational and mathematical tools, experimental protocols, theories, models, scientific concepts, and assumptions. As cultural products, the choice of models and resources depends on the evolving history of their reproduction. As mentioned in the previous chapter, generative models are currently popular for studying cognition, though this has not always been the case. These models have become scientific artifacts in a culture that increasingly reproduces and repurposes them in both science and society (see Martínez 2010).

Like many other models, generative models are instantiated in representational vehicles or media. A representational vehicle does not represent a phenomenon; rather, it represents a conceptualization in a tangible medium. For example, an equation can be instantiated in a tangible medium by constructing agent-based models. Weisberg distinguishes between the types of representation that models facilitate. Mathematical and computational models, such as generative models, are not structural representations that share relevant similarities with phenomena. Rather, they are representations of computational or mathematical descriptions of phenomena (Weisberg 2013). In my view, a computational or mathematical description of a phenomenon is a projection onto a target, not a representation of the phenomenon itself.

For Weisberg (2013), constructing models is a crucial step in scientific modeling. This is because scientists must first construct models before studying target systems. Weisberg (2013) identifies the following categories of target systems (examples are mine):

i. observable phenomena (e.g., the neural correlates of action observation).
ii. hypothetical systems (e.g., a neural mechanism that explains action observation).
iii. generic targets (e.g., evolution or learning).
iv. non-existent targets (e.g., an agent-environment coupled system).

Models of generic targets (e.g., the Atkinson-Shiffrin memory model; see Chapter 2.1) are not models of a physical system, but rather, they are models of behavior projected onto a generic system. Target-oriented modeling may encompass generic targets such as memory, visual perception, and sensorimotor coordination. Target systems must be conceptualized before they can be investigated. As discussed in this chapter, targetless modeling is a strategy that aims to conceptualize targets to make them amenable to scientific inquiry (see also Weisberg 2013, Chapter 7).

Model Construction in the Cognitive Sciences

According to Weisberg (2013), scientific models consist of a description and a structure. A model description specifies its function and components. However, as I understand it, modeling involves more than describing the model; it also involves conceptualizing a target system. In the context of the cognitive sciences, established frameworks provide scientists with specific conceptualizations for constructing cognitive target systems. For instance, the PP framework conceptualizes cognition as a probabilistic inference task. By transforming cognitive phenomena into tasks, these frameworks make them amenable to investigation using mathematical and computational tools, as well as other epistemic resources.

How does the practice of constructing scientific models to study cognition work? In my view, **target construction** can explain this practice. Often, modeling involves exploring unknown target systems or creating a scientific conceptualization of a target system that can be used for future research, as shown in section 3.3. This is arguably a strategy of analogical reasoning because scientists use familiar tools to explore unfamiliar targets (Bailer-Jones 2009). This practice involves attributing the properties of the known entity, i.e., the model, to the unknown entity, i.e., the target system.

The construction of perception models exemplifies this practice. Scientists construct these models using various epistemic resources, such as computer simulations, cortical data, concepts, and data derived from first-person experience. On the other hand, scientists construct computational, mechanistic or physiological, and qualitative models of perception for radically different purposes. These endeavors include constructing artificial perceptual systems, predicting the behavior of cortical networks in perceptual tasks, and describing the phenomenology of perception. The process begins with the conceptualization of a target, which is then instantiated in a representational medium.[9]

A constructed target does not need to capture the complexity of perception. Rather, it must be adapted to the specific problem into consideration, as illustrated by the definitions of perception in handbooks. For instance, a cognitive psychology handbook defines visual perception as "the processes by which a perceived, remembered, and thought-about world is brought into being from as unpromising a beginning as the retinal patterns" (Neisser 2014, 4). The description of visual perception in cognitive psychology deliberately excludes various elements of the phenomenon. It frames visual perception as a behavioral outcome of the processing of sensory inputs (retinal patterns) into outputs (perceived objects and scenes).

When modeling the physiology of visual perception, scientists often attribute the characteristics of mechanical information processing systems to perceptual systems. Mechanistic models of the visual system explain the physical realization of its function (from the retina to the visual cortex) by identifying how information is transmitted in different

[9] The representational medium of a generative model does not represent a physical cognitive system. Rather, this medium is an equation used to solve probabilistic inference tasks, also known as Bayes' rule. This rule is explained in Chapter 6. For now, it is sufficient to say that it is used to calculate the conditional probability of a hypothesis, i.e., the probability that the hypothesis is true given some evidence. This rule is fundamental to numerous probabilistic models used in cognitive science and is foundational to the view of the brain as an inferential or predictive system (Colombo and Seriès 2012). Whether Bayesian models are good at making predictions about the brain due to representational power is controversial. Vincent (2014, 6) points out that Bayesian approaches "are not always serious explanations of human behavior in all contexts. Instead, they act as important baselines."

perceptual tasks (Craver 2007). Unlike mechanistic models, informational processing or functional models do not require an explanation of the physical basis of the system under consideration. Instead, they conceptualize perception as an informational task and specify a system that realizes it. Functional models define algorithms or rules that orchestrate information processing. This chapter presents the argument that generative models are non-mechanistic. These models describe perceptual phenomena as inference tasks, identify their components (e.g., sensations, predictions, priors, and error signals), and specify the structure that organizes their interactions. However, they do not model the physical mechanisms underlying these processes.

Morrison and Morgan (1999) treat scientific models as autonomous systems with respect to theory and reality. Yet, they are related to both. As the authors state, "A model is independent of the thing it operates on, yet it connects with it in some way" (1999, 11). Models are not mere applications of theory to phenomena; they are independent of reality because their internal functioning does not merely replicate the processes observed in the real world. If models are to be considered autonomous, then how do they provide epistemic access to reality?

Godfrey-Smith (2006) proposes that epistemic access is granted through indirect representation of phenomena. While this may be the case with models oriented toward empirical targets, there are models that serve purposes such as theory development and that do not indirectly represent anything in the world. Furthermore, models can be deliberately distorted with respect to phenomena to gain access to behaviors and patterns of interest, even if these behaviors and patterns do not occur in nature. Also, how-possibly models can "depict things as they could be, but do not necessarily capture actual states of affairs in the real world" (Koskinen 2017, 493).

The following sections will examine the epistemological value of using generative models in science by analyzing two case studies. The first is the Action Observation Network, a generative model representing the communication of brain areas involved in action perception. This section will demonstrate how constructing this target-directed model facilitates formulating testable hypotheses and predictions. The goal is to show that generative models are neither representational nor mechanistic. They serve to address questions by transforming cognition in a task of probabilistic inference solved by the brain. The mathematical apparatus of generative models renders the target empirically accessible by enabling scientific predictions about brain activity data under the assumption that the brain minimizes prediction error. I argue that making scientific predictions about brain activity data is not the same as representing the structure of the brain.

The second case study is about the Agent-Environment Coupled System (AECS), a targetless model of niche construction. The construction of the AECS model shows that niche construction can be modeled as an agent's task of minimizing surprise, or prediction error, under specific assumptions. The epistemic benefit of constructing this targetless model is that it creates new models that can subsequently be employed in target-directed modeling. In other words, models of this nature are instrumental in theory development, broadly conceived as creating new conceptualizations and tools to explore old and new problems and domains.

3.2. The Action Observation Network Model

Biomotion Perception and the Action Perception System

The following section will present the research of neuroscientist Ayse P. Saygin and her collaborators at the Cognitive Neuroscience and Neuropsychology Laboratory at the University of California, San Diego (Saygin Lab). These neuroscientists conducted extensive research over several years on biomotion detection, action understanding, and the uncanny valley phenomenon, drawing on epistemic resources from PP. This section aims to demonstrate how these resources facilitated the interpretation of experimental results and neuroimaging data, the construction of models of hypothetical brain mechanisms, and the formulation of testable scientific predictions and hypotheses.

The researchers at the Saygin Lab defined biomotion detection (BMD) as the ability to perceive or identify the motion of biological entities. As they posit, this capacity enables animals to undertake a range of "tasks of biological significance, from hunting prey and avoiding predators, to imitation, social cognition, and theory of mind"(Blake and Shiffar 2007, as cited in Gilaie-Dotan et al. 2013, 457). Saygin (2007) conducted two experiments to identify the neurological correlates of BMD ability in a sample of 60 stroke patients with unilateral lesions. The researchers discovered that lesions in the Superior Temporal and Premotor Frontal areas negatively affect the performance of BMD ability. Gilaie-Dotan et al. (2012) identified a correlation between inter-individual variability in BMD ability and gray matter volume in the Posterior Superior Temporal Sulcus (pSTS) and ventral Premotor Cortex (vPMC). In a subsequent experimental study, Gilaie-Dotan et al. (2013) analyzed this variability using voxel-based morphometry, a neuroimaging technique used to examine focal differences in brain areas. This analysis was aimed at validating the hypothesis that "the

neuroanatomical structure of the pSTS and vPMC would predict individual differences in biological motion detection" (Gilaie-Dotan et al. 2013, 458).

In another study, Miller and Saygin (2013) posited that a singular mechanism underlies both BMD ability and action understanding, the latter of which is defined as the capacity to perceive intentions and comprehend intentional movements. The Action Perception System (APS), subsequently renamed the Action Observation Network (AON), serves as a model for a hypothetical neural mechanism responsible for action perception. The APS is a network model of brain regions that are active during tasks involving action understanding or BMD abilities (Saygin et al. 2012). The aforementioned system includes "the brain areas most commonly discussed in relation to action perception (i.e., lateral temporal, inferior frontal/ventral premotor, and anterior intraparietal cortex)," and is known as the Action Perception System (APS).

Saygin et al. (2012) conducted a study to determine whether the APS exhibits selectivity for human biological motion. This was based on previous findings suggesting that the same brain areas are involved in both biomotion detection and action understanding. In the experiment, participants were presented with three different pairs of images and motion. The pairs consisted of images that resembled human or robotic appearances and videos that resembled human or mechanical movements. The initial pair depicted a human appearance and motion (human); the subsequent pair exhibited a mechanical appearance and motion (robot); the final pair displayed a human appearance and a mechanical motion (android). The researchers classified the perception of human motion accompanied by a robot appearance as a **mismatch condition**, given the typical association between human motion and human appearance. Saygin et al. (2012) predicted that the perception of human appearance and mechanical motion would elicit anomalous activity in the APS. They confirmed the prediction because they found that the mismatch condition elicited distinctive responses in the Bilateral Anterior Temporal Sulcus.

Saygin et al. (2012) interpreted the above results in line with the PP assumption that the cerebral cortex generates top-down expectations of sensory experience and calculates sensory input based on the discrepancy between these expectations and actual sensory input. The distinctive responses elicited by the mismatch condition were treated by them as empirical evidence for prediction error signals. Moreover, they posited that the activity observed in the anterior temporal sulcus during the mismatch condition provides compelling evidence for the existence of predictive mechanisms in the human brain.

In addition, Saygin et al. (2012) proposed that PP could account for cognitive phenomena at the behavioral level. They suggested that the distinctive elicited by the mismatch condition could explain the phenomenon of the "uncanny valley," which they described as follows:

> It may seem like a good idea to make artificial agents look as human-like as possible, especially if they will be used in social settings. However, we soon encounter the 'uncanny valley': as an agent's appearance is made more human-like, people's disposition toward it becomes more positive, until a point at which increasing human-likeness leads to the agent being considered strange, unfamiliar and disconcerting (2012, 414).

The phenomenon of the uncanny valley can be observed in emotional responses to situations where there is a perceived mismatch between what is expected and what is actually experienced. Saygin et al. (2012) suggested that the uncanny valley phenomenon can be explained by a PP mechanism. This mechanism involves the generation of prediction error signals in response to a discrepancy between perceptual expectations and sensory input. The idea is that the mismatch between the observation of human motion and the expectation of mechanical motion triggers prediction error signals.

In a subsequent study, Urgen, Kutas, and Saygin (2018, 82) present an explanation of the uncanny valley phenomenon that appeals to a "violation of expectations in neural computing when the brain encounters almost-but-not-quite-human agents." If this is the case, then exposure to a human-like form will result in the prediction of human-like behavior, including human-like movement. Following this, the uncanny valley arises when these expectations are not fulfilled, such as when confronted with agents that possess human-like forms but exhibit non-human-like movements. In a study of action perception in humans, robots, and androids, Urgen, Kutas, and Saygin (2018) employed electroencephalography and event-related brain potential techniques. In contrast with the general prediction of distinctive activations of the APS elicited by the mismatch condition (Saygin et al., 2012), the scientific prediction in this case was more specific. The researchers predicted an increase in the amplitude of the N400 in the mismatch condition. The N400 effect is defined as

> a remarkable biomarker of human information processing (...) [It] is the human brain's response to any meaningful stimulus. It is a negative-going event-related brain potential, which peaks around 400ms after stimulus onset and is maximal in fronto-central regions of the human scalp to pictorial stimuli (Kutas and Federmeier 2011, as cited in Urgen, Kutas, and Saygin 2018, 182-183).

Urgen, Kutas, and Saygin (2018, 181) hypothesized that "the appearance of the agent would provide a context to predict her movement, and accordingly, the perception of the realistic robot would elicit an N400 effect indicating the violation of predictions, whereas the human and the mechanical robot would not." The results demonstrated that, despite the elicitation of the N400 effect among all participants,

> the amplitude of N400 (measured with area under curve between 370 and 600 ms averaged across all frontal sites) in the dynamic form was significantly greater than the static form for Android ($t(18) = 2.401$, $p < 0.05$). On the other hand, the static and dynamic forms did not differ either for Robot ($t(18) = 0.388$) or for Human ($t(18) = -0.346$) (2018, 185).

The results validated the scientific prediction of an increase in the amplitude of the N400 effect in the mismatch condition, as postulated in PP. Urgen, Kutas, and Saygin (2018) interpreted their results as preliminary evidence that "the mechanisms underlying the perception of other individuals in our environment are predictive in nature." However, does their successful prediction provide evidence for the existence of predictive mechanisms in the brain? I believe that the evidence does not support this conclusion. The reason is that a scientific prediction of an increase in the amplitude of the N400 effect does not provide insight into the underlying mental mechanisms nor constitute a causal explanation of APS behavior. Such a scientific prediction does not specify the components of a physical mechanism, nor does it elucidate the way these components interact to give rise to the observed patterns of behavior.

The scientists in the Saygin Lab believed that their findings supported the view of the brain as a predictive system. However, it is plausible that they were conflating a process model of a generic target system with a mechanistic model. In scientific contexts, mechanistic models serve as surrogate representations of structures and processes in physical targets. Yet their findings lacked sufficient detail about the physical components of the putative mechanism to be considered part of a mechanistic explanation. I discuss this problem further in Section 3.4, where I explore the epistemic gain of target-directed modeling.

The Action Observation Network

The Action Observation Network (AON) model (Urgen and Saygin 2020) is an instance of a model of hypothetical mental information processing mechanisms. It describes the activity of a network of regions (pSTS, parietal, and premotor) in the primate brain involved in visual perception of actions. The model conceptualizes this network as generative model, meaning the interactions among its parts work according to the principle of error minimization, transmitting prediction-error signals to one another. This network is activated during action perception but not object motion perception. Previous research showed that these regions were active during action perception (Saygin, 2007), but it did not explain how they communicate. Urgen and Saygin (2020) constructed the AON model to identify how these regions communicate under the assumption that this network is involved in the predictive task of perceiving actions.

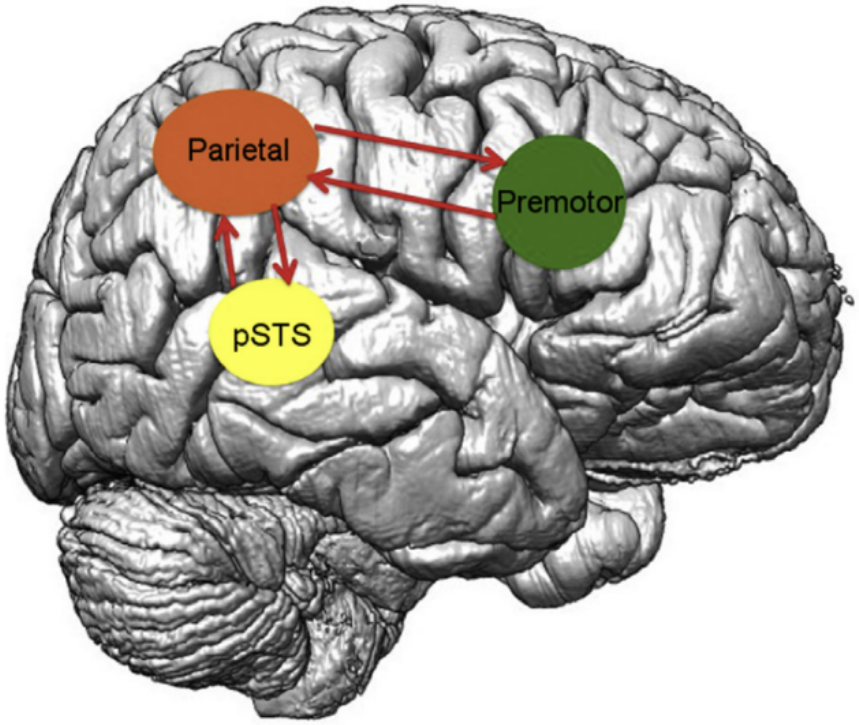


Figure 4: Anatomical Connectivity of the Three Nodes of the AON (Urgen and Saygin 2020, reproduced with permission)

Urgen and Saygin (2020) took a modeling approach to answer the question about how the AON components communicate. Their strategy involved constructing surrogate information processing architectures of potential organizational structures for AON node communication to determine whether these nodes' connections are forward or backward. The AON model incorporates theoretical notions from the PP, including the assumption of a hierarchical organization of information processing in the brain. This assumption was incorporated into the design of the AON model, which comprises three nodes that exhibit both forward and backward connections. Specifically, the lower node is associated with the pSTS region, the middle node with the parietal cortex, and the upper node with the premotor cortex. Additionally, the middle node has reciprocal connections with the other two nodes, enabling forward and backward connections.

The modelers then proposed possible information processing architectures of their communication. They specified that the pSTS-parietal-premotor network comprises a system of three nodes with feedback and backward connections. Urgen and Saygin (2020) first addressed the question of whether the parietal cortex (represented by the middle node) receives feedforward influence from the pSTS or feedback influence from the premotor cortex in the mismatch condition. Based on the PP assumption that the brain transmits error

signals in mismatch conditions, they postulated that the influence originates from the premotor cortex because this region emits error signals in mismatch conditions. To answer this question, Urgen and Saygin used Dynamic Causal Modeling (DCM), a PP modeling technique that is described in more detail below:

> The central idea behind dynamic causal modelling is to treat the brain as a deterministic nonlinear dynamic system that is subject to inputs and produces outputs. Effective connectivity is parameterized in terms of coupling among unobserved brain states (e.g., neuronal activity in different regions). The objective is to estimate these parameters by perturbing the system and measuring the response (Friston, Harrison and Penny 2003, as cited in Urgen and Saygin 2020, 134).

DCM is a technique for estimating the direct connectivity of brain regions. The method operates as follows:

i. First, modelers must describe several hypothetical information processing architectures of selected brain regions and hypotheses about how experimental manipulations modulate their connections.
ii. Subsequently, modelers must estimate three parameters: the intrinsic connections between the relevant regions, the experimental modulation of the intrinsic connections, and the strength of the resulting inputs.
iii. Ultimately, modelers must identify the most probable architectural configuration based on the available data. To accomplish this, they employ Bayesian model selection to simulate the behavior of these architectures (Vincent 2014).

Urgen and Saygin (2020) identified three potential information processing models for the pSTS-parietal-premotor network within the AON model construction. Previous scientific findings regarding the intrinsic connection within this network determined that the pSTS area serves as the input node in all hypothetical architectures. Moreover, the experimental modulation was consistently initiated by an action condition, namely the observation of an action. In accordance with the PP hierarchy assumption, the modelers postulated that the action condition triggered activations in the pSTS, which then propagated to the parietal and premotor areas. It was further assumed that these areas would send feedback information to pSTS.

The three candidate architectures exhibited disparate distributions of feedforward and feedback connections. In the initial configuration, the pSTS exerted bottom-up modulation on the parietal node. In the second, the premotor node performed top-down modulation on the parietal node. In the third, both bottom-up and top-down modulation occurred. In the simulations, all architectures demonstrated a response to the action condition. However,

there was a discrepancy in the way the mismatch conditions modulated the connections (Urgen and Saygin 2020, 136). The results indicated that the second candidate architecture was the most probable given the data, suggesting that the premotor node exerts a top-down influence on the parietal node during the mismatch condition.

The Saygin laboratory's research shows that empirical studies within the PP framework produce remarkable results. Specifically, designing an experiment to measure prediction error in mismatch conditions enabled researchers to discover an increase in the amplitude of the N400 effect. I will briefly summarize the epistemic gains associated with constructing the AON model and subsequent research conducted in the Saygin lab based on assumptions derived from PP.

i. The scientists at the Saygin Lab employed the PP assumption of a hierarchical organization of the brain to characterize the components of the communication of nodes in the AON model.
ii. The concept of prediction error prompted the formulation of a mismatch condition, which in turn facilitated the formulation of scientific predictions and hypotheses regarding the communication of the AON components.
iii. Ultimately, the researchers employed a technique known as DCM (Dynamic Causal Modeling), another PP method, to estimate the degree of connectivity between brain areas. This was achieved by conducting simulations on cortical data.

The second case study will further consider other epistemic gains of using generative models in science. Unlike the first case study, the next one is not directly concerned with advancing our empirical understanding of target systems. Rather, it focuses on developing scientific tools for investigating them.

3.3. The Agent-Environment Coupled System Model

Targetless Modeling

This section presents a case study of targetless modeling in the PP framework, namely the construction of the Agent-Environment Coupled System model (hereafter the AECS model) by Jelle Bruineberg, Erik Rietveld, Thomas Parr, Leendert van Maanen, and Karl Friston in 2018. The AECS is a targetless model in that it does not have a specific or generic

phenomenon as its target. Moreover, it does not contribute to our understanding of reality, at least not in a direct manner.[10]

What is the epistemic function of targetless modeling in cognitive science? I propose that it facilitates theory development, broadly understood as the creation of theoretical resources, such as concepts and modeling techniques. In the construction of the AECS model, Bruineberg et al. (2018) use an equation known as the Free Energy Principle (FEP) to create a model of niche construction. While this theoretical practice does not augment our understanding of reality, it can facilitate the development of epistemic instruments that can be used in target-directed modeling.

Scientists construct and manipulate models to explore observable phenomena (e.g., the neural correlates of action observation), hypothetical systems (e.g., the neural correlates of action observation), generic targets (e.g., evolution or learning), and non-existent targets (e.g., an agent-environment coupled system). In **targetless modeling**, the objects of study are the conceptualizations of target systems and their behavior, without consideration of any concrete target. "The only object of study is the model itself, without regard to what it tells us about any specific real-world system" (Weisberg 2013, 129). Bruineberg et al. (2018) constructed the AECS model to demonstrate that niche construction, an abstract behavior, can be modeled using the Free Energy Principle. Accordingly, the object of their modeling practice is the model itself, a model of a generic target system and its behavior.

The construction of the AECS model involved conceptualizing a generic agent-based system, attributing a structure to it, and simulating it. The model was used to simulate the behavior of a generic system of dependencies. In the cognitive sciences, targetless models incorporate the conceptualizations of target systems into tangible, manipulable models. According to Weisberg (2013, 131), targetless modeling facilitates the creation of a "more general modeling framework that can be used for target-directed modeling." The construction of the AECS model applies the FEP, which was originally developed in physics and theoretical biology, to the problem of niche construction. In doing so, scientists not only reproduce a model, but also repurpose its concepts and associated mathematical and computational apparatuses. The AECS model construction I present here demonstrates that targetless modeling enables scientists to apply an existing model to a new domain.

[10] My argument that modeling contributes to gaining an understanding of reality while not having access to it seems to be problematic. My understanding of reality is based on Hacking (1983), and follows the idea that we understand what is real by our ability to manipulate and intervene in the world. I have not thematized the concept further because this work is not intended to engage in debates on scientific realism.

The Agent-Environment Coupled System Model of Niche Construction

In contrast to the AON model, the AECS model does not provide specific predictions regarding the behavior of a target system. Rather, modelers constructed it to demonstrate that niche construction can be modeled using the FEP. "Niche construction is the process whereby organisms actively modify their own and each other's evolutionary niches" (Odling-Smee et al. 2003, as cited in Laland, Matthews, and Feldman 2016, 192). Although niche construction is a phenomenon observable in nature, the AECS model was not designed to target a natural phenomenon. The objective was to apply a formal equation to investigate the behavior of a generic agent in a synthetic environment. Consequently, the model conceptualized a behavior functionally, that is, independently of its physical realization. The goal was to model niche construction using a mathematical formalism.

Bruineberg et al. (2018, 162) provide an illustration of niche construction with the example of the emergence of a **desire path**, which is introduced as follows: "Rushing on their way to work, people might cut the corner of the path through the park. While initially, this might almost leave no trace, over time, a path emerges, in turn attracting more agents to take the shortcut and underwrite the path's existence." The emergence of a desired path is an example of developmental niche construction, which can be defined as "the construction of ecological and social legacies that modify an agent's learning process and development" (Bruineberg et al. 2018, 162). In other words, developmental niche construction is the creation or modification of environmental structures that result from unconscious or non-goal-directed effects of actions.

In contrast, goal-directed niche construction[11] refers to "the active modification of an environment such that the selective pressures on heritable traits change as a result of these modifications" (Bruineberg et al. 2018, 162). The abstract characterization of niche construction in the AECS describes this behavior as a process that emerges from local interactions between an agent and their environment. This will now be explained in further detail.

When modeling systems without specific targets, scientists typically proceed by describing a generic behavior (e.g., adaptation, reproduction, or evolution) and attributing a structure to it (Weisberg 2013). The AECS model construction proceeded in this way. The model specifies that niche construction is the global behavior exhibited by a system. In particular, it defines it as a form of adaptive behavior that emerges from the local interactions of a generic agent and a synthetic environment. The agent and the environment are a coupled

[11] Aaby and Desmond (2021, 1) point out that organisms "act as agents when niche constructing effects are a result of a goal-directed behavior over which the organisms have some degree of control."

system because their states mutually influence each other. In this context, attributing a structure involves using the FEP to formally represent the interactions among the parts of this system. As Weisberg (2013) points out, attributing a structure to a model is nothing more than the specification of arbitrary rules embedded in algorithmic formats or equations.

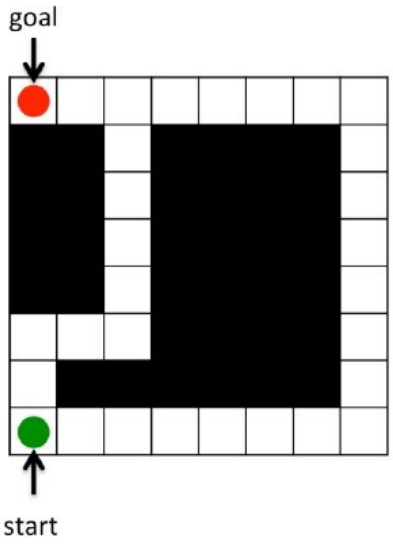


Figure 5: A graphical representation of the AECS model (adapted from Bruineberg et al. 2018, reproduced under Creative Commons license terms)

Figure 5 illustrates the AECS model, which comprises three components: the agent, the environment, and the interactions between them. In agent-based modeling,[12] agents are autonomous entities that make decisions in accordance with the rules they have been programmed to follow. The green circle represents an agent that can move in four directions—up, down, right, and left—from its current square to a neighboring square. An environment is the setting in which actions occur and may possess distinct characteristics. In the AECS model, the environment is represented by an 8x8 grid maze, comprising a goal square (red circle) and squares of varying openness, as indicated by their color (white for open and black for closed). interactions are either the effects of actions on the environment or the effects of the environment on the agent.

The objective of the agent in the AECS is to reach the target square, indicated by the red circle. In the initial stage, the agent is only aware of the open/closed status of its current

[12] Agent-based models have been employed in a number of fields, including "economics (Grazzini and Richiardi 2015), social sciences (Axtell and Epstein 1994; Epstein and Axtell 1996; Schelling 1971; 2006), and ecology (McLane, Semeniuk, McDermid, and Marceau 2011). They are interactive systems with self-organizing capacities (Bonabeau 2002; Niazi and Hussain 2009, as cited in Madsen et al. 2019, 301).

location; it is not yet aware of the full distribution of open/closed squares. Subsequently, upon transitioning to a neighboring square, the agent gains insight into the state of the square (i.e., open or closed) and utilizes this information to construct a model of the distribution of open and closed squares within the maze, thereby developing a representation of the environment. Should the agent proceed to a closed square, a penalty is incurred. As a result of the agent's movements, the distribution of open and closed squares in the model is subject to change. Therefore, the agent is obliged to update its model of the environment to develop optimal policies for reaching the goal square.

Cognitive scientists employ agent-based models as surrogate systems to investigate complex cognitive behavior. In this context, complexity is defined as a property of systems comprising numerous components that interact and exert influence upon one another. The behavior in question is complex because of the interdependencies between numerous variables. For example, social cooperation is a complex behavior because it involves the interaction of multiple agents with diverse personal and socio-cultural backgrounds. The limitations of linear, causal, and mechanistic models are such that they are unable to adequately capture the behavior of complex systems. For example, the AON model excludes significant aspects of action observation, including its situatedness within practical and interactive engagements with other agents. Agent-based modeling can address the shortcomings of alternative approaches for investigating and simulating cognition by enabling scientists to visualize the dynamics of complex systems through the simulation of interactions between local components (Madsen et al. 2019).

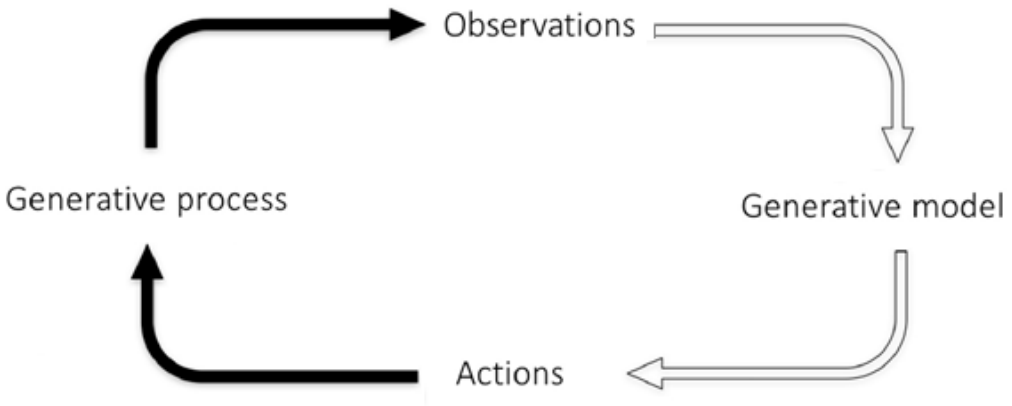
The modelers conceptualized the AESC as a coupled system, whereby the global behavior of the system is a result of local agent-environment interactions. If the agent can adapt its behavior to fit within the environment, it will reach the goal square with minimal or no penalty. For the agent to fit, it must learn the structure of the environment, and the environment must change its structure to better fit the agent (Bruineberg et al. 2018).

The modelers treat the AECS as an analogous model of niche construction, which is defined as a behavior in which an agent intervenes in an environment that, over time, becomes more suitable for the agent. In this manner, the modelers projected the properties of coupled systems to niche-construction. This was achieved by conceptualizing agent-environment interactions as a self-organizing behavior. In the AECS model, agent-environment interactions give rise to patterned behavior that resembles a feedback loop: "What the agent does changes the environment, which changes what the agent perceives, which changes the agent's expectations, which changes what the agent does (to change the environment)" (Bruineberg et al. 2018, 162).

In summary, the AECS model is a generic model of niche construction as a coupled system in which an agent adapts to an environment. In the model, the fit of this agent in the environment is achieved through a feedback loop, as the actions of the agent directly influence the state of the environment, which subsequently affects the state of the agent.

The Construction of a Scientific Model

As stated above, modelers used the FEP to construct the AECS system. The FEP is an equation with origins in thermodynamics and biophysics that has been used in neuroscience to characterize the brain as a system that minimizes surprise states through perception and action. In accordance with the formalisms of the FEP, the brain maintains non-equilibrium steady states (i.e., it conserves its average energy without reaching equilibrium) by constraining itself to a finite number of states that are perceived as viable (non-surprising states). Moreover, FEP systems are self-organizing in that their states are independent of external environmental states. However, environmental states can trigger surprising sensory states. Consequently, FEP systems must infer how environmental states cause surprising states to avoid them (a process called perceptual inference). Perceptual inference necessitates that systems establish a boundary between their internal and external states and construct a generative model of how external states elicit surprising states. FEP systems can reduce free energy in two ways: by modifying the parameters of the generative model and by a process known as active inference. The latter entails generating and pursuing policies that are anticipated to transform environmental states into non-surprising ones (Friston 2012b).[13]



[13] It is important to acknowledge that actions may fail to transform surprising states into non-surprising ones due to the uncertainty surrounding the distribution of states within the environment. For instance, "the action that reduces surprisal of temperature sensors depends on where the agent can find shade" (Bruineberg et al. 2018, 163).

Figure 6: An agent-environment coupled system under the Free Energy Principle

The generative model and the process correspond to the agent and the environment, respectively. The generative model is capable of inferring the hidden environmental states that give rise to surprising states (observations). A generative process can be defined as a distribution of hidden environmental states that give rise to surprising states. Actions that are intended to elicit non-surprising states may result in a modification of the distribution of hidden environmental states (adapted from Bruineberg et al. 2018. Reproduced under Creative Commons License terms).

Bruineberg et al. (2018) conceptualize the agent and the environment as generative models and processes, respectively. Since the AECS model was designed to simulate coupled systems, free energy is not relative to either the agent or the environment. Rather, it is relative to the entire system, which jointly minimizes it. The agent, or generative model, makes inferences about environmental states based on its beliefs about the distribution of those states. Furthermore, the agent forms and follows policies with the objective of minimizing its free energy. Figure 6 illustrates the inference process, showing how the agent determines which neighboring squares are open or closed to reach the square with the highest probability of being open (unless an exploration policy is followed). A generative process refers to the hidden distribution of environmental states that changes as a result of actions. The interactions of this coupled system are the effects of actions that modify the generative process, potentially transforming open states into closed states (i.e., surprising states into non-surprising states) or vice versa. Agent-environment interactions reduce the discrepancy between a generative process and a generative model as governed by FEP, contingent on the properties of the agent and the environment.

The question thus arises as to how the FEP addresses the above issue of fit. The FEP establishes a demarcation between the states of the generative model and those of a generative process. They communicate through observations and actions (see Figure 6), which the FEP characterizes as inferential. In this context, the fit problem is the task of an agent to reach a goal square by probabilistically inferring the distribution of environmental states (whether they are open or closed). The agent is equipped with perceptions, which are current state observations; priors, which are accumulated information about the environmental states distribution; actions, which are movements; and policies, which are actions, such as exploring the environment or exploiting prior knowledge to reach the goal square faster.

Policies are analogous to hypotheses (e.g., first move up, second move down, third move right). They are action plans to reach a specified goal state by leveraging the agent's prior knowledge via the exploration of the environment. The agent follows a general policy of

avoiding closed squares on each trial. However, sometimes the agent moves to closed squares, despite the presence of open squares in the surrounding area. This behavior can be attributed to the agent's failure to consider the distribution of open and closed squares. Furthermore, given that actions alter the distribution of squares within the grid maze, it is necessary for the agent to possess an additional policy through which it can explore the ways in which its actions affect the environment. As a consequence of the agent's actions, the environment undergoes changes that necessitate the constant optimization of the hypotheses and policies in place. To successfully complete this task, the agent must learn the distribution of environmental states and monitor how its movements and actions influence that distribution. The FEP offers a mathematical solution to this task of probabilistic inference, showing possible scenarios of fit between agent and environment.

Niche construction is in this context "the behavior in which agents plan the best route through the environment and then carve out that route" (Bruineberg et al. 2018, 176). In modeling the AECS under the FEP, the latter—a formal structure that specifies the rules governing agent-environment interactions—serves as the basis for describing niche construction systems. This type of structure permits the implementation of methodologies such as computer simulation. It is noteworthy that, through the simulation of the AECS model, Bruineberg et al. (2018, 162) have demonstrated that only specific systems with particular attributes can engage in niche construction. For example, it does not occur in the limiting case of a malleable agent and a rigid environment, where the agent learns the structure of the environment. In contrast, niche construction is observed when the agent is rigid, and the environment is compliant. In this scenario, the agent intervenes in the environment to align it with its expectations. It is noteworthy that the FEP provides a mathematical description of these properties. "The malleability of the agent and the environment can be given a concise mathematical description in terms of the prior beliefs. These prior beliefs reflect the influence sensory evidence has on learning. In other words, they determine the 'learning rate' or 'inertia' of both the agent and the environment" (Bruineberg et al. 2018, 162).

In this section, I have presented the construction of the AECS model as a process of using the FEP equation to simulate generic behavior in an agent-based modeling system. The objective in constructing this model was to develop one that could be used to simulate niche construction. The model describes the behavior of a generic agent interacting with an environment whose states modify the agent's states, and vice versa. The epistemic resources of PP, which will be discussed in greater detail in the next section, help scientists transform the cognitive phenomenon of niche construction into a solvable informational task. The final section considers the epistemic value of constructing the AON and the AECS models.

3.4. The Epistemic Gain of Modeling the Mind as a Predictive System

The Epistemic Gain of Target-Directed Modeling

In Section 3.2, I presented the AON model of pSTS-parietal-premotor network communication. I will now provide a brief overview of the role that PP epistemic resources play in the construction of this model and the scientific strategies involved in their construction.

My analysis of the AON model construction reveals the use of an idealization strategy in the conceptualization of the target system, namely the hypothetical pSTS-parietal-premotor system. The target was conceptualized as a network in which information is transmitted in both forward and backward directions. Moreover, an abstraction strategy was employed in the construction of the AON model, which entailed the exclusion of several attributes of the pSTS-parietal-premotor network.

The epistemic gains of the construction of this model are related to the production of scientific predictions and hypotheses about empirical phenomena. However, these strategies do not facilitate the acquisition of empirical knowledge about hypothetical cognitive mechanisms that are responsible for action perception, contrary to the original expectations of the scientists who developed the models.

Knuuttila and Morgan (2019) identify two distinct meanings of idealization present within the philosophy of science. Idealization is typically regarded as a part of a process of model construction. To illustrate, the process of idealization may be achieved by distorting the target, simplifying some of its features, or even obviating them entirely. An additional meaning pertains to a quality inherent to the models themselves. The idea of an idealized model implies that models are simplified, inaccurate, or distorted representations of the entities of which they are models. In my view, the practice of modeling the brain as either an information processing or predictive system is an instance of idealization, given that it involves a simplified, distorted, and inaccurate characterization of this organ's complexity. However, this strategy makes the brain amenable to scientific research.

The strategy of idealization involved in the construction of the AON model was more specific than the assumption that the brain is an information processing system. The idealization entailed modeling the communication between brain areas as if it were that of a hierarchical network, wherein higher nodes send backward signals that modulate forward signals from lower nodes. In light of existing knowledge regarding the brain areas involved, the modelers suggested that the middle nodes (the parietal nodes) would exhibit both types of

connections. It is only on the assumption that the brain is a hierarchical system that specifications about the types of connections in this system can be made based on existing knowledge or available data about the network under consideration.

As mentioned in Section 3.2, the AON model yielded testable hypotheses and scientific predictions about a brain network. The idealization involved in constructing the model enabled scientists to use empirical knowledge about the relevant brain areas to formulate initial hypotheses and predictions. In other words, the way a target system is conceptualized during model construction enables the integration of empirical knowledge in a scientific practice.

Notably, the PP concepts and assumptions were not limited to the idealization strategy in the Saygin Lab research. The concept of prediction error signal also played a role in designing experiments. While this is not directly related to modeling, it is important to mention because it demonstrates that scientific conceptualizations are relevant for characterizing target systems in scientific practices other than modeling. These practices can yield scientific predictions, an important outcome in scientific research. Saygin et al. (2012) based their experimental design on the PP assumption that a mismatch condition—such as a human figure accompanied by mechanical movement—would elicit error signals. They postulated that perceiving a mismatch condition would result in distinctive brain activations. Based on this premise, they discovered that the mismatch condition evokes distinctive responses in the bilateral anterior temporal sulcus. Subsequent studies have built on this work, including a study by Urgen, Kutas, and Saygin (2018), which accurately predicted an increase in the amplitude of the N400 effect in the mismatch condition.

Despite the accuracy of their scientific prediction, Urgen, Kutas, and Saygin (2018) misjudged the outcome of their experiment. They assumed that an increased amplitude of the N400 effect would serve as evidence for the existence of predictive mechanisms in the brain. However, I do not believe this to be the case because the aforementioned information processing and functional model is not mechanistic. Also, scientific predictions about the communication of brain areas are functional. In other words, they are not predictions about the physical implementation of such communication. Rather, these scientific predictions are based on data that abstracts many relevant details about the physical realization of the cognitive phenomenon.

The modeling of the AON system at the Saygin Lab was a significant contribution to the field, providing a deeper understanding of less well-studied brain areas. It is important to note that, like information processing models, AON models are not mechanistic and should not be considered models that make abstract representations of physical networks. While the AON model can be considered an idealized representation of the pSTS-parietal-premotor

network, it is incorrect to assume that scientific predictions and hypotheses about this network can be formulated based solely on the model's representational power. Using a PP technique, the modelers proposed a hypothesis about the communication patterns of the network's constituent areas. This technique transformed the problem of how these areas communicate into the problem of identifying the most probable information-processing architecture (feedforward, feedback, or both) for a given dataset. In summary, constructing the AON model demonstrated the utility of the PP modeling framework for empirical research because it can be used to formulate testable hypotheses and scientific predictions. However, scientists and philosophers must be careful not to conflate models with scientific *representations of* mental processes or mechanisms.

The Epistemic Gain of Targetless Modeling

In section 3.3, I presented a case study of the construction of a model without a target. As Weisberg (2013) asserts, target-directed models can be empirically informative insofar as they delineate the intended scope of the model and scientists elucidate how they provide an understanding of a target system. This may present a challenge, given that models are sometimes employed to examine poorly understood systems, or due to idealization assumptions (as discussed above), or unrealistic ones. For example, Lewandowsky et al. (2019) employ a Bayesian approach to model opinion formation as a strategy of accumulating evidence, yet their model is based on unrealistic assumptions about how individuals form beliefs about climate change. Notwithstanding the unrealistic assumptions, Lewandowsky et al. made accurate scientific predictions.

If the existence of a target is a prerequisite for attaining empirical knowledge about reality, the epistemic gain of targetless modeling should be evaluated in light of other factors. As previously stated, the objective of the AECS model was to demonstrate that an abstract conceptualization of a phenomenon known as niche construction can be instantiated in an agent-based system through the application of a specific equation. This has no direct influence on the generation of scientific predictions and hypotheses. Therefore, it does not contribute to our empirical understanding of reality. Nevertheless, these models can be instrumental for future target-directed modeling practices.

The assertion that the construction of such models is epistemically valuable is contingent upon the premise that models must first be constructed before they can be utilized for the investigation of targets. The argument has also been made that targets are themselves constructed within research practices. In target-directed modeling, a constructed target system is already available. In contrast, in targetless modeling, this is not necessarily the

case. Thus, it seems reasonable to view targetless modeling as a practice oriented towards the construction of a target system in a manner that renders it suitable for model-based research. In the case of the AECS model construction, this practice entailed conceptualizing niche construction as an information processing task. While alternative conceptualizations of niche construction may exist (for instance, ecological ones), the conceptualization in place was an informational one. The objective was to transform niche construction into a task that could be solved using mathematical and computational epistemic resources.

The epistemic gain of model construction in the case of the AECS model is the creation of instruments for modeling target systems. It is only as a result of a modeling practice that involves the application of an equation to niche construction, which is conceptualized as an informational problem. The equation in question, namely the FEP, is just one among many other mathematical tools that could be employed in the construction of models of or simulation of cognitive behaviors.

Models such as those being produced under the umbrella of PP have the potential to become important epistemic resources in the cognitive sciences. They may also provide a means of addressing the dilemma faced by many cognitive scientists between conducting qualitative or quantitative research. It is noteworthy that research in the field of cognitive science, which is situated within the broader tradition of 4E cognition, has predominantly employed qualitative or first-person methodologies. Nevertheless, even for those who are most enthusiastic about these approaches, it is not possible to gain a complete understanding of the full range of factors involved in complex cognitive behaviors through qualitative methods alone. As Gallagher notes,

> Cognition is not simply a brain event. It emerges from processes distributed across the brain-body-environment. The mind is embodied (…); from a first-person perspective embodiment is equivalent to the phenomenological concepts of the lived body. From a third-person perspective the organism-environment is taken as the explanatory unit (Gallagher 2017 quoted in Fingerhut 2020, 354).

To my understanding, targetless modeling offers a way to address the aforementioned dilemma. It enables the creation of epistemic tools that facilitate quantitative science and numerous other potential epistemic gains. Apart from PP, few approaches have been used to address embodied cognition problems, such as dynamic systems theory (Chemero 2009; De Haan 2020). Although the PP framework has emerged as a promising quantitative counterpart to 4E cognition, its association with it is currently somewhat *ad hoc* (see Allen and Friston 2018).

Since PP provides formal tools for modeling cognitive problems as probabilistic inference tasks, one could argue that they could complement other substantive theories of cognition for similar purposes. In the case of the AECS model, however, the conceptualizations of niche construction employed in other theories are irrelevant to the model's abstract characterization of this behavior. The objective of constructing the AECS model was to provide an abstract cognitive behavior with a mathematical apparatus. This is a prerequisite for target-directed modeling and quantitative research on niche construction.

As a concluding remark, it seems pertinent to discuss some misinterpretations of the epistemological implications of PP that have arisen in the absence of a perspective on model construction. As previously discussed, the AECS model construction employs FEP to characterize a cognitive behavior. This represents a case of targetless modeling, which does not contribute to the advancement of empirical knowledge about the mind, but rather provides epistemic resources for investigating it. The AECS model construction facilitates the creation of tools that transform complex cognitive problems into simpler informational problems. However, as the resulting models are formal mathematical tools, further assumptions must be made in modeling practices using them before they can be used to attain empirical knowledge (Sims and Pezzulo 2021).

Despite its formal nature, the informational perspective on cognitive processes, as proposed by PP and the FEP, does imply the transfer of certain assumptions about informational systems to cognitive phenomena. These assumptions are inherent to the modeling process. One such assumption is that a cognitive agent must occupy a distinctive position within a state space to be adaptive. This is an informational assumption, not a cognitive one. Friston (2017) further elaborates on this assumption as follows:

> According to physicists, complex systems can be characterised by their states, captured by variables with a range of possible values. In quantum systems, for example, the state of a particle can be described by a wave function that entails its position, momentum, energy and spin. For larger systems, such as ourselves, our state encompasses all the positions and motions of our bodily parts, the electrochemical states of the brain, the physiological changes in the organs, and so on. Formally speaking, the state of a system corresponds to its coordinates in the space of possible states, with different axes for different variables.

One potential explanation for the aforementioned misunderstandings is the language used by proponents of the FEP. Although it is a formal principle, its inspiration from physics is relevant to how it conceptualizes cognitive agents as dynamic systems with state-space trajectories. Friston seems to suggest that cognitive agents are truly dynamic systems, rather than this being merely one fruitful way of conceptualizing them. Another assumption implicit

in the construction of FEP models of cognition is that biological systems are self-organizing. This idea is elaborated upon below:

> Living systems are unique in nature, since among all self-organising systems, they seem to maintain their organisation when facing environmental perturbations. Most self-organising systems dissipate the gradients around which they emerge: a lightning bolt, for instance, effectively destroys the gradient in electrical charge that gave rise to it. Organisms, strikingly, not only self-organise but manage to persist across time as self-organising systems (Ramstead et al. 2018). Heuristically, we can say that organisms expect to be in their characteristic phenotypic states; surprising deviations from these expectations must be avoided to maintain the system within viable (i.e., phenotypic) states (Ramstead, Kirchhoff and Friston 2020, 229).

Applying the FEP to the characterization of cognitive systems frames them as self-organizing systems that distinguish between surprising and non-surprising states. Importantly, informational models do not constrain the ascription of cognitive content (Wiese 2017, 715). The constraints that the FEP imposes on systems are informational, not cognitive or biological (Wiese 2017; Andrews 2021). These constraints depend on the additional assumptions scientists make when modeling cortical activity, biological systems, or cognitive behavior. I argue in Chapter 6 that the FEP is a model template in and of itself.

Proponents of the FEP have different views on its epistemic role. Friston treats the FEP as a transcendental, even metaphysical, principle of life rather than focusing on explaining its scientific function. This principle defines biocognitive systems as "self-organizing systems that can exhibit consciousness" (Beni 2021, 6411). I am skeptical of these views given that principles are susceptible to numerous philosophical interpretations, some of which may be metaphysical, as Beni (2021) has suggested. In his words,

> Some interpretations regard the FEP as a fundamental hypothesis or theory with somewhat realist tendency to represent nomic relations between the brain and the environment (Allen and Friston 2018; Wiese and Metzinger 2017). However, it is also possible to argue that being an axiom or postulate that is grounded in theoretical physics and biology makes FEP a scientific principle (Colombo and Wright 2021). In this vein, it may be (and indeed has been) argued that FEP provides a principled explanation that serves as a guide to which things may or may not accord, rather than a hypothesis that provides a *bona fide* description of the target system (Beni 2021, 6410).

As will be discussed in the following chapter, my proposed modeling approach is compatible with more substantial views of PP and the FEP. Specifically, it does not eliminate the

possibility that cognition is predictive in nature or that cognitive systems are self-organizing entities driven by free energy minimization. My view is that these assumptions are not integral to constructing PP or FEP models of cognition. In other words, these assumptions are part of the methodology scientists use to solve cognitive problems, rather than assumptions about the phenomena they are modeling. In PP, these assumptions impose informational constraints rather than physical constraints in the characterization of target cognitive systems. In Chapter 6, I discuss the nature of the FEP as an epistemic resource that imposes informational constraints on systems.

Chapter 4 - Ontological Agnosticism

The idea that cognition is predictive modeling has been the subject of considerable debate among philosophers with different conceptions of the mind. Two representative accounts that have emerged in the last decade are Conservative Predictive Processing and Radical Predictive Processing. Both assume that cognition operates through predictive modeling but have different understandings of how mental models work. For the conservatives, mental models are similar to mental representations. For the radicals, the mind's predictive activity is non-representational and non-brain-bound; it extends into the environment.

In this chapter, I defend an agnostic stance on the aforementioned positions and their ontologies of the mind, which are said to be integral to Predictive Processing. To do so, I critically discuss key philosophical notions important to philosophical debates about the cognitive sciences. These include theories of representationalism and 4E cognition, as well as doctrines of internalism and externalism. I examine how proponents of PP have revisited these ideas.

I argue that, when considered as modeling assumptions rather than defining features of the mind, these views can have an epistemic function in scientific practices. Like other conceptualizations in scientific practices, the concepts of mental representations, internal mental states, and mental models are not primarily designators of mental entities or structures. Rather, they are theoretical resources that serve to construct ontologically neutral scientific models.

Since the role of concepts in modeling practices has already been addressed in previous chapters, this chapter presents a discussion of my perspective on mental modeling and predictions as mental kinds. I refer to my perspective as **ontological agnosticism**. Unlike anti-realism regarding mental representations, internal states, or mental models, my agnostic stance does not reject the possibility of their existence. I merely propose that one can be agnostic about the reality of mental models yet still use scientific models to study cognition.

Ontological agnosticism is the view that the science of PP neither implies nor supports the existence of mental models. This is because PP is a modeling framework comprising various epistemic resources useful for constructing models. These resources include conceptual

ones that can be drawn from various philosophies of mind. When employed in a scientific context, these resources do not function as designators of mental properties, entities or structures. Ontological agnosticism opens the possibility of ontological pluralism. Once we acknowledge that the sciences of the mind do not imply a specific ontology of cognition, we can accept that they tolerate various substantive views of cognition.

The chapter is divided into three sections. Section 4.1 provides an overview of key views that have informed discussions of mental modeling in the philosophy of the cognitive sciences. These views include mental representationalism, enactivism, embodied cognition, extended cognition, embedded cognition, internalism, and externalism. These views will be the subject of subsequent sections' discussions.

Section 4.2 critically examines Conservative Predictive Processing (CPP), a realist, representationalist, structuralist, and internalist view of generative models. The chapter elaborates on these notions in the context of PP, particularly on how they serve to characterize generative models. Finally, a series of objections to this interpretation of generative models are presented.

Section 4.3 interprets generative models as epistemic artifacts inspired by the extended cognition thesis. Generative models allow scientists to conceptualize systems in unique ways without specifying any material realization of such systems. Importantly, using them influences the way scientists think and imagine cognition. This view justifies my ontological agnosticism. To argue that generative models are mental entities is to conflate an operational description of a system that is valid within a scientific practice with an ontological description of cognition. This operational description can be used to model any cognitive system, regardless of whether it is representational, enactive, embodied, extended, or embedded. This is why it is ontologically neutral.

4.1. Mental Representations in the Cognitive Sciences

Cognitive science emerged as a discipline in the mid-twentieth century as a response to the dominant tradition in psychology known as behaviorism. Based on Rey (1997), Graham (2023) distinguishes between methodological, psychological, and logical behaviorism. Methodological behaviorism posits that the objects of study in the sciences of the mind are observable behaviors, and that the discussion of mental events or states contributes nothing to our understanding of behavior since they are not objects of empirical investigation. Psychological behaviorism posits that behavior can be understood as a causal

sequence of stimuli and responses, driven by learning processes. In logical behaviorism, mental states are conceptualized as behavioral dispositions, and the very notion of a mental state is regarded as a mere linguistic construct employed to describe behavior.

Cognitive science originated as a response to behaviorists' attempts to explain behavior based solely on observable stimulus-response sequences. To understand this, consider the following question: Why is using a pen to write a memo on a piece of paper qualitatively different from moving a pen by accident? One action can intuitively be seen as intelligent, while the other cannot. While our intuition suggests this, it does not explain why. A behaviorist could argue that the difference between writing something and accidentally moving the pen is due to a learning process involving reinforcement routines. Specifically, they might suggest that humans acquire writing skills through a process of repetition involving punishment or rewards.

Such an answer would have been inadequate for the earlier cognitive scientists. For them, the defining characteristic of cognition is that it involves an internal process in the cognitive agent that is triggered by stimuli and that precedes, and shapes, the reaction. This process enables the cognitive agent to perceive environmental objects, such as the pen or paper, as tools for writing. Other analogous processes facilitate the execution of skilled movements by the cognitive agent to write on the paper. The act of selecting specific symbols in a particular sequence to convey a specific meaning is an example of a more complex cognitive process.

Early cognitive science theories sought to explain how the mind and/or brain cognize in terms of complex representations and computational procedures (Thagard 2005, ix). Though not the primary focus of this chapter, I briefly describe these computational procedures because they are integral to the concept of representation in the cognitive sciences. According to these early theories, cognition involves a computational process analogous to that performed by Turing machines.

> A Turing machine is an abstract model of an idealized computing device with unlimited time and storage space at its disposal. The device manipulates symbols, much as a human computing agent manipulates pencil marks on paper during arithmetical computation (Rescorla 2020).

Following the analogy of the Turing machine, thinking or cognition is similar to running a particular program or set of rules, given a sensory input, to produce a particular output. In the example above, to perceive a pen, our cognitive machine, after receiving some sensory input, would run a program that transforms the sensory signal into a mental image or

representation, e.g., visual data from the retina is processed into information about the spatial location and shape of the pen.

What is unique about Turing machines is that their computations operate on syntactic structures. Machines such as computers can recognize symbols by their syntactic properties (Slors, de Bruin, and Strijbos 2015). As syntactic structures, the symbols mechanically manipulated by Turing machines lack content. For this reason, the Turing machine analogy fell short of explaining human cognition. The early cognitive scientists postulated that the computations of the human mind are qualitatively different from those of Turing machines because the mind manipulates symbols that have content (Bermúdez 2014).

To explain how human minds come to know the world and acquire such content, cognitive scientists postulated the existence of mental entities called **mental representations**, which Cain (2016, 17) describes in the following:

> A first idea was widely endorsed by philosophers during the modern period of the seventeenth and eighteenth centuries. This is the idea that the mind is populated by representations. A representation is a symbol residing in the mind, and mental representations are involved whenever a person is in a particular mental state (for example, has a belief or a perceptual experience) or executes a mental process (for example, thinks or recollects a past event). Just as external representations such as spoken or written words, pictures or maps have meaning or content, so too do (mental) representations, and the content of a mental state is inherited from the content of the representation involved in having it. One prominent question about mental representations is why they have the content that they have. One answer is that mental representations are imagistic in nature and their meaning is grounded in resemblance, so that, for example, the representation involved in having thoughts about horses means horse in virtue of resembling a horse.

Following the example introduced earlier, the act of writing a memo on a piece of paper using a pen requires the writer to internally construe these environmental objects as mental images that represent them by virtue of some resemblance relationship. This resemblance is what gives the mental representations of paper and pen their content. But having a mental representation is not enough to perform a cognitive act. The writer must also have a mental state, e.g., the belief that the pen is a tool for writing, which allows her to perform certain actions, e.g., using the pen to write.

Both mental representationalism and computationalism were unified in a program known as functionalism, or computer functionalism. This program was mainly developed by Putnam

(1960). Its central claim was that mental states such as beliefs or desires are computational states of the brain. The realization of mental states was relative to the functional organization of mental states, that is, how they relate to each other, and how the brain physically implements them was less relevant to functionalists. Functionalism unified mental representationalism and computationalism by treating mental states both as functional states or computational operations that manipulate syntactic structures according to certain rules, and as semantic structures that guide cognitive behavior. Mental representations guide behavior because the content they carry is not only about the appearance of objects, but also about the function they perform. For example, the mental representation of a pen is not just an internal image of the physical appearance of various pens, but also the belief that the pen is a tool for writing.

Finally, although functionalism holds that the material realization of cognition is to some extent irrelevant and that what matters is the internal organization of cognitive processes, there have been many accounts of mental representations that attempt to naturalize mental content, that is, to explain how it is physically realized. In these accounts, what has been called the syntactic structure of mental representations has a corresponding physical instantiation in a representational vehicle. To give an example, according to Poldrack (2020), the fusiform area enables object perception by associating sensory information with mental categories of objects. In this case, the mental categories and sensory information would act as representational content, while the brain areas involved in the cognitive act would correspond to the representational vehicle.

Similarly, Egan (2014, 115) defines mental representations as (...) "mediating states of an intelligent system that carry information" (Markman and Dietrich 2001, 471) and ascribes two properties to them: "(1) they are physically realized and have causal powers; (2) they are intentional, in other words, they have meaning or representational content". Egan distinguishes between the physical structure that carries the content, the representational vehicle, and the semantic or representational content carried by the vehicle. Bechtel (2008, 159) clarifies that the term representation can, in principle, refer to either the representational content or the representational vehicle:

> The term representation characterizes the relation between states within the information processing system, (...) referred to as the vehicle of representation, and external objects or events, referred to as the content of the representation. The term representation refers to both the vehicle and its content.

To recapitulate, representationalism in the cognitive sciences is the thesis that the mind forms internal representations of the environment to guide behavior. These representations possess syntactic and semantic properties. Their syntactic structure, embedded in any

representational vehicle, is analogous to symbols, maps, or mental models, as postulated by PP. Their semantic properties are responsible for endowing the representation with content, for example, by making a symbol stand for an environmental object. In what follows, I will elaborate on why mental representations are considered internal structures.

Internal Representations

In the cognitive sciences, internalism is the view that cognitive processes or mental states are primarily caused by internal factors, such as neural processes and other mental states, rather than environmental factors. Internalism is also a thesis about the boundaries of cognitive systems — that is, an explanation of how they are physically realized. According to the internalist hypothesis, the boundaries of cognitive systems extend to the brain or to the brain and body. This view excludes the possibility that the environment plays a role in cognitive processes except as the source of inputs. The content of mental states, such as beliefs or desires, is determined by their relation to mental representations.

A methodological implication of internalism is that an adequate explanation of cognition can be provided by focusing on brain activity and inner mental representations alone. The position that only neural or brain states give rise to cognitive processes is currently being advanced by proponents of neural representationalism (Bechtel 2016). An analysis of this theory serves to illustrate the understanding of mental representations within the context of internalist views. Thomson and Piccinini elaborate on the concept of neural representations within this theory, as follows:

> When neuroscientists discuss sensory representations, they mean activation patterns in the nervous system that carry information about the current environment, are part of a broader mapping between neural states and states of the world, and help guide behavior. For instance, activity in the topographic map of the visual environment contained in the primary visual cortex may carry the information that the traffic light has just turned green, and this activity is used by the brain to decide to do things like pull the foot off the brake pedal (2018, 198).

Neural representationalism asserts that mental states are correlated with specific patterns of neural activity in the brain that encode the representational content. Within neural representationalism, mental representations are patterns of brain activations that are triggered by environmental states. The shape, size, and color of what is being represented—in the example above, a pen or a piece of paper—are encoded in the activity of the brain. In neural representationalism, the representational vehicle of mental representations are brain or neural states.

Do activation patterns in the brain, triggered by environmental objects and states, represent anything? Not necessarily, because correlation does not imply representation. Hutto and Myin (2014) offer an alternative critique, suggesting that the theoretical construct of neural representation may lead to the "constitutive-explanatory fallacy." In short, this is the tendency to conflate theoretical descriptions of phenomena with potential explanations. Scientists who assume that neural representations map something in the world commit this fallacy by conflating a scientific description of brain behavior that invokes neural representations with an explanation of that behavior. These are examples of the numerous critiques that the notion of neural representation has received. The prevailing philosophical view is that neural representations are theoretical constructs (see Sprevak 2013; Thomson and Piccinini 2018).

Neural representationalism is one of many attempts to naturalize mental content — that is, to identify the physical structures responsible for producing mental representations. While this approach has the potential to clarify the origin of mental content, it cannot explain why specific cognitive content is ultimately acquired. This is important because content is acquired selectively rather than indiscriminately. Cognitive beings collect and process information relevant to their survival and adaptation to their respective environments.

Teleosemantics (Millikan 2009) addresses this explanatory gap. This theory posits that the content of mental representations is a consequence of the evolutionary history of agent-environment coupling. Consider, for example, the perception of snakes and spiders as threatening and dangerous, which often elicits a fearful response. According to Teleosemantics, this response is explained by the evolutionary history of interactions between humans and these species. Specifically, the fear that these animals evoke in humans is an evolved psychological protection mechanism. As demonstrated by Hoehl et al. (2017), this psychological protection mechanism manifests in various forms beyond phobias. For instance, the researchers observed that infants exhibited increased pupil dilation when shown images of spiders and snakes. This dilation does not occur when infants view images of other animals or objects.

Although Teleosemantics provides a compelling explanation of the evolutionary and historical origins of internal representations, it neglects to address the role of culture in cognition. This significant omission has been highlighted by Gardner (1985) as a dimension of cognition that has been systematically ignored in scientific research. Among the first cognitive scientists to suggest a role for cultural artifacts in cognitive processes was Hutchins (1995). He argued that representational content is not an internal phenomenon but is instead directly accessible through external representational media. For instance, the

content of maps is not stored internally because it is already accessible through these external artifacts, which facilitates navigation tasks.

Before elaborating on the concept of external representations, it is important to explain why neural representationalism is an internalist perspective on mental representations. As mentioned earlier, internalism asserts that cognition occurs within the boundaries of the brain or brain-body. Internalism does not deny that cognition takes place in an environment; rather, it draws a boundary between the environment and the cognitive system. In other words, internalism does not imply that the environment has no influence on the generation of representational content. From the perspective of neural representationalism, environmental factors trigger sensory representations, making the environment a necessary condition for generating mental representations. Teleosemantics, as introduced above, is not internalist because it considers agent-environment coupling to be the fundamental unit of analysis in the sciences of mind. As will be discussed below, externalist views either assume that the content of mental representations is determined by environmental factors or that the environment itself is a constitutive part of the cognitive process.

External Representations

There are several externalist perspectives within the cognitive sciences. Among them, I will now discuss distributed cognition, which posits that the cognitive system is distributed across minds and the cultural and physical environment (Hutchins 1995). Some externalists argue that the mind itself is partly in the environment. Extended cognition, which will be discussed in connection with enactive, embodied, and embedded views of cognition, is the view that the environment is a constitutive element of the cognitive process. Although both views are compatible, distributed cognition provides a framework for conceptualizing mental representations as external to the cognitive agent, whereas extended cognition is often situated within a broader non-representationalist research program in the cognitive sciences. Although this distinction is imperfect, it provides a rationale for discussing distributed cognition at this point and presenting extended cognition afterwards.

It should be noted that externalist views of cognition are not necessarily anti-representationalist. The hypothesis that environmental objects and other minds are part of the cognitive process can be framed in a representationalist fashion. In a navigation task, for instance, a cartographic map may serve the function of representing the territory. In this case, the necessity of an internal representation of the territory performing an analogous function to that of the cartographic map is questionable, given that the external artifact already fulfills the representational function that enables the agent to orient herself in the

territory. Nevertheless, distributed cognition does not rule out the possibility that internal representations may still play a role in navigation and other cognitive tasks.

Distributed cognition posits that information and knowledge are distributed "between and across internal and external representations" (Zhang and Patel 2006, 334). It is important to note that distributed cognition is not a thesis about the distributedness of agency, consciousness, or thinking. Rather, it is the scientific assumption that to investigate the behavior of certain cognitive systems, it is necessary to examine how information is distributed and propagated across its components (Zhang and Patel 2006, 334). In distributed cognition, information is offloaded into environmental structures, which can facilitate external representations and enable various cognitive processes. The following example illustrates this view:

> Let us consider multiplying 965 by 273 using paper and pencil. The internal representations are the meanings of individual symbols (e.g., the numerical value of the arbitrary symbol “5” is five), the addition and multiplication tables, arithmetic procedures, etc., which have to be retrieved from memory. The external representations are the shapes and positions of the symbols, the spatial relations of partial products, etc., which can be perceptually inspected from the environment. To perform this task, people need to process the information perceived from external representations and the information retrieved from internal representations in an interwoven, integrative, and dynamic manner (2006, 334-335).

This example indicates that external representations can be directly perceived from the environment. The same can be said of cartographic maps, which provide a perceivable representation of a territory. Similarly, it can be argued that the appearance and shape of an object are indicative of its function. In consequence, rather than being internally represented, such a function can be directly perceived in the object. In light of externalist accounts such as distributed cognition, it can be argued that the external format of a representation is of paramount importance for it to perform its cognitive function. The "representational effect" may be taken as an example:

> The *representational effect* refers to the phenomenon that different isomorphic representations of a common formal structure can cause dramatically different cognitive behaviors. One obvious example is the representation of numbers (...). We are all aware that Arabic numerals are more efficient than Roman numerals for multiplication (e.g., 73×27 is easier than LXXIII×XXVII), even though both types of numerals represent the same entities–numbers (Zhang and Norman 1994, 271).

The example of the representational effect demonstrates that the format of external representations has a significant impact on the way cognitive tasks are performed. If this is the case, it follows that cognitive agents develop distinctive ways of doing, thinking, and imagining when they interact with external environmental objects (see Fingerhut 2020).

The internalist and externalist perspectives on mental representations need not be mutually exclusive. However, they address distinct facets of cognitive processes. For internalist views, the representational vehicles of cognition are understood to be inner content-bearing structures. In contrast, externalists posit that these representational vehicles can also be distributed in the environment. Ultimately, externalist perspectives contend that the form and format of external representations enable distinctive ways of cognizing that the mind cannot accomplish solely through internal processes (Kirsch 2010).

Embodied, Enactive, Extended and Embedded Cognition

The research program known as Embodied, Enactive, Extended, and Embedded Cognition (4E Cognition) began in the 1990s with the publication of *The Embodied Mind* by Francisco Varela, Eleanor Rosch, and Evan Thompson. Since then, 4E Cognition has gained prominence within the cognitive sciences. One could argue that, in contemporary cognitive science, it is as relevant as computation- or brain-centric approaches. What follows is a focus on the specific meaning of each core notion of the program and how they motivate a new conceptualization of cognition together.

To illustrate the meanings ascribed to the terms "enactive," "embodied," "embedded," and "extended cognition," Carney employs the following example:

> Consider my planning of my monthly finances using an electronic calculator. Depending on the 4E theorist, this process can be described as embedded cognition (it facilitates thinking by causally exploiting an object in the environment) or extended cognition (the calculator constitutes part of my cognitive apparatus). However, the action also enacts a world: by budgeting for the future, certain items in the environment are disclosed to me as affordable or not affordable. Finally, my physical and cultural embodiment as a human being shapes my cognitive processes, such that I am most comfortable using decimal arithmetic (I have ten fingers) and will avoid thinking with degrees of precision greater than is practically useful (enculturated knowledge) (2020, 77-78).

This example shows that 4E cognition is incompatible with two classical cognitive science assumptions: that human cognitive processes are driven by the internal organization of mental representations, and that the material realization of cognition is irrelevant. In 4E

approaches, the cognitive agent is situated in a meaningful world wherein environmental and cultural structures are either causally relevant to the cognitive process or elements of the cognitive system. In the aforementioned example, an electronic calculator relieves the user of the burden of performing complex mathematical calculations when planning a budget only if the action is culturally and bodily situated.

Cultural practices facilitate acquiring the skills necessary for using calculators, including the habit of using them for tasks such as budget planning. Furthermore, inhabiting a cultural environment is essential for developing the manual dexterity necessary for using a calculator. The 4E cognition program challenges the functionalist assumption that the material realization of cognition is irrelevant. Rather, the program emphasizes that cognitive processes are only possible through the coupling of the brain, body, and environment.

In *The Embodied Mind*, Varela, Rosch, and Thompson challenged the dominance of computationalist and representationalist perspectives on cognition within the field. They proposed an alternative enactive approach to cognition, which is based on the following ideas:

i. Cognition is not confined to the brain. It is not merely the computational processing of abstract semantic structures, either.
ii. The unity of cognitive systems and processes lies in the interaction between the brain, body, and environment.
iii. Cognitive structures emerge from interactions between organisms and their environments.

Having explained 1 and 2, the discussion will now focus on emergence. Before proceeding, it is necessary to clarify one point. Enactivism distinguishes cognition from action (Aizawa 2017 argues the opposite). This is because embodied action is not merely the production of an observable output. Rather, it is a movement that creates new meaningful environmental structures, thereby opening new possibilities for action. Enaction is as an action that instantiates a new realm of possibilities for a cognitive agent. For example, the skillful use of an electronic calculator to plan one's monthly finances results in the creation of a cultural artifact called a "budget" that provides the agent with a means of predicting what future costs should be avoided. This artifact thus serves to guide the agent's future actions in ways that would otherwise remain unseen.

Although the planning of finances illustrates how cognition is enactive, embodied, extended, and embedded, it is perhaps a less representative example of enactivism in its original formulation. According to this formulation, cognition occurs in the interaction of the agent and the environment. While the agent's movements generate environmental structures, the

generated environmental structures enable the agent to perceive and act in distinctive ways. This initial formulation was guided by a focus on the coupling of perception and action. In particular, the idea was that perception "is best understood not as some purely inner event but as something that is enacted via an agent's skilled sensory motor behavior" (Clark 2014, 212).

Clark refers to the original formulation of enactivism as the **strong sensorimotor model of perceptual experience**, attributing this view to Alva Noë:

> Perceiving, according to Noë (2004), is like painting. Painting is an ongoing process in which the eye probes the scene, then flicks back to the canvas, then back to the scene, and so on, in a dense cycle of active exploration and partial, iterated cognitive uptake. It is this cycle of situated, world-engaging activity that constitutes the act of painting. Seeing (and more generally, perceiving) is likewise constituted (Noë claims) by a process of active exploration in which the sense organs repeatedly probe the world, delivering partial and restricted information on a need-to-know basis. It is this cycle of situated, world-engaging, whole animal activity that is the foundation, on Noë's account, of perceptual experience (2014, 212-213).

Perception, in analogy to painting, requires an active exploration of the environment through the various sensory organs, which is determined both by the sensorimotor skills of the agent and by the purposes that motivate the cognitive act. If we consider, for example, the skilled act of typing on a computer keyboard, we can see that visual perception is directed towards the screen, while tactile perception is focused on the keyboard. Such an act depends on the writer's skills; in the case of someone who is not used to typing on a keyboard, the task is not only to type on the keyboard, but also to learn how to operate this device. Because of this difference in purpose, visual perception is not directed exclusively to the screen, since the user must visually locate the keys corresponding to the symbols to write correctly. Then, depending on the purpose of the cognitive act, as well as on the agent's sensorimotor skills or, so to speak, the coupling between the agent and the environmental artifact, cognition is performed in one way or another.

In summary, enactivism understands cognition as a kind of action that is situated in an environment and creates meaningful structures in it. In this sense, cognition is not a response to a stimulus, but an interaction with a meaningful environment. Given this dynamic character, cognition cannot take place only in the brain but requires a constant shaping of the environment through action. Continuous actions on the environment create patterns that regulate or coordinate future interactions, i.e. they stabilize into new cognitive structures that modulate future actions.

Embodied cognition (EC) is often, but not always, understood as a complementary thesis to enactive cognition. In what follows, I present a general understanding of this thesis within the framework of 4E cognition. The embodied cognition thesis has also been developed independently of this framework. Shapiro (2007) describes EC as a general research program in the field that emphasizes the role of the body in explaining cognitive abilities. He further distinguishes three research aims of EC:

> First, steps in a cognitive process that a traditionalist would attribute to symbol manipulation might, from the perspective of EC, emerge from the physical attributes of the body. Second, rather than conceiving of cognition as the churnings of a brain isolated from the body and environment in which it is situated, EC might attempt to account for the content of cognition by appeal to the nature of the body containing the brain. Third, instead of viewing cognition as beginning with stimulation of afferent nerves and ending with signals to efferent nerves, cognitive processes or states might extend into the environment in which the organism lives (Shapiro 2007, 340)

The common emphasis on the role of the body in cognition is often associated with anti-representationalism. However, this is merely a strategy for explaining the body's role in the cognitive process, either as a constraint or enabler of various actions. Shapiro (2007, 340–341) examines auditory perception as an example. He notes that, while "larger distances between ears provide greater auditory acuity," the density of the matter between the ears because sounds of varying frequencies will behave differently when traveling through a given medium." These constraints are relevant to how the auditory system processes information without symbolic representations. In this case, the auditory system does not require a representation of the distance between the ears because their actual physical distance enables auditory accuracy.

Extended and embedded cognition are other two research programs in the cognitive sciences. For extended cognition, mental processes take place not only within the brain and the body, but also involve other minds, environmental structures, and artifacts. Extended cognition differs from distributed cognition or content externalism, as introduced earlier, in that it argues not only that the environment is causally relevant to the content that minds acquire, but that it is constitutive of the cognitive system (Clark and Chalmers 1998). The thesis, then, is that cognition itself extends beyond the mind insofar as the environment shapes or gives its distinctive form to the cognitive act.

What does it mean for cognition to occur beyond the boundaries of the brain and the body? According to extended cognition, environmental objects and structures such as tools, social institutions, and other minds actively contribute to cognitive processes. A prominent example provided by Clark and Chalmers is the "Otto and Inga" experiment, in which Inga

relies on her biological memory to remember where a museum is located, while Otto, who has Alzheimer's, relies on his notebook to find the museum's address. Clark and Chalmers argue that Otto's notebook is an integral part of Otto's memory system, just as Inga's biological memory is an integral part of her cognitive machinery. The example illustrates how an external tool can be integral to a cognitive task, extending it beyond the brain.

Finally, embedded cognition is similar to extended cognition in that it emphasizes the role of the environment in cognition. In embedded cognition, however, cognition is offloaded into environmental objects and structures, rather than the environment itself being part of the cognitive process. Embedded cognition asserts that the body and the physical and social environment facilitate cognition. An example is the arrangement of kitchen utensils. Not only does the arrangement of these tools support a cognitive task, but it also arguably speeds it up in time-sensitive situations, making cooking more efficient. It also eliminates the mental effort of remembering where they are. People who cook regularly place the tools and ingredients they often use, such as spatulas, oil, or knives, in easy-to-reach places for smooth cooking.

As we have discussed, 4E cognition is a research program in cognitive science that suggests that cognition occurs in the interplay of brain, body, and environment. Enactive cognition conceptualizes cognition as a form of action; embodied cognition focuses on the body as an enabler and constrainer of cognitive processes; embedded cognition stresses the ways in which cognition is offloaded into environmental structures; and extended cognition considers environmental structures and artifacts as elements of cognitive systems. In the next section, I explain how PP has positioned itself within the debate over mental representationalism and 4E cognition, to discuss later whether it favors either of these views.

4.2. Predictive Processing vis-a-vis Mental Representationalism

In my view, the most optimistic view on the potential of PP to solve the mental representationalism vs. 4E cognition debate in cognitive science can be found in Allen and Friston (2018). According to them, PP argues that cognition is not only based on internal representations, but also functions through active engagement with the environment, which they call 'active inference'. Allen and Friston (2018) believe that PP is compatible with 4E cognition because, for both, the mind is distributed in brain-body-environment interactions and predictions improve over time in response to environmental states. For them, however,

PP also entails mental representationalism because it seems to rely on internal or cortical generative models that map environmental structures.

I find this view optimistic because it suggests that PP bridges the gap between internalist and externalist views of cognition, providing a unified model for the study of cognition. Before discussing this view in light of my ontological agnosticism, I first discuss the inconsistencies of thinking of generative models as internal representations and then examine the concept through the lens of 4E cognition.

In Chapter 2, I have argued that generative models are strictly a type of scientific model used in research to investigate various target cognitive systems by framing them as probabilistic inference tasks. Generative models of the brain conceptualize it as a system that not only responds to stimuli but also predicts what will be encountered in perception or what the outcomes of actions will be. For internalist, generative models are mental representations gather information of the structure of an environment. In this section I will explain and critique this interpretation of generative models as representations, which is known in the literature as Conservative Predictive Processing (hereafter CPP).

CPP is a representationalist and internalist interpretation of the concept of generative models. It is representationalist because it considers generative models to be mental structures that map or represent environmental structures. It is internalist because it assumes that generative models are functionally realized in brain organization, i.e., mental structures analogous to mental states and neural representations. Proponents of CPP are therefore realist about generative models and assume that the science of PP aims to discover how predictive processing routines are actually implemented in the brain.

For CPP, a generative model can make accurate predictions if it correctly maps the structure of the environment (Gładziejewski 2016). CPP thus interprets generative models in light of the cognitivist ideas discussed earlier, suggesting that predictive routines rely mainly on internal models that create representations of reality that improve over time. Specifically, they become more accurate by feeding the model with prediction error signals.

CPP entails internalism because it suggests that the functional organization of generative models is physically realized in brain processes. This also implies that the mental content that cognitive agents use to make predictions is contained in the brain. Gładziejewski (2016) compares the mental content of generative models to the content of maps used in navigation tasks. This analogy suggests that accurate predictions depend on the ability of mental structures to map environmental structures. If the model is accurate, it will enable good predictions for navigating territory. The difference between mental models and cartographic maps is that the former can be improved by learning from the mismatch

between the two structures. In the long run, thus, a generative model becomes more accurate. In CPP, minimizing prediction error amounts to increasing the representational accuracy of the models. In Gładziejewski's words:

> Intuitively speaking, prediction error is most successfully minimized if the system selects a correct hypothesis about the causal etiology of incoming sensory data and uses this hypothesis as the basis for its top-down predictions. For example, assuming that what one is in fact observing is a tiger, then the hypothesis that the inflow of sensory information has been caused by a tiger will generate (on average) smaller prediction error than alternative hypotheses-including any that attribute the causal origins of the incoming signal to a plush toy or a domestic cat. Importantly, the size of the prediction error is always modulated by the precision that the system attributes, within a given context, to sensory signals coming from the world (2016, 562).

This example captures the specific way in which generative models are representational. Accuracy is a property of representations. Another property is complexity. In the example above, a simple model (one with few parameters) cannot serve to predict that the person is observing the tiger. The representation is complex as it requires a concept of tiger, a concept of observation, and a concept of the situation (it is most likely that one is observing a tiger in the zoo than on the streets). It is only thanks that generative models are complex systems that they can serve to make predictions. A complex map of a territory, for example, can be detailed in that specifies every single relevant place in a territory. However, if it is too complete it may make navigation slower since the person would need more time to plan the route.

Though Gładziejewski's analogy is problematic as it involves two ontologically different representations, that is, external and internal representations, it is nonetheless useful for understanding in what sense generative models can be representations. Both cartographic maps and generative models are "(a) structural representations that (b) guide their users' actions, (c) do so in a detachable way, and (d) allow their users to detect representational errors" (2016, 566). Gładziejewski (2016) suggests that these four properties make generative models truly representational. I first present them to then raise objections to CPP.

(a) Generative Models as Structural Representations

Cartographic maps are structural representations that track the relevant structures of a territory. The nature of a navigation task determines what is structurally represented by the map. For example, while tourist maps highlight the landmarks of a territory, urban traffic maps of the same territory focus on features such as transportation connections. Thus, maps structurally represent the territory in different ways depending on their purpose. Similarly, generative models are structural rather than complete representations. To be

accurate, they need to capture the environmental information relevant to specific cognitive tasks, not all its complexity. They must also track relevant changes in the environment so that agents can perceive and act efficiently.

(b) Generative Models as Action-Guiding Representations

Cartographic maps are external representational artifacts that guide navigation tasks. Similarly, generative models are internal representations that guide perception and action. To do so, they must be accurate representations.

(c) Generative models as Detachable Representations

Generative models are detachable because they can serve multiple cognitive tasks (detachability is understood here as reusability). Since they are detachable, the agent can change its model if it does not serve to make good predictions.

(d) Prediction Error as Representational Error

If generative models are representations, then they can misrepresent. They misrepresent when they trigger error signals. Error signals are representational errors that feed the system to be more accurate in the future.

Objections to Conservative Predictive Processing

This section presents objections to the internalist, structuralist, and representationalist approaches to CPP.

(a) Contra Internalism

For internalists, cognition is brain-bound. This implies that environmental structures such as social institutions, other minds, and artifacts are not cognitive. For example, Gładziejewski (2016, 566) characterizes cartographic maps as ''a particular type of prototypical (and non-mental) representational structure.'' Thus, for generative models to be mental structures, they must be realized within the boundaries of the brain. In line with this, Williams notes that the term "models" is used in a literal sense in this context, referring to "physical structures that structurally resemble their targets" (2018, 164).

Internalism considers the content of generative models to be knowledge stored in the brain:

> In any case, what might it mean to say that an organism or system learns and deploys a model? Clark (2016) everywhere advances the idea that cognizers acquire knowledge that can be appealed to in order to causally explain their subsequent behavior. Clark (2016) talks of model-driven systems in the following terms: as "using stored knowledge" (6); as "knowledgeable consumers" (6); as acquiring "bodies of

> knowledge" (68); as something that "learns and deploys a generative model" (22); as building up "the sensory scene using knowledge" (25, 17); as "the brain using stored knowledge to predict" (27); of successful perception requiring "the brain to use stored knowledge and expectations" (79); and so on (Clark 2016 as cited in Hutto et al. 2020, 179)

Clark (2016) adheres to an internalist view of knowledge. Specifically, he suggests that generative models rely solely on stored internal knowledge, rather than environmental information, to make predictions. However, information is often found in the environment or environmental objects. Using a map for navigation tasks, such as locating a specific place, is an instance of this. In this example, information stored in an external representation is crucial to accomplishing the task.

Internalist versions of PP mistakenly assume that generative models must be internal mental entities realized in the brain that map environmental structures. Hutto et al. (2020) express skepticism about the need for generative models to explain cognition on the assumption that these models are internal mental structures. The point is that an internalist must provide evidence for how the brain physically implements generative models. The problem, however, is that there is currently no explanatory account of such realization. As Hutto et al. note:

> To take such talk [that about mental models] seriously as part of an explanation, we are owed the details of precisely what changes in the system occur when it acquires a model – when it learns a model – and how such structural changes amount to the acquisition of knowledge that can causally explain and drive the relevant responding. There is a reason to doubt that this can be achieved (Hutto and Myin 2017, 26–53). Yet without a substantive proposal about how acquired knowledge can causally explain behavior, talk of the system's learning and using models provides no explanation, only the illusion of one (2020, 179).

In my view, Hutto et al. (2020) are correct in not viewing generative models as labels for mental structures, but rather as a heuristic for investigating open problems and questions. This heuristic does not itself specify whether cognition is realized in the brain, body, or environment. According to Hutto et al. (2020), concepts such as mental models must be part of a detailed causal description of behavior to be considered to be part of a serious science, a standard that PP does not meet. Hutto et al. thus conclude that the idea of mental modeling has no scientific value. Nonetheless, the case studies in Chapter 3 suggest that PP, when understood as a modeling framework rather than a theory of mental modeling, provides scientists with distinctive epistemic possibilities.

(b) Contra Structuralism

Within CPP, structural representationalism is the idea that mental generative models structurally represent environmental structures relevant to perception and action (Williams 2018). Specifically, it posits that:

i. generative models are structural representations.
ii. the cognitive agent's niche or environment has structures.
iii. the structure of generative models maps the structure of the environment.

Each of these assumptions is controversial. Let me first elaborate on the idea of mapping. According to Williams (2018, 166), "predictive brains exploit cortical activity as a stand-in for the ambient environment with which to anticipate its sensory effects and support viability-preserving intervention." In his view, it is the capacity of generative models to map relevant environmental structures that makes accurate predictions possible.

In my view, an inconsistency of the structuralist view is that it appeals to a notion of objective environmental structure while simultaneously arguing that the structure captured by mental representation is one relevant to perception and action. In other words, the structuralist is committed to the idea that there is an objective representation of the environment, while at the same time arguing that this representation is relative to the interests and practical skills of the cognitive agent.

Moreover, if the idea that the represented objective environmental structure is relative to the interests and practical skills of cognitive agents is taken seriously, it follows that the positing of two structures is unwarranted. This is because an environmental structure is relevant only if there is a cognitive agent to whom it is relevant. If this is the case, then the subjective representation of the agent already contains the relevant structure of the environment. Put more simply, the notion of relevant structure is incompatible with the notion of objective structure.

One could accept this but still argue that environmental structures are relevant to a single agent, but to many of them, and thus are objective as public entities existing in the world. An example might be social institutions that are relevant to social life. They are objective in the sense that they have publicly and tacitly known rules and organizations that do not depend on the individual's representation of them. Even in this case the mapping from internal to external structure seems unwarranted, since the structure of social institutions would be embedded in the environment and thus not need to be represented internally. I discuss this notion of embeddedness in the following critique of CPP's representationalism.

(c) Contra Representationalism

So far, we have taken representationalism in PP to be the view of generative models as internal mental structures that perform representational functions that depend on the capacity of these models to map environmental structures relevant to perception and action. Representation, in this view, is a necessary condition for accurate predictions. Critics of CPP are not necessarily opposed to concepts such as prediction or model, but question whether they must be representational. Orlandi (2018, 2383), for example, distinguishes detectors (of which predictions are an instance) from representations on the grounds that

> carrying information is not sufficient for representation. Mere detectors, or trackers can carry information in some naturalistic sense, without being representations. Detectors are typically understood as states that monitor proximal conditions and that cause something else to happen. As such, they are mere intermediaries that are found in many biological and chemical systems. What is important about detectors is that they are active in a restricted domain, and they do not model distal or absent conditions. The behavior of a system that uses only detectors is mostly viewable as controlled by the world. What the system does can be explained by appeal to present environmental situations. Magnetosomes in bacteria and cells in the human immune system that respond to damage to the skin are examples of detectors of this kind.

Orlandi is right to question the need to postulate mental representations without a clear specification of a representational function. Predictions, he suggests, can be understood as detectors with non-representational functions. Myin and van Es (2020) go even further, suggesting that the notion of representation is unwarranted in PP. They elaborate a critique in response to Gładziejewski's (2016) analogy of generative models and cartographic maps, which for them conflates internal and external representations. A shortcoming of this analogy is that

> [...] neural models are thought to precede the practices that representations are a product of. This marks a clear difference between cartographic maps and neural models, and it is due to this analogy that cartographic maps may be said to represent a target terrain when being used in the relevant social practice, yet neural models may not lay claim to the same representational status (2020, 16).

According to Myin and van Es (2020), the concept of representation is inappropriately used in neural representationalism. However, this does not seem to be the case with generative models. Still the analogy does not explain why generative models are representational. To explain how generative models make predictions without relying on the concept of representation, Myin and van Es (2020) turn to embedded cognition. This theory posits that cognitive agents can make predictions because they are sensitive to environmental

regularities. These regularities exist because cognitive agents transform environments into niches. In this context, generativity is not a property of internal states but of environmental structures created through agent-environment interactions. For the embedded view, environments become predictable because sociocultural agents systematically model them. Importantly, environmental structures not only model the environment, but also constrain the agency of cognitive beings.

A final critique of the representationalist view of generative models comes from the philosophy of science. As noted in previous chapters, generative models are first and foremost scientific models. As such, they are representational only if they are constructed and manipulated in a way that allows for representation. In other words, representation is not a property of scientific models, but an accomplishment of modeling practices. Models whose target systems are generic cognitive behavior, such as perception or attention, are not representational unless we understand representation as the instantiation of a conceptualization in a particular diagrammatic, written, or symbolic medium.

Challenging the understanding of generative scientific models as representations, Tee (2020) proposes that generativity is the ability to create new structures or models based on data. Thus, rather than representing reality, generative scientific models function as constructive tools for scientists that allow for various other epistemic possibilities rather than representation, as the cases of the AON and AECS models presented in the previous chapter have shown. Regardless of whether generative models are viewed as mental structures or scientific artifacts, the notion of representation seems insufficient to explain their generativity. In the next section, I present and discuss views of generative models that do not appeal to representations to account for their generativity.

4.3. 4E Cognition and Ontological Agnosticism

Predictive Processing vis-a-vis 4E Cognition

In my view, conservative Predictive Processing is untenable because of its internalist, structuralist, and representationalist assumptions. Furthermore, CPP's mental modeling realism fails to acknowledge that generative models are epistemic tools. Generative models can be associated with any theory of cognition because they are generic. This section discusses how generative models have been incorporated into 4E cognition theories. In the concluding part of this section, I articulate my ontological agnosticism—the view that generative models are ontologically neutral epistemic artifacts.

Radical Predictive Processing (RPP) is an interpretation of PP that aligns with 4E cognition. RPP challenges the assumption that generative models are internal mental representations. Proponents of RPP argue that while 4E cognition offers a comprehensive account of cognitive processes, RPP provides epistemic resources for modeling and experimentation in empirically oriented research. Yet, 4E cognition provides tools for analyzing generative models as epistemic artifacts.

RPP remains a realistic interpretation of mental modeling. The difference between RPP and CPP concerns what is understood by a cognitive system and process. Unlike CPP, RPP posits that the fundamental unit of analysis in cognitive science is the brain-body-environment system. Furthermore, it treats cognitive processes as enactive, embodied, embedded, and extended, as I will now discuss.

RPP builds on enactivism to contend that perception and action are coupled. Nöe explains this idea in the following:

> In normal perceivers, sensations are smoothly integrated with capacities for thought, and for movement; so, for example, we naturally turn our eyes to objects of interest, we modulate our sensations with movement in a way that is responsive to thought and situation. A sharp sound makes us turn in the direction from which the sound emanates. A ball rushes toward us and we reflexively duck. A person speaks to us, we turn to him or her (Nöe 2004, 6).

In the case of a sharp sound, predictions serve as motor commands that direct bodily movements (Clark 2013). In this example, the perceiver's body is actively engaged in the process of perception, as evidenced by the rapid movement of the eyes in the direction from which the sharp sound originates. In the absence of an association between the sound and a potential threat, for example, when multiple thunders are heard during a thunderstorm, a subsequent sharp sound is likely to be perceived as less alarming. Consequently, the probability of a new motor command directing the gaze towards the sound is reduced.

RPP is also based on the 4E cognition assumption that cognitive structures emerge from patterned agent-environment interactions. Hipólito and van Es (2022) further elaborate on this assumption as follows:

> With bodily experience and action, infants first enact the world. They acquire simple motor skills such as walking, reaching, or kicking their legs. These tasks are learnt because infants have some motivation to reach a goal: getting across a room to grab a toy, for example [...]. Through everyday embodied actions, such as poking, squinching, banging and so on, the child gathers understanding about their movements in the environment.

An agent's practical understanding of an environment is shaped by the interactions it has with it, gradually turning the environment into a niche that offers possibilities for perception and action. Rather than requiring representations of goals, actions require sensory-motor coordination that depends on an agent's knowledge of its world.

Clark (2015b, 20) proposes the following role of generative models in RPP:

> The task of the generative model in all these settings is to capture the simplest approximations that will support the actions required to do the job – that means taking into account whatever work can be done by a creature's morphology, physical actions, and socio-technological surroundings. Such approximations are constrained to "provide the simplest (most parsimonious) explanations for sampled outcomes" (Friston, Adams, and Montague 2012). There is thus no conflict with work that stresses biological frugality, satisficing, or the ubiquity of simple but adequate solutions that make the most of the brain, body, and world.

In Clark's view, generative models are not mental representations but rather sensorimotor schemata that coordinate goal-directed perceptual tasks. His perspective is arguably aligned with enactivism, which posits that the brain plays "an important role in the ongoing dynamical attunement of the organism to the environment. In social interaction, for example, brain processes integrate into a complex mix of transactions that involve moving, gesturing, and engaging with the expressive bodies of others" (Hutto et al. 2020, 165). In conclusion, while RPP implies a mental model ontology, its concepts are compatible with 4E cognition.

As illustrated in the case of the AECS model in Chapter 3, PP offers a framework for examining the interactions between an agent and an environment as tasks of probabilistic inference. They are used to characterize the interactions of an agent that models an environment and is simultaneously modeled by it. By modeling the environment, an agent transforms it into a predictable niche a field that provides and constrains possibilities for action. As with the AECS model, generative models can be used to conceptualize the coupling of agent and environment. However, they do not specify that cognition must occur in the brain or extend into the environment. To substantiate either of these claims, modelers must make additional assumptions about the systems they are studying. It is crucial to highlight that, in contrast to CPP, a 4E perspective on generative models can also treat them as epistemic resources in the sciences of the mind, rather than as models that represent the inferential or probabilistic nature of cognition.

Generative models as Extended Epistemic Artifacts

The philosophical debates on PP have not adequately addressed the scientific nature of generative models. Instead, they have focused on questions such as whether generative models are internal or external mental representations or parts of enactive, embodied, extended, and embedded cognitive structures (Allen and Friston 2018; Hipólito and van Es 2022). 4E cognition, particularly extended cognition, provides a framework for analyzing generative models as epistemic artifacts. In this framework, generative models extend the cognition of scientists by constraining and affording distinctive ways of thinking on systems, addressing questions, and solving problems. For an overview of the 4E approach to scientific practice, see Sanches de Oliveira, van Es and Hipólito (2023).

Extended cognition does not inherently conflict with representationalism. As discussed earlier, distributed cognition recognizes that external tools and artifacts, such as cartographic maps, can have representational functions. Whether these representations are genuinely cognitive depends on the definition of cognition used. According to CPP, external representations are not cognitive because they are not part of the brain or body. In contrast, for extended cognition, environmental factors are genuine components of cognitive systems; thus, epistemic artifacts, such as scientific models, are also cognitive too.

Cartographic maps illustrate how epistemic artifacts constrain and enable cognitive processes. First, these maps' cognitive function depends on the context of their use, such as in navigation tasks. Second, using the artifact appropriately constitutes an embodied act because it requires the agent to visually perceive and manually manipulate the artifact. Third, information is embedded in the artifact rather than stored in the brain. It is accessible only to an agent who is enculturated and knows how to use maps. Finally, the success of a navigation task depends on factors beyond the agent. Even if the agent is skilled at using maps, they may fail to reach the intended destination due to the map's inaccurate representation of the territory.

Artifacts facilitate or impede agency based not only on how they are manipulated, but also on their design and how they are constructed. In the case of cartographic maps, specific visual features such as scales, symbols, and grids serve to guide navigational tasks. In this sense, artifacts "embody kinds of knowledge that would be exceedingly difficult to represent mentally" (Latour 1986, as cited in Hutchins 1995, 96). They facilitate distinctive modes of action, cognition, and imagination. Navigation apps like Google Maps facilitate navigation by providing a structured route representation that users can access without needing to retain extensive information in their memory. However, Google Maps has its own set of challenges; it can potentially lead users down dangerous routes.

As these examples illustrate, the extended approach's key distinction is its view of artifacts as instruments that afford distinctive possibilities for cognition and action, not merely as external representations. Scientific models are material instruments that enable particular practices. As discussed in previous chapters, models allow scientists to conceptualize objects in ways that enable the objects to be studied or make predictions about their behavior.

In philosophical discussions about the predictive mind, generative models have not been considered epistemic artifacts that extend scientists' cognition. Instead, the literature has relied on philosophical debates such as those over mental representationalism and the reality of mental models. Even when treated as scientific models, generative models have been understood as representational instruments. Tee's account of generative models from the philosophy of science, discussed in Section 2.3, takes an alternative approach. Drawing on the idea that models provide epistemic possibilities based on their construction and manipulation, Tee (2020) challenges the notion that models are generative merely because they represent a target system. Tee defines generativity as the ability to construct a new model in scientific research based on an existing model. This perspective aligns with an extended cognition approach to these artifacts.

Tee (2020) presents the case of Peschard's constructive abstract models to illustrate the generation of new models. Peschard's work focuses on the use of abstract models of fluid dynamics as tools for generating new scientific models. By applying a mathematical term describing the continuous coupling of oscillators to the Stuart-Landau model, scientists generate the new Ginzburg-Landau (GL) model, which captures the 3D dynamics of coupled wakes. As Tee indicates,

> Peschard demonstrates this in her example of a model of a wake in fluid dynamics, showing that a 2D wake and its model are serving as a basis for the construction of a system of coupled wakes and its new model. By coordinating the original model to its 2D wake experimental system via the specification of the coupling parameters, the coordinated model-target plays the role as an instrument for the generation of a new model and the corresponding new target (i.e., a system of coupled wakes) (2020, 26).

In the cognitive sciences, generative models such as the AECS introduced in Chapter 3 have been used to mathematically simulate the dynamics of niche constructing agents. These models can then be used to study how agents learn skills by interacting with other agents and artifacts in sociocultural environments. As an illustration, one might consider the development of mathematical reasoning in children. This form of reasoning, which is essential for social interaction in modern cultures, requires the development of sensorimotor skills. Children do not learn to count by relying on internal mental

representations; rather, they are likely to use their hands and visual symbols of numbers (Moeller et al. 2012). The use of visual symbols for numbers serves as a means of facilitating the completion of counting tasks. By using these tools, children exploit the physical properties of the cultural artifacts of their niche. As Shvarts et al. (2021) note, "for mathematical actions, not only bodily potentialities but also affordances of cultural artifacts take part in the constitution of an action." These practices can be considered generative since environments and cultural artifacts model the cognitive process, and the resulting cognitive process have an impact in the sociocultural environment.

By assuming that the elements of a cognitive system (specifically, the brain, the body, and the environment) model each other, enactive, embodied, extended, and embedded cognitive processes can be seen as generative. In generative cognitive systems, the actions of the agent model the environment and vice-versa. This happens at different levels. The cognitive practices of social agents transform their niches, while environmental resources, such as artifacts and tools, shape the cognitive tasks they perform. Thus, generativity in extended cognition refers to two things: first, the activity of modeling the environment, and second, the role of the cultural environment in modeling human behavior. Cognitive agents take advantage of environmental opportunities and transform themselves by their actions. For example, changing one's diet is not enough to determine what, how much, and when to eat. Individuals also rely on environmental resources, including public information about healthy eating, media content, and applications that provide behavioral guidance. Sociocultural environments provide opportunities for action, and technological artifacts are integral to human action and interaction. As a result of their habitual use, agents become enculturated.

The idea of modeling the agent can be conceptualized within the framework of 4E cognition as a form of enculturation, defined as the process by which an individual becomes part of a culture. This process is driven by interaction with other sociocultural agents and participation in technological environments. Fabry (2018, 2483) characterizes it as a "temporally extended transformative acquisition of cognitive practices in the cognitive niche." An enculturated agent can perform distinctive cognitive tasks that, from this perspective, are enabled by the properties of the environments it models.

In summary, generative models can be conceptualized as extended cognitive artifacts that offer distinctive epistemic possibilities. In the cognitive sciences, they have been used to model cognitive processes as generative processes in which an agent models the environment, and the environment models the agent. In other words, generativity in this context implies two key aspects: cognitive systems model their environments to make them predictable, and the environment constrains the realization of cognitive processes.

Ontological Agnosticism

Ontological agnosticism in this dissertation is the view that generative models are ontologically neutral epistemic artifacts rather than mental entities or structures. As epistemic artifacts, they do not impose cognitive constraints on the systems they model. In other words, these models are strictly neutral with respect to the material realization of cognitive processes, adding arguably mathematical or computational but not cognitive constraints to the target systems of which they are models. I discuss and justify this position in this concluding part, but first I contextualize it within the broader framework of this chapter.

Ontological agnosticism is my response to the philosophical debates about generative models and the inferential and predictive nature of cognition. The views in question are realism, anti-realism, instrumentalism, internalism, and 4E cognition. The points under discussion are whether mental models are mental kinds and, if so, how they are realized. There is no point in going through the details of this debate again, especially since, as I will now argue, ontological agnosticism is a means of undermining the very need to determine which are the correct views among them.

Ontological agnosticism implies that PP tolerates competing philosophies of mind. This is because its theoretical constructs (e.g., models, predictions, inferences, prediction errors), rather than representing mental entities, processes, or mechanisms, serve the epistemic function of conceptualizing target systems in scientific practices. Since generative models in PP articulate these concepts into tangible epistemic artifacts, the characterizations of cognition afforded by their use open epistemic possibilities for exploring, manipulating, or controlling systems. However, their epistemic function is not to substantiate any particular view of cognition over another.

My justification for ontological agnosticism is as follows: Scientific models, as introduced earlier, are extended cognitive artifacts. This is also true of generative models, insofar as they are human creations that extend scientists' cognition, enabling them to conceptualize cognitive and other target systems in specific ways relative to the way these models are constructed and manipulated. Thus, generative models, like other epistemic artifacts, provide distinctive ways of reasoning about target systems.

As discussed earlier, generative models allow scientists to model target cognitive systems as generative processes, i.e., processes in which a system produces a new model based on an existing one. This is a generic and operational description of a system that has been used in the sciences to study perception and other cognitive phenomena. Rather than providing a realistic characterization of these phenomena, it serves to frame them as if they were generative. Only under this heuristic can the mathematical properties of generative models

be used to study a target system. A crucial implication of this "as if" mode of reasoning is that it does not specify the material realization of the target system. Put more simply, by framing a system as generative, scientists do not specify that the system is realized in the brain, the body, or the environment. In principle, generative models, as seen in Peschard's model presented earlier, can be used even beyond cognitive science research.

If generative models do not add cognitive constraints to the systems they are used to model, but only mathematical constraints, as Wiese (2017) points out, then the only constraints they impose is that a target cognitive system is framed as probabilistic and inferential tasks (as discussed in Chapter 2). Thus, the debates about representationalism and 4E cognition in the context of generative models are not about cognitive constraints. It follows that the notion of generative models is strictly neutral with respect to what is at stake in these debates.

My ontological agnosticism undermines the need to resolve debates about whether PP favors representationalism, internalism or 4E cognition. I have argued that generative models are scientific artifacts that serve to model target cognitive phenomena as tasks of probabilistic inference. 4E approaches to cognition help us articulate an understanding of the epistemic uses of these models in extended scientific practices. By characterizing target cognitive systems as generative, they offer a particular way of conceptualizing target systems.

From a 4E perspective, I have described the generativity of scientific models as the ability to produce new models based on existing ones (Tee 2020). In the cognitive sciences, such models have been used to model distributed cognitive systems, such as sociocultural agent-environment systems. Generativity in this context means the capacity for model building. In agent-environment coupled systems, the patterned actions of cognitive agents model the environment, transforming it into a sociocultural niche, while the structure of the niche constrains while enabling or thwarting the agents' actions. In this sense, an agent-environment coupled system engaged in niche construction can be described as generative (see Bruineberg et al. 2018).

The PP literature has overlooked this conceptual dimension of generative models, focusing mostly on their mathematical properties (e.g., in the use of Bayesian statistics in the analysis of cognitive problems) or taking these mathematical properties as structural representations of cognitive structures. My ontological agnosticism opens new ways of thinking about different cognitive systems, while preserving the possibility of using the mathematical apparatus of PP to investigate them.

The idea of generativity, as I have argued, can be helpful in modeling complex cognitive behavior, or the cognitive behavior of distributed brain-body-environment systems, or even cognitive systems that involve processes of enculturation. For example, the interaction of humans and cultural environments can be seen as a generative process, as humans are modeled by their socio-cultural niches to develop skills such as tool use, language, or mathematical reasoning. By acting in cultural environments, enculturated agents in turn transform their sociocultural niches by creating new cultural artifacts, available environmental tools whose use shape the human mind. In another context, Husserl described such processes as generative, arguing that they involve a "historical and intersubjective becoming" (Steinbock 1995, 63).

Let me conclude this chapter by suggesting a possible explanation for why generative models and other theoretical constructs in PP, such as predictions, prediction errors, or inferences, which belong to the conceptual dimension of scientific modeling, have been conflated with mental entities, mechanisms, or structures. This category mistake arguably occurs because the metaphors used to study cognition become standard or default ways of thinking about reality. The most famous example of this mistake is the mainstream view of the brain as a computing machine, which has spread beyond science and is arguably a common sense understanding of the brain in Western cultures.

From the perspective I advocate here, the above theoretical constructs are not designators of cognitive structures, but epistemic instruments that afford distinctive ways of theorizing. According to Toon (2023), humans use theoretical constructs, such as mental representations and internal states, to form beliefs about themselves and others. Theoretical constructs are epistemic instruments that facilitate the acquisition of knowledge or mediate our understanding of reality.

Interestingly, some of the concepts, metaphors, analogies, and models used in science transcend the realm of scientific cognition and become cultural artifacts. In *The Idea of the Brain*, Cobb (2020) examines the various scientific metaphors used to study the mind. These metaphors include hydraulic machines, telegraph networks, and, most recently, computers, which have been standardized in the past and present to conceptualize the mind, even outside of scientific contexts.

Cobb ascribes an epistemic function to these metaphors: "Over the centuries, each layer of technological metaphor has added something to our understanding, enabling us to carry out new experiments and reinterpret old findings" (2020, 4). Interestingly, Cobb also suggests that "by holding tightly to metaphors, we end up limiting what and how we can think" (2020, 4). In the cognitive sciences, not only metaphors, but arguably theoretical constructs and other epistemic tools, extend scientific cognition by enabling distinctive ways of thinking and

imagining cognitive systems. If generative models are metaphors for the study of cognition, it is crucial to evaluate what kinds of reasoning they enable and what they thwart. Cobb warns that there is a price to be paid for clinging to metaphors. For example, if generative models are treated as real mental structures, one may forget their properties as scientific models, conflate the properties of these models with the properties of phenomena, or worse, ascribe to them the properties of other theoretical constructs (e.g., assuming that generative models are mental representations).

The conflation of generative models and cognitive structures arguably proceeds as follows: a cognitive behavior is first conceptualized as a probabilistic inference task of a cognitive system. For example, perception, as discussed in 2.3, is conceptualized as the behavior of a system that predicts upcoming sensory signals on the basis of a model of the environment. Once the system is conceptualized in this way, and findings about perceptual phenomena can be made about it, the conflation occurs when perceptual phenomena are attributed the properties of the model. Put simply, they are assumed to be probabilistic and inferential in nature.

Proponents of the PP often recognize that descriptions like the above are not necessarily serious descriptions of the mind. However, it is often the case that ontological statements are made based on them, and this has been the case in the philosophical debates about mental modeling explored in this chapter. Strictly speaking, generative models are epistemic artifacts used in science. In the philosophical debates, however, they have been interpreted as cognitive structures, or as concepts that represent minds as model builders or predictive engines. They have also been understood as structures that support either a representationalist or a 4E understanding of cognition. By approaching them as epistemic artifacts, I have argued for an agnostic view of these interpretations, that is, I have argued that PP tolerates competing ontologies of the mind but does not necessarily support any of them.

Chapter 5 - A History of the Predictive Mind

The objective of this chapter is to support an ontological agnosticism regarding the idea that the mind makes predictions by modeling the world and itself. The argument is put forth that the notion of models in PP principally refers to scientific models and is exempt from ontological commitments to the existence of mental models. This argument is developed through an investigation of the genesis of the central concepts of PP, which are those of inference, error, and model. A thorough examination of the genesis of these concepts reveals that, in their initial formulations, they were conceived as scientific analogies or descriptions of abstract and fictional systems. That is to say, the concepts that allegedly refer to hypothetical mental models are in fact ontologically neutral. Furthermore, an analysis of the historical development of these concepts reveals their role as scientific instruments providing various epistemic possibilities, none of which consist in representing real mental entities or structures.

In Section 5.1, I examine Helmholtz's theory of perception. I address the origins of the bottom-up and top-down approaches to cognition, which challenge the view of perception as a passive aggregation of sensations, and the complementarity of mechanistic and psychological accounts of perception. Additionally, the section delves into the genesis of the PP concepts of priors and perceptual inference within Helmholtz's theory of perception. Priors, understood in the PP framework as mental contents that facilitate predictions about future perceptual contents, were conceptualized by Helmholtz as mental contents that regulate the synthesis of sensations into perceptual units. Helmholtz's theory posits that perception involves both bottom-up and top-down dynamics. While prior knowledge dictates the interpretation of objects and processes in the environment in perception, sensations provide the content that fine-tunes perception. This conceptual framework has been adopted by contemporary PP research as a central modeling assumption in the scientific study of perception.

In the remainder of this section, a close examination of the notion of perceptual inference in Helmholtz's theory is undertaken. It is argued that Helmholtz utilized perceptual inference as a scientific analogy to bridge the explanatory gap between physiological and psychological accounts of perception. This analogy enabled him to describe perception in terms compatible with a mechanistic account without committing to a literal description of

perception as an inferential process. This analysis lends support to my ontological agnosticism about mental models by demonstrating that scientific concepts can offer a serve to make workable characterizations of cognitive target systems without making assumptions about the underlying nature of cognitive phenomena.

In Section 5.2, the theory of systems of cybernetics is introduced, with a focus on the seminal contributions of Ashby. Cybernetics is characterized as a novel scientific paradigm, one that does not seek to understand reality but rather to control observer-relative systems. Additionally, the argument is made that cognitive systems are instances of cybernetic systems. This perspective challenges the established view of cognitive systems as natural or mental kinds, suggesting that they are, in fact, constructs created within specific scientific practices. As abstract systems, cognitive systems acquire their characteristics through stipulations within scientific modeling practices. A notable point is that, similar to cybernetic systems, cognitive systems are not confined to any given material realization. They function as epistemic instruments for modeling practices aimed not at representing, but at gaining control of observer-relative target systems. This section explores how the PP concepts of error, model, and active inference stem from this novel scientific approach to questions and problems.

The conclusion relates these discussions to the understanding of key concepts of PP, such as that of a generative model as an ontologically neutral epistemic tools rather than mental kinds. Mental modeling realism, following this, stems from the conflation of a tangible and intersubjective epistemic artifact with a mental internal structure with properties similar to the artifact. Whether the correspondence between their properties is because the artifact captures a mental structure is something I am agnostic about.

5.1. A History of Perceptual Inference

This section examines the origins of contemporary top-down scientific approaches to cognition in Helmholtz's theory of perception. It is argued that Helmholtz's conceptualization of prior knowledge in perception and his analogy of perceptual inference have shaped the development of top-down approaches to cognition, of which PP stands out as a prominent case.

As discussed in Chapter 2, the terms bottom-up and top-down in neuroscience and PP describe models of information processing within the brain and nervous system.[14] Both

[14] As discussed in Chapter 2, generative models in PP are organized as hierarchical systems. However, the notion of the brain as a hierarchical system is not unique to PP. A brain hierarchy serves as a scientific heuristic for exploring the organization of brain networks, in which higher cortical

models assume that information flows through a hierarchy of layers in the nervous system. In bottom-up processing, inputs from lower layers are transmitted upward in an ascending or feed-forward fashion. For example, when light hits the retina, photoreceptors convert it into electrical signals that are sent to higher brain regions. This is a typical feed-forward connection. In contrast, top-down processing involves descending or feedback connections, where higher brain areas influence and modulate activity in lower sensory regions (Rauss and Pourtois 2013). Top-down approaches suggest that in this interplay, the brain integrates sensory input with prior knowledge and expectations.

As discussed in Chapter 2, the concepts of bottom-up and top-down play a role in the psychology of perception. At the level of psychological explanation, the aim is not to describe the mechanisms that regulate the processing of sensory information. Rather, it is to explain perception as an active behavior that is shaped by expectations and experience. In this context, top-down cognition refers to the anticipations in perception that are allegedly involved in the making of phenomenal experience and action.

For Engel, Fries, and Singer (2001), the concept of top-down cognition is incompatible with prevailing neuroscientific accounts of perception, which they argue should aim to explain perception only by relying on bottom-up mechanisms. Walsh et al. (2020, 242) note the following about these standard neuroscientific accounts: "For many years, the dominant theoretical framework guiding research on the neural origins of perceptual experience has been provided by hierarchical feedforward models in which sensory inputs are passed through a series of increasingly complex feature detectors." These standard accounts assume that perception is best understood when it is conceptualized as a bottom-up process in which sensory signals ascend in a feedforward fashion from the retina to the visual cortex. For them, this is a crucial assumption in the science of perception.

In the cognitive sciences, computational and informational approaches to perception frequently adhere to this assumption. A notable example is Marr's (2010, 31) definition of visual perception as "a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information." According to Marr's conceptualization, sensory input functions as raw data that undergoes processing to eliminate irrelevant details. This processing is orchestrated sequentially and in a feed-forward manner. In Marr's theoretical framework, the initial stages of sensory processing function independently of higher-level cognitive top-down modulation of

layers are associated with more specialized and differentiated functions than lower layers. It is important to note that within these hierarchies, information is not necessarily transmitted in a strictly top-down fashion from higher to lower levels. For a comprehensive review of the concept of hierarchy in brain research, see Hilgetag and Goulas (2020).

perception, thereby rendering the process inherently bottom-up. At the higher levels of the hierarchy (e.g., the fusiform area in the case of face recognition), the brain synthesizes raw sensory data into perceptual objects according to a sequence of fixed rules.

Phenomenologists, ecological psychologists, and proponents of 4E cognition have challenged the bottom-up approach to perception (see Gallagher and Zahavi 2012; Gibson 2015; Hutto and Myin 2017). These frameworks characterize perception as an active process involving continuous interactions between the brain, body, and environment. PP also challenges the bottom-up approach, but from a different angle. Rather than neglecting the passive dimension of perception, it characterizes it as a mixed bottom-up and top-down process (Seth 2014). Its top-down dimension amounts to expectations modulating the processing of raw sensory data. In that sense, for PP prior knowledge is relevant at the earlier stages of sensory processing.

As we have seen in the previous chapter, the scientific models in PP are neither phenomenological nor psychological, but information processing models. Their purpose is to characterize the transmission of information in the brain and nervous system as bidirectional, with top-down signals modulating bottom-up sensory signals before ascending to higher levels of the hierarchy, and bottom-up signals modulating top-down predictions (Friston 2005). As Keller and Mrsic-Flogel (2018, 425) indicate, "a brain area at a higher level of the hierarchy sends a top-down signal to an area at a lower level in the form of a prediction of the bottom-up input to that area."

In what follows, I will explore how this mixed approach to cognition and perception has its origins in Helmholtz's theory of perception.

Helmholtz' Theory of Perception

Several theories in the history of philosophy and science have approached perception and cognition in a manner similar to that of PP. This section will examine the theory of perception put forth by Hermann von Helmholtz, a neo-Kantian scientist and philosopher. Helmholtz is frequently cited as the primary precursor of PP, given that the analogy of perception as a form of inference—widely considered a cornerstone of PP—originated in his work.

During his lifetime (1821-94), Helmholtz published works in psychology (on visual and auditory perception), physiology of human vision (invention of the ophthalmoscope), physics (Helmholtz free energy in thermodynamics), and philosophy. His *Treatise on Physiological Optics* was considered one of the main references in the physiology of vision in the nineteenth and twentieth centuries. The three volumes of this book were published between 1856-66 and covered topics such as depth perception, motion perception, and

color vision (Pouyan 2014; Helmholtz 1977). In the third volume, "The Perceptions of Vision," he proposed that perception is an unconscious inference.

According to Helmholtz, perception is the cognitive act of apprehending external objects as ideas. Perception is not merely a passive imprinting of a mental image triggered by sensations. Rather, as a cognitive or mental act, it involves the intelligent recognition of external objects from sensations. These perceived objects are understood as ideas or representations inferred from sensory experiences. According to Helmholtz, psychology is the scientific discipline that investigates these cognitive processes, which have been previously referred to as the top-down cognitive or mental dimension of perception (Meyering 1989).

Despite the importance attributed to this top-down dimension of perception, Helmholtz (1925, 1) insisted that perceptions are triggered by sensations, stating: "The sensations aroused by light in the nervous mechanism of vision enable us to form conceptions as to the existence, form, and position of external objects. These ideas are called visual perceptions." As psychology does not deal with the mechanisms by which sensations operate, Helmholtz argued that this science could only partially explain perception. In his view, an empirically based psychology of perception must be complemented by a physiological explanation of the mechanisms of perception.

Helmholtz's conception of a science of perception entailed the integration of two distinct approaches. In contemporary terms, he characterized it as an interdisciplinary subject in the midst of the physiological and psychological sciences. Notably, he put forth the idea that the physiological theory of perception could serve to restrict the scope of psychological explanations by enabling scientists "(...) to see clearly the connections between the phenomena [perceptions] and to arrange the facts in their proper relation to one another" (1925, 2). This approach was instrumental for him to avoid *ad hoc* assumptions about the psychology of perception. This emphasis on empirical research arguably distinguishes his approach from that of modern rationalists such as Kant.

As previously stated, while the physiology of vision is concerned with the nature of physical stimuli and the physical mechanisms of perception, the psychology of perception is concerned with ideas. The psychology of vision, for instance, examines the recognition of forms and positions of external things, i.e., the content of visual perception that is not derived from sensations. Helmholtz contended that in visual perception, the activity of the mind is to make a synthesis. This synthesis involves the integration of sensory input with prior knowledge, a process that is governed by specific rules. The most elementary of these rules is the following:

> The general rule determining the ideas of vision that are formed whenever an impression is made on the eye, with or without the aid of optical instruments, is that such objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism, the eyes being used under ordinary normal conditions (1925, 2).

According to this rule, the synthesis of prior knowledge and sensations is an act of imagination, whereby the mind imagines that something is present in the visual field to infer that something in the world is causing the sensory impressions. This act of imagination only works when the organ of perception works under normal conditions; however, many factors can alter these conditions. A practical example is pressing the finger on the upper left or right side of the eye. This action may result in the perception of a dark circle, yet it may not lead to the recognition of an external dark circle in the world that is causing the sensations. According to Helmholtz, this discrepancy arises given that the attribution of reality to external objects is a judgement, as opposed to a mere passive association of sensory impressions.

Helmholtz distinguished between the mode and the qualia of perception, defining qualia as the subjective experiences of sensory phenomena, such as the perception of sweetness, redness, or brightness. The mode of perception, in turn, encompasses the broader framework within which qualia manifest, including the visual, tactile, or acoustic modes. Both qualia and modes are not properties of external objects. Rather, they are the result of the interaction of external objects and sense organs, as Helmholtz (1977, 121) himself explained, "our sensations are, in fact, effects produced in our organs by external causes; and how such an effect is expressed depends, of course, quite essentially upon the nature of the apparatus on which the effect is produced."

Similar to the PP framework, Helmholtz's theory of perception is a mixed top-down and bottom-up approach to perception. In the following discussion, I will address how this theory explains the phenomenon of transforming sensations into perceptions. Helmholtz conceptualized this psychological process using the scientific analogy of inference.

Inference as a Scientific Analogy

Helmholtz's theory of perception stands in opposition to associationist accounts by treating it as a bottom-up and top-down phenomenon. In simpler terms, for Helmholtz's, the mind is active in perception, and to account for the psychology of this act, he contended that minds rely on past experiences to make inferences about the causes of sensory impressions. The following exposition will elaborate on the concept of perceptual inference and discuss its

implications for the interpretation of inference in PP. It will be argued that this concept is a scientific heuristic, rather than as a designator of a real property of cognitive phenomena.

Helmholtz's theory of perception delineates instances in which sensory organs may constrain perception, potentially resulting in misjudgments of perceptual contents. However, Helmholtz argued that false perception does not stem from a malfunctioning of the sensory organs. Rather, it is the result of a misjudgment of the material presented to the senses (Helmholtz 1925). Illusions are defined as the perception of non-existent objects. Helmholtz (1867) posited that these illusions can be attributed to the profound influence of experience, training, and habit on our perceptions (cited in Hohwy 2013, 117). Prior knowledge provides perceivers with contextual cues that influence their perception and sometimes leading them to perceive things that are not present.

According to this theory, the occurrence of illusions can be attributed to mismatches between two types of contents perception is based on and that are indistinguishable in phenomenal experience. These two types of contents are bottom-up signals that are received by the sensory organs and top-down prior knowledge (Meyering 1989). Prior knowledge plays a pivotal role in guiding judgment, leading to the recognition of objects. Conversely, bottom-up sensory signals are instrumental in shaping fine-grained perceptions by integrating prior knowledge with the present context. It is not feasible, for instance, to spatiotemporally orient oneself exclusively based on prior knowledge. Instead, the ability to orient oneself in space and time is dependent on the current sensory impressions.

Helmholtz's concept of perceptual inference explains how sensory impressions and prior knowledge converge to produce perceptual experience. According to him, this inference occurs unconsciously, as it is not part of the phenomenal experience. For him, perceptual inference is an analogy useful for understanding perception. His position was that perception is not literally an inference, and that inference is not something that minds consciously do.

In the context of the inference analogy, perception is construed as consisting of a major premise and a minor premise, with the perception of an object being regarded as the result of a syllogism. In particular,

> [...] the major premiss is formed from a series of experiences, each of which has long disappeared from our memory and also did not necessarily enter our consciousness formulated in words as a sentence, but only in the form of an observation of the senses. The new sense impression entering in present perception forms the minor premiss, to which there is applied the rule imprinted by the earlier observations (Helmholtz 1977, 132).

In accordance with the inference analogy, perception is similar to a logical syllogism. The major premise of this syllogism comprises prior knowledge or past experience, the minor premise consists of the current sensory input, and the conclusion integrates both elements to infer the most probable cause of the sensations. To illustrate this, let the major premise be "dogs are barking animals," and let the minor premise be "I hear a barking." In this context, the inferred perception could be "a dog is causing the barking I hear." The presence of a dog in the visual horizon is not necessary for inferring that a dog is barking. This perceptual inference could be deceived if one later perceives that another human mimicking the barking of a dog was causing the sensory impressions.

The inference analogy suggests that perception is an interpretative act and that it does not provide direct access to reality. In the absence of the major premiss (prior knowledge), perceivers lack contextual cues to interpret the causes of sensory impressions, as the analogy suggests. Conversely, as this interpretation is contingent upon the contents of the minor premiss, namely sensations, and they result from the encounter of the world and our sensory organs, the resulting perceptual object cannot be a direct apprehension of reality. In summary, perception is indirect due to the combined influence of both cognition and sensory experience.

Helmholtz's sign theory of sensations provides a framework for understanding the why perception is an indirect access to reality. As Helmholtz (1977, 128) asserted, "the objects extant in space appear to us clothed in the qualities of our sensations. To us, they appear red or green, cold or warm, to have a smell or taste, etc., whereas after all these qualities of sensation belong only to our nervous system and do not reach out at all into external space." The concept of qualia refers to the qualities of sensations experienced by the organs. In other words, they are not objective properties of external objects and are better understood as signs of external objects.

> Inasmuch as the quality of our sensation gives us a report of what is peculiar to the external influence by which it is excited, it may count as a symbol of it, but not as an image. For from an image one requires some kind of alikeness with the object of which it is an image —from a statue alikeness of form, from a drawing alikeness of perspective projection in the visual field, from a painting alikeness of colours as well. But a sign need not have any kind of similarity at all with what it is the sign of. The relation between the two of them is restricted to the fact that like objects exerting an influence under like circumstances evoke like signs, and that therefore unlike signs always correspond to unlike influences (Helmholtz 1977, 121-122).

According to this theory, sensory experiences do not constitute images of external objects; rather, they are the result of sensory impressions mediated by the sense organ, which

renders sensations in a different modality from its causes. Sensations can evoke, but not represent, the external objects that caused them. In this regard, Helmholtz proposed that sensations function as signs of their external causes.

Helmholtz (1977) posited that qualia facilitate the formation of "an image of the lawfulness in the processes of the actual world" (122). To illustrate, the color and scent of a flower serve as indicators of its state of blooming. Qualia, such as smell, sound, or color, provide cues for spatiotemporal orientation and for the interpretative dimension of perception. According to PP, sensations play a comparable role in the refinement of predictions, serving in the updating of priors. They serve as a means of adjusting expectations about the environment in light of evidence and tracking environmental changes, thereby enabling the generation of efficient predictions for achieving various cognitive tasks.

PP draws from Helmholtz's theory of perception the analogy of inference, which is important to note as Helmholtz (1867) himself recognized that this is an analogy and that perception is not a conscious act of logical reasoning. Having made this clarification, it is worth pointing out that PP also shares the view that perception is an indirect access to reality. According to PP, sensory signals are caused by unobservable factors, and the mind must predict these factors to understand what is causing the signals. These predictions are informed by prior knowledge and refined through sensory input, resulting in a synthesis of bottom-up sensations and top-down expectations. This synthesis is a hallmark of both Helmholtz's and PP's conceptions of perception, where the mind identifies a phenomenon based on a combination of expectations and sensations.

While Helmholtz considered the inference analogy to be a useful heuristic for the scientific study of perception, more realistic interpretations of the analogy have emerged in the PP literature. This is not the case for all scholars, as evidenced by the divergent views on the analogy. Some scholars, such as Parr et al. (2018), consider it to be a useful analogy, while others, such as Orlandi (2018), suggest that it is an unfortunate one. Helmholtz's interpretation of the analogy, however, provides a solid foundation for defending an ontological agnosticism on the analogy. The question is essentially whether there is any justification for claiming that perception is inferential in nature. Helmholtz adopted an instrumentalist interpretation of inference, explicitly stating that mental processes are not inferential in nature. Although accepting a more substantive view of inference requires some sort of justification, the same need does not exist for the instrumentalist view, since the analogy is just a useful heuristic in scientific contexts.

In my view, it is more reasonable to treat the analogy of inference as one among a multitude of possible research assumptions. Indeed, scientists have long utilized analogies for investigating cognitive problems. The conceptualization of cognitive behaviors as

mechanisms, a notion derived from physics, is a common practice in scientific inquiry. As will be discussed in the subsequent section, cognitive behaviors have also been conceptualized as systems, which are abstract constructs with specific characteristics attributed to the phenomena under investigation.

The epistemic virtues of analogies cannot be established *a priori*. Scientists must engage with analogies and leverage their benefits to achieve specific purposes. The analogy of inference is but one among a myriad of possible assumptions in the scientific study of cognition. The epistemic virtues of an analogy are not determined by its accuracy in representing reality; rather, they lie in its ability to facilitate scientific progress. In the context of analogy inference, the usefulness of an analogy does not necessarily imply that perception is inferential in nature. Perceptual inference is a conceptualization that helps the construction of a target cognitive system within the context of a scientific practice.

For Helmholtz, the analogy of inference proved instrumental in establishing a link between psychological and mechanistic accounts of perception. In the context of PP, it has been instrumental in facilitating the utilization of diverse statistical instruments for the examination of perceptual problems, both at the psychological and neural levels (see Hohwy 2013). During its evolution, the analogy has acquired new functions and contents and has lost some others. The analogy of inference in PP has evolved from a logical to a probabilistic one. This transformation can be attributed to the influence of Bayesian statistics in PP, a topic that will be discussed in Chapter 6.

5.2. Cybernetic and Predictive Processing

A New Science of Systems

The cybernetic theory of systems has exerted a profound influence on various scientific disciplines since the mid-twentieth century, playing a pivotal role in the development of the cognitive sciences (Meyering 1989). The seminal cybernetic conferences organized by the Macy Foundation between 1946 and 1953 are widely regarded as the genesis of this field of research. These conferences were interdisciplinary gatherings of scientists from various disciplines, philosophers, and mathematicians who sought to establish a comprehensive framework for the study of the human mind, which would later come to be known as the cognitive sciences.

The development of new concepts, methods, and tools by cyberneticians has led to novel avenues for exploring cognition. Noteworthy contributions include Wiener's and Shannon's

work, which laid the foundation for the foundational concept of information processing in the cognitive sciences. For Wiener, cybernetic was a novel science about control and communication in animals and machines (Wiener 2019). According to Wiener, cybernetic investigates systems that manifest behaviors that are regular and purposeful. A notable illustration of this is thermoregulation in mammals, which employ mechanisms such as sweating to maintain a constant body temperature, or seeking shade in the case of reptiles, with a similar objective.

Systems are defined as entities composed of parts that interact or communicate with each other. Cybernetics is the scientific study of interaction or communication of the parts of systems that leads to regular or controlled observable behavior. For cyberneticians such as Ashby, systems were observer-relative theoretical constructs. These systems were approached by Ashby and other cyberneticians as epistemic instruments, serving as theoretical constructs for the manipulation and control of various observable behaviors. Machines, for example, can be regarded as artificial systems that manifest observable behavior because they are designed to do so. In a similar vein, humans and animals can be regarded as systems, as certain observers (e.g., scientists) attribute purposeful behavior to them.

According to Ashby (1956), the study of control and communication in animals and humans was merely one of the potential applications of cybernetics, which he conceptualized as a general theory of systems. The fundamental concern of cybernetics, as articulated by Ashby, was not the nature of systems *per se*, but rather their behavior. According to Ashby, systems are not "things but ways of behaving" (1956, 1). Within the context of cybernetics, the concept of **behavior** refers to the observable output of patterned interactions among system components. Systems theories aspire to control and predict various behaviors. As Heylighen and Joslyn emphasize:

> Cybernetics is the science that studies the abstract principles of organization in complex systems. It is concerned not so much with what systems consist of, but how they function. Cybernetics focuses on how systems use information, models, and control actions to steer toward and maintain their goals, while counteracting various disturbances. Being inherently transdisciplinary, cybernetic reasoning can be applied to understand model and design systems of any kind: physical, technological, biological, ecological, psychological, social, or any combination of those (2003, 155).

Despite the contemporary disuse of the term "cybernetics" in scientific discourse, the concepts developed by cyberneticians have exerted a profound influence on contemporary scientific reasoning. These include information processing, models, feedback, control, and systems. Beyond its contributions to the development of concepts, cybernetics also

pioneered a novel approach for framing scientific problems. By shifting the focus from the nature of things (the question of what a thing is) to their function (what they do), a new kind of theoretical framework that emphasizes controlling and predicting observable behaviors was created.

In this novel framework, the parts of a system are not concerned with corresponding to the actual parts and interactions of an entity. Rather, they are concerned only with the parts necessary for a system to produce observable behavior. Similarly, the contemporary use of the concept of a system in cognitive science, informed by this cybernetic concept, does not refer to cognitive entities, mechanisms, or structures; rather, it deals with processes and tasks created in scientific practices. These theoretical constructs are used to create models that predict, control, or simulate observable behaviors.

As theoretical constructs, systems are not empty vehicles for the objective representation of phenomena. As Ashby argued, systems possess observer-relative properties and laws. From a modeling perspective, these properties and laws are embedded in the epistemic instruments employed for exploring reality. These properties and laws are incorporated into system models that shape the conceptualization of target cognitive behaviors. This conceptualization is commonly referred to as "system model," which is an epistemic instrument that serves to construct various models for exploring and intervening in reality.

In a system model, an entity, process, or phenomenon is defined as a system composed of interacting components. The construction of a system model process is as follows:

i. An observable behavior is stipulated to be the target domain of a system; that is, the observer behavior is taken to be what the system does. In this context, the concept of function refers to the way the patterned interaction of system components gives rise to a desired behavior.
ii. The variables, components, or parts of a system are defined. Cybernetic approaches conceptualize entities as systems composed of interconnected parts, and the idea is that the communication between these parts leads to regular global behavior. The regular behavior of a system is understood to result from the interaction of its parts, and when a change occurs in one of these parts or its interactions, it affects the others, thereby producing a feedback loop. In certain instances, these alterations may result in the system's failure to function or exhibit regular behavior, thereby leading to its breakdown.
iii. The final step involves identifying the patterns of interactions among these components that result in the desired behavior, thereby enabling control over the system.

Cybernetics is arguably the precursor of the cognitive science (Pickering 2010; Lowe et al. 2018). This is because it developed a framework for the systematic and quantitative study of cognition. Notably, this framework facilitated the conceptualization of observable cognitive behavior because of the interactions among components of a system, thereby *representing* cognition *as* information processing. However, this influence was not fully acknowledged at the early years of the cognitive sciences. The research program of cognitivism, in its initial stages, distanced itself from cybernetics by accusing it of portraying cognition as a mechanical stimulus-response process (Lecas 2006).

Cognitivists also accused Cybernetics of its anti-mentalist implications, although it maintained the assumption that cognitive beings are information processing systems (Varela, Thompson and Rosch 2016). Cognitivism contended that cognitive systems within Cybernetics are deterministic automata and that cyberneticians provide no criteria for distinguishing human and artificial intelligence. Cognitivism shifted the focus from solely relying on the notion of information processing, proposing that the mind employs internal mental representations to transform information into meaningful structures that facilitate perception and action (Haugeland 1978).

It is my understanding that the sciences of the mind have drawn more from cybernetics than the concept of information processing (see Chemero 2000). The conceptualization of cognition as a systemic function, the scientific practice of studying cognition through the construction of models, and the delineation of variables or components that result in a desired behavior belong to a modeling approach to cognition that was pioneered by cyberneticians. A salient distinction between cyberneticians and cognitivists pertains to their approach toward mental kinds. Cyberneticians, in contrast to cognitivists, did not attempt to determine mental kinds through theories of cognition. They, for instance, did not argue for the necessity of mental representations, for the scientific explanation of cognition.

The notion of information processing proved to be pivotal in the development of computational approaches in early cognitive sciences research. The analogy between cognition and computation stems from the foundational work of Alan Turing, who argued that the mind is a computational system that is not constrained by its physical realization (Turing 1950). This cybernetic concept was subsequently incorporated into a cognitivist theory known as Computer Functionalism or Strong AI program (see Searle 1980). The crux of both theories lies in the view that distinguishing between human and machine intelligence is not a straightforward task, as the material realization of cognition is, to a certain extent, irrelevant (Nath 2020). This implication has been the subject of substantial criticism, particularly from proponents of 4E cognition. Notably, this idea was also contentious among cyberneticians (Miller 2003; Dewhurst 2019).

The heterogeneity of cybernetic views becomes evident in the epistemology and ontology of systems. Cyberneticians hold divergent views on the nature of knowledge that a science of systems can deliver, as well as on the ontological status of systems. For instance, Wiener was relatively skeptical about the potential of cybernetics to explain the mind. In a philosophical paper, he advocated for a kind of relativism that

> admits the existence of certainty, of any degree of certainty short of absolute certainty. Though it considers that even the best approximation is subject to criticism, it does not regard this as preventing us from giving the brevet rank of absolute certainty to items of our knowledge, and using them as a basis for the criticism of other knowledge, without, at the same time, criticizing them. And it will not permit the relative certainty of our scientific knowledge to be degraded to the rank of mere 'appearance' at the behest of any metaphysical theory (1914, 577).

For Rosenblueth and Wiener, system models serve as proxies for reality. Despite Wiener's relativist stance on scientific knowledge, he postulated that system models maintain some correspondence with real entities. Both Rosenblueth and Wiener conceptualized system models as abstractions, asserting that "no substantial part of the universe is so simple that it can be grasped and controlled without abstraction. Abstraction consists in replacing the part of the universe under consideration by a model of similar but simpler structure" (1945, 316). By focusing on the parts of a system that are relevant for producing a behavior, system models abstract away from the complexity of phenomena to "approach asymptotically the complexity of the original situation" (1945).

Rosenblueth and Wiener argued that system models facilitate the acquisition of approximate knowledge by virtue of their partial structural correspondence with their targets. According to them, "a formal model is a symbolic assertion in logical terms of an idealized relatively simple situation sharing the structural properties of the original factual system" (Rosenblueth and Wiener 1945, 317). Despite their perspective on scientific knowledge as being partial and susceptible to criticism, they had a realist view of systems.

Rosenblueth and Wiener (1945) did not assume that structural correspondence is something that models, by virtue of any default partial correspondence with their targets. According to them, the belief that a natural entity is a system merits justification if it is to be given serious consideration within the scientific community. In their words, "All scientific problems begin as closed-box problems, i.e., only a few of the significant variables are recognized. Scientific progress consists in a progressive opening of those boxes." (Rosenblueth and Wiener 1945, 319).

The perspectives held by Rosenblueth and Wiener (1945) concerning the nature of knowledge in system theories serve as an illustration of the debates on the ontology and epistemology of systems. For the purposes of this chapter, it is essential to note that Wiener's relativism is distinct from the ontological agnosticism that is being advocated for herein. While acknowledging the limitations of scientific knowledge, Rosenblueth and Wiener (1945) propose that scientific knowledge is only approximate, subject to criticism and refinement. In their view, scientific models are partial representations of reality.

Ontological agnosticism, in contrast, eschews the assumption that system models are good models of reality because they map structures that exist in nature. According to this perspective, control and predictability are not contingent upon the ability to accurately represent reality. In my view, system models function as epistemic instruments, designed to yield specific forms of knowledge. Investigating entities or processes as systems necessitates the establishment of several assumptions, which transform phenomena into predictable and controllable research objects. My agnosticism concerns whether research objects are representations of natural and mental kinds.

Ashby's Systems Theory of Cognition

W. Ross Ashby (1903-72), an English psychiatrist, is widely regarded as one of the founders of cybernetics. His most important contribution to the field is his theory of adaptive systems, which was crucial to the development of the cognitive sciences and PP. In the following pages, I argue that in his works Ashby's anticipated the PP's concepts of error, model, and active inference while treating them as epistemic instruments in the construction of system models. For Ashby, the purpose of cybernetics was not just to conceptualize phenomena as systems, but also to materially instantiate such idealized systems in concrete representational vehicles.

According to Ashby, system models are neither scientific representations of phenomena nor representations of the structure of phenomena. Ashby leveled accusations against the science and philosophy of his era for their failure to make a clear distinction between phenomena and structure. He argued that this oversight led to the misuse of these concepts. From Ashby's standpoint, cybernetics does not concern itself with the study of phenomena; rather, it focuses on the analysis of systems.

Systems are idealized structures that are relative to some observers (scientists). Prior to their utilization within the sciences, system models must first be constructed. To this end, scientists must first define a behavior of interest and stipulate that such behavior is the result of the interactions between system components. Any phenomena, including the

human body, society, or physical processes, can in principle be described as systems (Krippendorf 2009). However, not all system descriptions are of equal epistemic value. For instance, the social phenomenon of segregation can be conceptualized as a systemic problem within a population. However, the statement "segregation is a systemic problem" is superfluous if the relevant parts of the population and the relevant interactions that lead to segregation are not clearly specified.

According to Ashby (1956), the distinction between systems and phenomena does not imply that they are unrelated; rather, the point is that system models must be first constructed before they can be used for any epistemic purpose related to the understanding of phenomena. As Weisberg (2007) explains, multiple scientific models can be about the same target system, or the same model can be used to explore various targets. This principle extends to system models, as they, too, are constructs that pertain to theories rather than reality. The use of cybernetic systems enables scientists to gain control and predictability over observable behaviors that are created in, though not controlled by, the research practices. Ashby, according to Hacking's (1983) terminology, was anti-realist about systems, but not phenomena, treating systems as instruments for gaining control over them.

In Ashby's system modeling approach, phenomena are analogized with mechanisms. Mechanisms exhibit regular and reproducible behaviors. The aim of system modeling is thus to gain control and predictability of systems' behaviors. This can be achieved by an observer who comprehends how a mechanism is assembled and how the relevant interactions among its components result in desired system states.

Ashby's system modeling approach is arguably a historical antecedent to the mechanical view in the cognitive science, which holds that "the mind should be thought of as a kind of causal mechanism, a natural phenomenon which behaves in a regular, systematic way, like the liver or the heart" (Crane 2003, 1). This approach was regarded by Ashby as a scientific heuristic. By this, it is meant that the mechanistic approach to the mind is not a reductionist perspective treating cognitive phenomena as epiphenomena stemming from physicochemical states in the brain.

Analogizing cognitive phenomena with mechanical systems constitutes a scientific strategy that, by making certain assumptions, opens epistemic possibilities in the exploration of cognition. As Crane (2003) has emphasized, the mechanical view paves the way for an exploration of the material realization of cognition. For Ashby, this perspective enables control and prediction of system behavior, rather than the representation of the real structure of cognitive phenomena. Ashby (1956) distinguished sharply between phenomena and systems, asserting that the laws governing the former are independent of the laws governing the latter. Thus, the laws that govern systems pertain to theory—that is, to how

scientists impose constraints on their theoretical constructs—and do not apply to phenomena, at least not directly.

Ashby (1979) believed a mechanical perspective had the potential to explain intelligent behavior. He was a proponent of a materialist school of psychiatry, which espoused the view that mental phenomena have a physicochemical basis (Pickering 2009). The aforementioned school distinguished mechanistic approaches from metaphysics of mind. According to Ashby, mechanistic approaches are concerned with observable behaviors and processes. In contrast, metaphysics of mind addresses questions such as the nature of consciousness, will, and spiritual life. According to Ashby, these metaphysical questions are beyond the scope of the sciences of the mind (see Asaro 2008).

Error as a Concept of Information Theory

Having introduced the system modeling approach pioneered by Ashby, it is now possible to elaborate on how this approach advanced several of the main concepts of PP. According to Clark, the central idea of PP finds a clear precedent in Ashby's work: "The whole function of the brain is summed up in: error correction" (Ashby n.d., quoted in Clark 2013, 181). Ashby conceptualized cognitive systems as regulatory systems, thus treating them as entities that behave in a regular, reproducible, and predictable manner. For both PP and Ashby's scientific models of the mind, error and prediction error signals serve as how the components of a system communicate to produce cognitive behavior.

In Ashby's work, the term "error" is employed in certain instances to denote the errors of a trial-error mechanism that dynamic systems utilize to reach equilibrium. This notion is an integral component of the analogy between systems and mechanisms that was previously outlined. However, this notion of error falls outside the scope of his theory of adaptation, as discussed below. As Lowe et al. (2018, 153) observe, "in a complex, dynamic, and hazardous environment, purely trial-and-error behavior is not viable—the organism risks damage or even destruction if it repeatedly tries the 'wrong' behavior."

In approaching regulatory systems, and then adaptive systems, Ashby employed another notion of error related to entropy in information theory. Cybernetic models conceptualize systems as entities that produce an observable behavior due to the relevant interactions of their components. Information theory contributes to this framework by specifying these interactions in quantifiable terms. Within this theoretical framework, these interactions are conceptualized as forms of communication, understood as the transmission of messages from sources of information to intended recipients (Shannon 1948). Messages are encoded as signals and transmitted through communication channels.

In system models, messages are not semantic structures but rather signals that inform about the probability of events, states of the system, or values. According to Shannon, the informational value of a message is associated with the surprise of a state, also referred to as informational or Shannon entropy. If a message conveys information about a state with a known probability of occurrence, it is considered unsurprising and thus offers no new information. Conversely, if the message conveys information about a state that is improbable, it is surprising and informative. In this context, Shannon entropy can be defined as the quantity that measures the amount of information added to a system by a message, or surprise.

Building on Shannon's information theory, Ashby applied the scientific concept of entropy to new areas of interest, namely cognitive and biological systems. Ashby (1991) employed Shannon entropy as an epistemic instrument to quantify the communication among components of regulatory systems. These systems, as their nomenclature suggests, exhibit a constant behavior due to the stability of the interactions between their components. The concept of stability will be discussed subsequently. For the sake of brevity, it is sufficient to note that such behavior is regular because stable interactions ensure that the system settles at specific desired values.

Consider, for instance, the human body temperature system, which employs diverse regulatory mechanisms to maintain a constant temperature. In regulatory systems of this nature, the messages transmitted from one component to another do not contribute to the acquisition of new information. Consequently, the entropy of the system is near zero. In the words of Ashby:

> In regulation, the 'message' to be transmitted is a constant, i.e., has zero entropy. Since this matter is fundamental, let us consider some examples. The ordinary thermostat is set at, say, 70 ºF. 'Noise', in the form of various disturbances, providing heat or cold, threatens to drive the output from the value. If the thermostat is completely efficient, this variation will be completely removed, and an observer who watches the temperature will see continuously only the value that the initial controller has set (1991, 412).

Thermostats, defined as devices that measure the temperature of specific spaces, function as regulatory systems. They are responsible for the regulation of temperature, ensuring that the ambient temperature of a space remains within the desired values. A furnace, for instance, serves as an example of a thermostat because it adjusts the rate of gas combustion to regulate the heat output within a room. This is an example of a self-regulated thermostat, in which alterations to the burn rate are contingent on a desired value indicated by a thermometer, which is an integral component of the mechanism. A thermostat driven

by the mechanism of error correction would take the discrepancy between the desired and the actual temperature as an input for updating the burn rate.

As regulatory systems, thermostats execute actions to preserve their states at desired values. According to Ashby's theory, other regulatory systems, such as the brain, do not measure a physical quantity but rather an informational one. Consequently, these systems are not governed by physical laws but by informational principles. In both scenarios, regulatory systems approach a state of zero entropy when their states align with the desired values (Ashby 1991). Entropy, in this context, is a measure of informational surprise, which serves to regulate the behavior of the system when it deviates from the desired values. In such instances, components of the system initiate the transmission of entropic messages. Entropy can also be a contributing factor to the noise in a transmission of a message. However, regulatory systems exhibit tolerance to noise.

In the PP framework, informational entropy is referred to as **prediction error**. The concept of prediction error adopts divergent meanings, contingent on its application in cognitive or biological systems modeling. From an informational perspective, prediction error is defined as a quantity that quantifies the amount of information conveyed by a message. Within the context of scientific models of cognition in PP, prediction error functions as a measurement of the discrepancy between the predicted and actual sensory states. The magnitude of prediction error is directly proportional to the degree of discrepancy between an actual sensory state and its prediction. In the context of modeling biological systems, prediction error is referred to as "free energy," representing a measurement of surprise within a generative model. In the PP framework, biological systems are conceptualized as self-organized entities that maintain their viability by resisting potentially deleterious states. Free energy, in this theoretical framework, serves to quantify the viability of biological systems. This will be further discussed in Chapter 6.

The utilization of the concept of error, irrespective of its application to biological or cognitive systems, presupposes a characterization of systems as informational within the context of the PP framework. Consequently, the notion of error within the PP framework does not offer any insights into the nature of cognitive or biological systems. Instead, it functions as an epistemic instrument, facilitating the quantification of specific variables within these systems. As previously mentioned, this characterization is rooted in Ashby’s theory of systems. In this theory, error is regarded as a constituent of an observer-dependent target system, and its utility depends on its ability to help scientists achieve control over target systems. As discussed earlier, cyberneticians were primarily interested in controlling, simulating, and predicting the behavior of systems, not in their nature.

Adaptive and Stable Systems

In developing a cybernetic theory of cognition, Ashby focused on the concept of adaptation, on the assumption that this behavior distinguishes intelligent from non-intelligent beings. In his major work, *Design for a Brain*, Ashby defines adaptive behavior as follows: "A form of behavior is adaptive if it maintains the essential variables within physiological limits" (1960, 58). An example of such a variable is the temperature of the human body. When the body's temperature surpasses certain thresholds, the body enters a state of risk, which compromises its integrity. Adaptive systems employ various regulatory mechanisms to maintain certain variables within viable limits. For these systems to execute such regulatory actions, they must possess the capacity to discern potential threats to their viability. In this context, error signals constitute information that the system variables are in non-viable states. Ashby presents the behavior of a retina as an exemplar:

> The retina works best at a certain intensity of illumination. In bright light the nervous system contracts the pupil, and in dim relaxes it. Thus the amount of light entering the eye is maintained within limits.
>
> If the eye is persistently exposed to bright light, as happens when one goes to the tropics, the pigment-cells in the retina grow forward day by day until they absorb a large portion of the incident light before it reaches the sensitive cells. In this way the illumination on the sensitive cells is kept within limits (Ashby 1960, 60).

In this example, the retina behaves adaptively because it makes certain changes to perform the stable visual function despite variations in the system's states, which are the levels of light exposure. This type of action is called control, or in some cases motor control, and occurs when a system detects error signals. In Ashby's words:

> To be adapted, the organism, guided by information from the environment, must control its essential variables, forcing them to go within the proper limits, by so manipulating the environment (through its motor control of it) that the environment then acts on them appropriately (1960, 82).

As a behavior of biological regulatory systems, adaptation consists in the ability of a system to act either in its internal states or in the states of its environment. This understanding of adaptation is arguably the same as the one of the PP framework:

> Minimizing free energy (and therefore entropy) enables an organism to resist the dispersive effects of "external forces acting upon it" to ensure "it is in continuous equilibrium with the forces external to it" (Pavlov 2010).

> In other words, minimizing free energy reduces the discrepancy (e.g., prediction error) between sensations and their predictions. This discrepancy can be reduced by changing predictions – through perception – or by selectively sampling sensory inputs that were predicted – through action (Pezzulo, Rigoli and Friston 2015, 18).

Ashby's cybernetics and the PP framework both conceptualize organisms as information processing systems, with the capacity to detect and intervene in internal or external states. These frameworks share the common premise that the reduction of informational error is essential for adaptive behavior. For Ashby, this information-based approach proved indispensable for the establishment of a quantitative biological science. Ashby raised several concerns about the understanding of adaptation in the sciences of his time, which lacked precisely the instruments for approaching it quantitatively:

> But the concept of "adaptiveness," though important, has several marked disadvantages. Firstly, it is vague. It is very difficult to define exactly what we mean by the word. Secondly it is not quantitative; and although one can recognize, roughly, degrees of adaptation yet there is, at present, no means of expressing this quantitatively (...).
>
> It would clearly be better if this concept could be changed for another which would be equivalent to it as far as its essential features are concerned but which would be free from these objections.
>
> It is suggested in this paper that the concept of "stable equilibrium" may perhaps be equivalent to it (Ashby 1940, 478).

To approach the concept of adaptive behavior from a quantitative perspective, Ashby transferred the notion of equilibrium, originally developed in the field of physics, to the domain of systems theory. According to the Encyclopaedia Britannica (2019), equilibrium is "the state of a system in which neither its state of motion nor its internal energy state tends to change with time." Ashby adopted Lorentz's (1927) conceptualization of equilibrium, which was defined as "the state of a system in which neither its state of motion nor its internal energy state tends to change with time." In his words, "by a state of equilibrium we mean a state in which it [a system] can persist permanently" (quoted in Ashby 1947, 54). In contrast, stable equilibrium is a behavior exhibited by dynamic systems, defined as systems whose states undergo change over time. According to Ashby, "It is only when we disturb the bodies and observe their subsequent reactions that the concept [of stable equilibrium] develops its full meaning" (1940, 479). To illustrate this point, consider a pendulum in a state of rest, as opposed to one in motion. The former is said to be in equilibrium, while the latter is centered around a stable equilibrium position.

In Ashby's theory, the concept of stable equilibrium serves as a scientific analogy to understand adaptive behavior. He contended that, in contrast to the concepts of adaptation in the science of his era, the notion of stability is "[...] purely objective; it avoids all metaphysical complications of 'purpose'; it is precise in its definition and lends itself immediately to quantitative studies" (Ashby 1940, 478). This analogy enabled the conceptualization of the behavior of adaptive systems as movements towards a stable equilibrium position. The concept of stability refers to dynamic systems that react to "any disturbance from the equilibrium position (...) [by calling] (...) forth opposing forces that restore the system to its initial state" (Pickering 2009, 217).

Ashby's theory of adaptation made use of a concept from physics to model biological systems. Specifically, Ashby imagined the capacity of adaptive systems to respond to perturbations as if they were restoring an initial equilibrium (Asaro 2008). Ashby, indeed, "insisted that what it means to survive, and therefore to be living, can be captured by the general systems concept of stability" (Froese and Stewart 2010, 15). In his conceptualization, the essential variables of biological systems are centered around a stable equilibrium position corresponding to viable states. To illustrate, consider again the stable equilibrium position of the adult human body temperature, which is 37°C. Temperatures near this degree pose no threat to the viability of the system; however, an extreme temperature of 42°C can potentially cause brain damage.

The preceding example illustrates that regulatory systems, such as the human body temperature system, sustain their states at desired values unless one or more of their components become damaged. To illustrate this point, consider the scenario in which a thermostat's malfunctioning component results in its failure. Consequently, the thermostat would be unable to accurately measure the temperature of a physical space. A system's breakdown occurs when it is unable to maintain a stable equilibrium position in response to alterations in one or more of its components.

Regulatory systems do not cease to function when one or more of their components are damaged. However, the system's reactions to these disruptions can lead to novel behaviors. Consequently, the system is, in a strict sense, a new one, given that systems are defined by what they do. Asaro (2008, 158) asserts that such alterations can be formally determined "in mathematical theory of mechanisms, [if] the equations or functions that previously defined the system no longer hold true."

Models in Cybernetics and Predictive Processing

To recapitulate the previous points, in Ashby's cybernetic theory, biological systems are conceptualized as adaptive systems whose behavior is driven by regulatory mechanisms. These mechanisms ensure stability by controlling environmental conditions when the states of a system are not in proximity to a stable equilibrium point. The stability of a biological system is contingent on its internal organization and environmental conditions. Consequently, alterations in either the internal or external states can generate breaks in the system. Such breaks do not imply that the system ceases to exist; rather, the system adopts a new, different behavior.

An example of a change in the internal organization of a system that can cause a break in the human body system is the abrupt fluctuations in blood pressure. An example of an environmental condition that can cause it is being in a space devoid of oxygen. For Ashby, biological systems demonstrate resilience to breaks. This suggests that while certain alterations may compromise the system's functionality, not all are necessarily fatal.

In light of this, Ashby proposed the categorization of biological systems as ultrastable systems, which are regulatory systems capable of tolerating instability in their global behavior and breaks in their parts. Ultrastable systems are comprised of a reacting part and an environment. Ashby's characterization of ultrastable systems is as follows:

> Two systems of continuous variables (that we called 'environment' and 'reacting part') interact, so that a primary feedback loop (through complex sensory and motor channels) exists between them. Another feedback, working intermittently and at a much slower order of speed, goes from the environment to certain continuous variables which in their turn affect some step-mechanisms, the effect being that the step-mechanisms change value when and only when these variables pass outside given limits. The step-mechanisms affect the reacting part; by acting as parameters to it, they determine how it shall react to the environment (Ashby 1960, 98).

In contrast to stable systems, ultrastable systems resist breaks without altering their global behavior, even if local behaviors (i.e., the behaviors of their parts) undergo change. To achieve this, ultrastable systems depend on feedback mechanisms, as articulated by Lowe et al. (2018, 149): "Ultrastability refers to the requirement of multiple (at least two) feedback loops—behavioral and internal ('physiological')—in order to achieve equilibrium between an organism and its environment."

As previously mentioned, Ashby pioneered a modeling approach to systems. While for him the identification of concepts to describe phenomena was essential, e.g., the concept of ultrastability to describe biological systems, he also believed that such conceptualizations

must be instantiated in representational media for them to be epistemically valuable. To this end, Ashby constructed the Homeostat, a scientific model of ultrastability that demonstrated the physical possibility of this behavior.

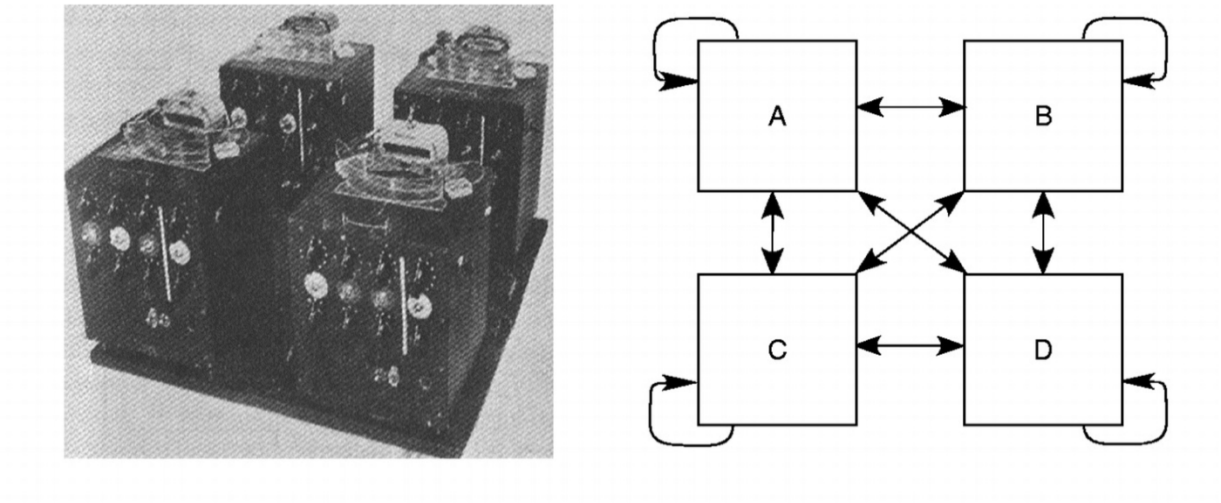


Figure 7: The Homeostat (reproduced from Cariani 2009, under Taylor & Francis terms).

The Homeostat is a machine composed of four subsystems that maintain a dynamic equilibrium through feedback loops. Cariani (2009, 142) explains how it works in the following:

> Each box-shaped Homeostat module had a pivoting magnet at its heart that moved a plate. The plate swung within a semicircular chamber filled with distilled water that was situated on the top of the box. Electrodes were situated at the ends of the chamber. Depending on the orientation of the magnet, the plate's position vis-à-vis the electrodes would change, and hence its voltage V with respect to them. The four plate voltages constituted the four control variables to be stabilised. In each module, the orientation of the magnet itself depended on the summed effect of the magnetic fields created by four coils (A–D), which were driven by currents from their respective plates in the different modules (amplified via triodes). If the control voltage V of the plate in a given module deviated from its intermediate null state, which was about halfway between the electrodes, a 'uniselector' stepping switch was activated and a new, randomly chosen circuit was put in action in place of the previous one. The uniselector for a given module switched several circuit parameters at once: four resistances that effectively determined the relative strengths of magnetic fields in the four coils, and four capacitances that changed the temporal characteristics of the circuits in which the coils were embedded. Thus, at any given time the structure of

> the Homeostat could be characterised in terms of 32 parameters, i.e. 4 modules × (4 resistances+4 capacitances) per module.[15]

The construction of the Homeostat demonstrated that ultrastability is physically realizable. This concept refers to the stable equilibrium resulting from the interactions of two or more systems with feedback connections. In the Homeostat, stable equilibrium is represented by a set of target values, with the model's task being to maintain controlled variables in proximity to these target values. As Cariani (2009, 143) explicates, "the Homeostat evaluated whether a particular set of parameters made a 'good controller' vis-à-vis a particular environment."

The relevance of the Homeostat to the field of cognitive science was largely overlooked until the rise of PP. By demonstrating the possibility of ultrastability, the Homeostat became the first scientific model to physically instantiate adaptive behavior or homeostasis (i.e., the maintenance of an internal organization of a system by keeping its essential variables within certain limits) in a changing environment. In summary, the Homeostat is not a scientific model that shows that biological systems are ultrastable. Instead, it functions as a model that instantiates ultrastability (regulation by feedback mechanisms) through a representational medium. Specifically, it represents the behavior of a system that exhibits tolerance to perturbations or changes in its components. While the Homeostat does not directly represent any specific cognitive or biological process, the model construction practice itself demonstrated the physical possibility of such a behavior. Ashby's used the Homeostat to study the brain. In the field of PP, analogous models have been employed to study cognition (Seth 2015).

---

[15] A thorough elucidation of the Homeostat's functioning lies beyond the purview of this chapter; for an in-depth exposition, refer to Cariani (2009). Cariani's (2009, 142) explanation of how the Homeostat works continues as follows: "Given the four uniselector switches, each with 25 distinct positions, the Homeostat therefore had 25 × 25 × 25 × 25 (390,625) clearly defined combinations of uniselector states that determined the analogue control parameters and the associated behaviour of the device. Particular combinations of parameters in interaction with a particular external signal could lead to stability or to chaotic instability. While the ‘digital’ part of the device, the uniselector switch, had clear, discrete functional states, the analogue part was less exhaustively specified. Although Ashby notes that the parameter space of the Homeostat at any given time could be characterised in terms of 32 parameters (16 resistances +16 capacitances), in fact other factors, such as the friction of the magnet pivot or external magnetic fields or subtle nonlinearities in the operations of the triode, could also come into play to influence the dynamics. Since the Homeostat proved to be an ultrastable device, designed to be impervious to almost all conceivable perturbations, including signal inversions, these additional factors likely made no difference in the device's adaptive control functioning, since the device always eventually successfully settled on a stable configuration. In terms of specification, in contrast with the digital uniselector states, strictly speaking, the analogue portions of the device were only partially defined."

A final point of discussion regarding Ashby's contributions to PP pertains to the ambiguity of the notion of model, a topic that has been thoroughly examined in this dissertation. Within the PP framework, models function as both generative scientific models and putative cognitive structures. In contrast, Ashby's conception of models remains constrained within the domain of scientific models, referring to a component of ultrastable systems. This conception of model proved instrumental in the conceptualization of generative models in PP, a topic that will be elaborated upon subsequently.

To maintain their global behavior despite local behavioral changes, ultrastable systems rely on feedback mechanisms and control processes. The manifestation of these processes necessitates the presence of a regulator or controller and a system to be controlled. According to Ashby's argument, for this process to occur, a regulator must be a model of the system.

Regulators and controlled systems possess their own states. The complexity of a system is relative to the number of states it possesses. Complex systems necessitate a regulator that can model their states to be controlled. Conant and Ashby (1970, 89) posited that "the relation between regulation and modelling might be much closer, that modelling might in fact be a necessary part of regulation." Building upon this premise, Ashby advanced a law for system modeling. The law of requisite variety stipulates that the condition for regulators to control complex systems is to create models. Ashby (1991) defines variety as the number of possible states of a system. The law asserts that only variety can absorb variety. In essence, it asserts that a regulator or control mechanism must possess a minimum number of states equal to the number of states of a system to effectively control it or maintain its stability.

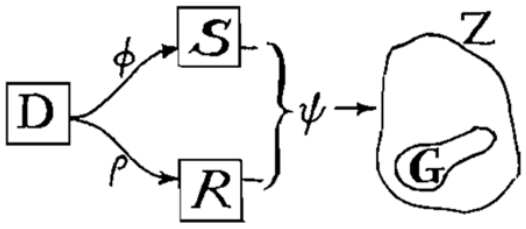


Figure 8: A regulator and a controlled system

D are environmental disturbances that affect the states of a system. φ are events in the system S triggered by D, and ρ are events in the regulator R triggered by D. ψ are interactions between S and R, that produce the set of events Z. G is the subset of events in Z produced by effective control exerted by R to S (Conant and Ashby 1970, reproduced under Taylor & Francis terms).

According to the law of requisite variety, R must have equal or more states than S for the subset G, the controlled states, to have any states. If R is a good regulator, the difference between Z and G states is near zero. Ashby argued that the effective control requires R to be isomorphic to S. In Figure 8, ψ triggers states in G if there is a one-to-one mapping between R and S states.

To conclude this detour into Ashby's cybernetics, it is necessary to elaborate on the relevance of the cybernetic notion of a model with respect to the concept of a generative model. In previous chapters, it was proposed that cognition in PP is conceptualized as a form of model-building. While this occurs in scientific conceptual work, generative models are used as instruments to investigate cognitive problems in scientific practices. Ashby was a pioneer in this approach to scientific problem-solving, as cybernetics was for him both a conceptual framework for characterizing phenomena as systems and a means of instantiating these conceptualizations in concrete representational media, such as the Homeostat.

While an ambivalent notion of model was arguably already present in Ashby's cybernetics, the idea of prediction-based modeling is certainly new in PP. Indeed, Conant and Ashby (1970) anticipated the view of cognition as model building by arguing that models are an essential part of complex regulatory systems. As Seth (2015, 8) notes, they "build on the law of requisite variety by arguing (and attempting to formally show) that the nature of a controller capable of suppressing perturbations imposed by an external system must instantiate a model of that system." Ashby used the term model to refer to the regulator of a self-organized system, which corresponds to a model of its optimal organization.

Finally, Ashby argued that models are essential to certain types of systems. As mentioned earlier, systems are relative to certain observers. This means that system laws, such as the law of requisite variety, are rules of system models, not real entities. Nevertheless, Conant and Ashby (1970, 90) believed that using these models in science could improve our understanding of the brain:

> The suggestion has been made many times that perhaps the brain operates by building a model (or models) of its environment; but the suggestion has (so far as we know) been offered only as a possibility. A proof that model-making is necessary would give neurophysiology a theoretical basis and would predict models of brain operation that the experimenter could seek. The proof would tell us what the brain, as a complex regulator for its owner's survival, must do.

Since the emergence of PP, the conceptualization of cognition as predictive modeling and the brain as a statistical machine has developed in various directions, oscillating between

anti-realist and realist views. My ontological agnosticism regarding mental modeling realism is inspired by Ashby's idea that models are components of systems. I have attempted to demonstrate that these models are epistemic instruments for exploring reality. In my estimation, Ashby's theory offers a middle ground by suggesting that models function as epistemic instruments that facilitate the formulation of scientific problems concerning the mind.

Why should we be Ontological Agnostics?

In this chapter, I examined how Helmholtz's theory of perception and Ashby's systems theory anticipated the core concepts and analogies of PP. These include the analogy of perception as inference, and the concepts of prediction error and model. My analysis of the history of these concepts supports to an agnostic view of mental modeling because it reveals them as ontologically neutral epistemic instruments.

Cybernetics pioneered a modeling approach to scientific problems. This approach transforms complex problems into simpler ones that can be controlled and predicted. Cyberneticians distinguished between cognitive behaviors defined within scientific practices and cognitive phenomena in their full complexity. This distinction is significant for understanding cognition in contemporary theories, such as PP. It implies that scientific models of cognitive systems, such as generative models in PP, do not represent cognitive phenomena. Rather, these models are artifacts that open epistemological possibilities, such as controlling or predicting observable behaviors in research practices.

Just as systems theory is observer-relative, the idea of perception as inference is a scientific analogy. Helmholtz found this analogy instrumental for studying perception. The notion of perceptual inference, which originated in his work, has been controversial in the PP literature. Helmholtz adopted an instrumentalist rather than a realist interpretation of the analogy. This analogy is just one possible assumption in the scientific study of cognition. Its epistemic value lies not in its accuracy, but in what it enables scientists to accomplish. Another analogy has been discussed in this chapter: the one between stable equilibrium and adaptive systems in cybernetics. Analogical reasoning is a pervasive strategy in cognitive scientists' theorizing. It is a means of importing instruments from other disciplines to understand, simulate, and predict cognitive behavior.

Ashby's modeling approach to adaptation pioneered an informational approach to cognitive and biological systems. This approach conceptualizes the behavior of these systems as maintaining certain system states at desired values. Building on this idea, PP adds that cognitive systems use prediction error to perceive and act. This chapter's historical analysis

has revealed this concept's origin in information theory and its connection with the concepts of entropy and surprise. Currently, this concept is part of the constellation of epistemic resources used in PP to investigate cognition. As such, it does not refer to any kind of physical or natural information.

Ashby also pioneered a modeling approach to scientific problem-solving by creating a surrogate model known as the Homeostat. He argued that adaptive behavior could only become the subject of scientific investigation by conceptualizing it as observable behavior resulting from the patterned interactions of system components. His modeling approach consisted of instantiating this conceptualization in a representational medium, the Homeostat. The construction of this model was epistemically informative in that it demonstrated the physical possibility of ultrastability; a generic behavior conceived in the conceptual dimension of a modeling practice.

Ashby's influence extends to the concept of a generative mental model in PP. In particular, he argued that models are a necessary component of complex regulatory systems. He conceptualized adaptation — a generic concept of intelligence that applies to any sentient being — as the behavior of these systems. In this context, a model is a component of a regulatory system. The concept of a regulatory system was then used to investigate the brain, framing it as an organ that "helps the body cope with an unpredictable environment, to adapt to whatever comes along" (Pickering 2009, 216).

Models are components of regulatory systems that specify the optimal distribution of system states. Without models, regulation would not be possible because there would be no way to determine the values needed to maintain a system's internal organization. In my view, this concept of a model is a clear precursor to the concept of a generative model in PP. Both serve the purpose of developing a functional characterization of the brain and cognition as a system's behavior. As Chapter 3 has shown, these conceptualizations are the first steps scientists take to develop testable hypotheses and predictions about the behavior of this organ.

Finally, the history of concepts that became influential in PP suggests that they were not originally considered to be related to mental phenomena, processes, or mechanisms. For example, Helmholtz's analogy of inference was instrumental in developing an understanding of the phenomenon that facilitated its scientific study. In turn, cyberneticians developed the system model, arguably the standard instrument for conceptualizing cognitive problems in contemporary cognitive sciences.

But just because a scientist conceptualizes a problem in a certain way does not mean that the phenomena are represented by that conceptualization. It is more accurate to think of

conceptualizations as creative processes that produce the target systems of science. As this chapter's detour into systems theory has shown, research objects are created in scientific practices and are made epistemically accessible only thanks to these creative processes.

Chapter 6 - Transdisciplinarity in Predictive Processing

In this chapter, I justify my interpretation of Predictive Processing as a modeling framework within the sciences of the mind. I conceptualize modeling frameworks as constellations of epistemic resources that are relevant to the construction and manipulation of scientific models. These resources include concepts, assumptions, knowledge, skills, and mathematical, computational, and other methods. According to my account, research traditions in the cognitive sciences, including PP, are better understood as modeling frameworks. This is because scientific modeling is the default strategy for analyzing problems and addressing questions in this field. From an epistemological perspective, frameworks are interesting because they frame cognitive target systems in specific and standardized ways.

This chapter focuses on the transdisciplinary epistemic resources of the Predictive Processing modeling framework. My goal is to demonstrate that models and target systems in Predictive Processing are transdisciplinary objects created through epistemic practices that blend conceptual, mathematical, and computational resources. Target systems acquire the properties of the epistemic resources used in their construction, such as the properties of model templates. In this context, I analyze Bayesianism and the FEP as two prominent epistemic resources that impose statistical and informational constraints on Predictive Processing models.

As the term suggests, frameworks transcend disciplinary boundaries, facilitating the resolution of problems across diverse scientific fields. The transfer of epistemic resources is exemplified by the use of concepts that are not bounded by disciplines, such as prediction error, inference, models, and free energy, in the study of cognitive and biological systems in Predictive Processing. This chapter will also explore model templates, which are mathematical and computational structures used across the sciences to construct models. I conceptualize model templates as minimal platforms for constructing model and target systems. By relying on templates and other epistemic resources, modeling frameworks provide scientists with a *modus operandi* for framing and solving problems.

Section 6.1 discusses whether cognitive science is an interdisciplinary or transdisciplinary field of study. I raise skepticism about the idea that this discipline aims to provide multilevel

explanations of complex cognitive phenomena. Instead, I argue that cognitive science is a transdisciplinary field of study that relies on frameworks for constructing models of transdisciplinary target systems created in modeling practices. From this, I infer that the discipline pursues epistemic gains beyond multilevel and mechanistic explanations of complex phenomena.

Section 6.2 presents two templates of the Predictive Processing modeling framework: Bayesianism and the Free Energy Principle. I argue that Predictive Processing provides a general conceptualization of target systems, and that these templates offer mathematical and computational resources for investigating those target systems framed in statistical and mathematical terms. Together, they facilitate the construction of models and target systems. Finally, I critique metaphysical interpretations of Predictive Processing which disregard the epistemic implications of using models in the cognitive sciences.

6.1. Transdisciplinarity in the Cognitive Sciences

Predictive Processing as a Modeling Framework

The nature of PP as a scientific enterprise is a matter of debate. Scholars refer to it as either a theory, a framework, or a scientific model. In previous chapters, I suggested that PP is best understood as a modeling framework. I expand on this idea in this section.

Definitions of the concept of a framework are relatively scarce in philosophy of science literature. This does not mean that frameworks have not been the subject of philosophical reflection; they have, albeit indirectly, in discussions of similar notions, such as paradigm, stance, research program, research tradition, or epistemic system.

Recently, Partelow (2023) defined frameworks in science as integral yet vague tools for empirical inquiry and theory development. Partelow characterizes them as vague because the concepts and relationships associated with a framework are not always clearly defined or consistently applied. Partelow focuses on frameworks used in scientific research and science-based policy, particularly the Social-Ecological Systems framework. Originally developed to study human-environment interactions, it has since been repurposed in diverse fields such as medicine, psychology, economics, and the humanities (Colding and Barthel 2019). This suggests that frameworks are not bounded by disciplinary constraints. In the Social-Ecological Systems framework, for example, the same tools used to study psychological problems can also be used to study economic ones.

As the term suggests, frameworks serve to frame scientific problems in specific ways. According to Partelow (2023), this framing process depends on the concepts and relationships belonging to a framework. While this is true, I believe the framing process also depends on other relevant factors not mentioned in Partelow's work. For example, frameworks in the cognitive sciences rely on various representational devices that embed conceptualizations into tangible models. Therefore, concepts and their relationships, though integral, are insufficient for addressing specific theoretical and practical scientific problems.

As I have suggested in previous chapters, conceptualizations provide workable characterizations of target systems. Information theory, for instance, has been instrumental in conceptualizing biological phenomena as information processing problems. However, the framing of biological phenomena as information processing problems alone would not lead to the construction of scientific models unless scientists also include other epistemic resources, such as assumptions, analogies, model templates, and methods. Biologists' sophisticated technologies and methods, not just information theory concepts, are integral to investigating biological phenomena framed as informational problems.

Based on the above, I define frameworks as constellations of conventionally related epistemic resources that serve the purpose of framing and solving scientific problems. Unlike other accounts, mine takes into account both the conceptual and material resources of frameworks. As constellations of epistemic resources, frameworks do not establish a one-to-one pairing of concepts with other epistemic resources. Rather, scientific problem framing and solving result from the creative and opportune use of epistemic instruments. How diverse epistemic resources are drawn together depends on the specific problem being addressed, the scientists' skills and knowledge, the availability of instruments and technologies, and the scientific community's habits, values, and conventions.

In previous chapters, I have argued that PP is best understood as a framework. However, the PP literature has failed to acknowledge this constellation of epistemic resources, conflating the entire framework with specific methods, models, or theories within it. This explains why, over the past decade, numerous names have been given to the framework, including the 'Bayesian brain hypothesis' (Knill and Pouget 2004), the 'Predictive Mind' (Hohwy 2013), the 'Cybernetic Bayesian Brain Hypothesis' (Seth 2015), or 'Action-Oriented Predictive Processing' (Clark 2013, 2016; Kirchoff 2018).

Defining modeling frameworks as constellations of epistemic resources suggests that despite their respective foci, research practices under their umbrella rely on similar assumptions, concepts, models, and methods. For instance, research conducted within the PP framework typically employs Bayesian statistical methods (Orlandi 2018). Though

alternative methods could, in principle, be used (Colombo, Elkin, and Hartmann 2021), the tradition of using these methods is what makes the Bayesian method a crucial epistemic resource of the framework.

According to my account, principles, theories, and other epistemic resources within the same constellation are not hierarchically dependent on one another. This view contrasts with the hegemonic views in the PP literature. These views hold that the Free Energy Principle is a scientific principle from which theories and models, such as PP, are derived. My account, in contrast, suggests that epistemic resources within a framework—such as theories, principles, and models—are drawn together to frame and solve scientific problems. Accordingly, there is no need to assume any hierarchical dependencies within them. I discuss the idea of hierarchies in theories at the end of this chapter.

Three epistemic resources of the PP framework deserve particular attention: Two are Bayesian methods and the Free Energy Principle (also called active inference), discussed in the second part of this chapter. The third is Predictive Coding, a computational method that has gained significant traction in recent decades. Corlett, Mohanty, and MacDonald (2020, 531) define Predictive Coding as "an encoding strategy in signal processing whereby the expected features of an input signal are suppressed and only unexpected features are conveyed." Predictive coding is used to model prediction error in information processing systems. Interestingly, this method has been conflated with the entire framework. As Hohwy (2020, 210) notes, "It is common to see PP equated with predictive coding." Due to the frequent and indiscriminate interchange of these terms in the literature, distinguishing between them is somewhat arbitrary. Thus, using one term to refer to either the method or the framework is permissible, provided the intended meaning is explicitly stated. Here, I define PP as the framework and Predictive Coding as an epistemic resource, or a method pertaining to it.

Modeling frameworks not only provide a *modus operandi* for constructing models but also play a role in constructing target systems. As previously stated, research conducted under a framework's umbrella uses certain epistemic resources to investigate various target systems in conventional ways. In doing so, the properties of these epistemic resources are projected onto different target systems. This projection implies that the target system is *represented as* a particular kind of thing: a research object constrained by the properties of the epistemic resources involved in its construction.

Since these resources are used in standardized ways, the resulting target systems of these models become standardized, too, with some adjustments, and are constrained similarly. Thus, diverse target systems modeled under the same framework become similar or

structurally similar. This is not due to structural similarities between target systems but rather an implication of using the same epistemic resources to investigate them.

My account of modeling frameworks explains why diverse targets, such as perception, attention, and mental disorders, are conceptualized as probabilistic inference tasks in PP. PP models of cognition are not models of cognitive phenomena, but rather, they are models that *represent* cognition *as* probabilistic inference. Models relate to target systems through projection. In other words, target systems are *represented as* if they had the properties of the models. Through modeling, scientists transform complex cognitive phenomena into solvable cognitive tasks.

Since frameworks provide standardized conceptualizations of cognition, and projections are relative to frameworks, there are as many established conceptualizations of cognition as there are frameworks in the sciences of the mind. As discussed in Chapter 2, PP conceptualizes cognition as predictive, model-based, and inferential tasks. Research under its umbrella has been used to model visual perception (Itti and Baldi 2009; Rao and Ballard 1999) and perceptual learning (Fiser, Berkes, Orban, and Lengyel 2010), among other cognitive phenomena (Corlett, Mohanty, and MacDonald 2020).

Since scientific models are the primary epistemic instruments for exploring cognitive problems, I refer to frameworks in the cognitive sciences as modeling frameworks. These frameworks are transdisciplinary because their epistemic resources are not bound by disciplinary constraints. In the context of the cognitive sciences, this means that phenomena studied by more than one discipline are studied using the same repertoire of epistemic resources.

Scientists draw on a range of epistemic resources to investigate different cognitive behaviors. For models to be epistemically useful, they must be constructed as systems of dependencies that exhibit behaviors of interest (Knuuttila 2021)—that is, behaviors analogous to cognitive behaviors. However, the behavior exhibited by a scientific model does not have to correspond directly to a cognitive behavior found in nature.

According to Bailer-Jones (2009), scientific analogies help explore unknown phenomena using knowledge from more familiar domains. In PP, the analogy of inference helps scientists idealize cognitive phenomena as the behavior of probabilistic systems in uncertain environments — a behavior that can be represented using statistical tools. As I have argued in previous chapters, models of cognition have cognitive tasks as their target systems. These tasks are constructed based on idealizations, reductionist heuristics, or analogies belonging to a framework. Cognitive tasks are epistemic things (Rheinberger 1997) that are created during modeling processes and tailored to available epistemic resources.

Cognitive tasks are idealized behaviors that can be instantiated in tangible representational media. A minimal representational function can be ascribed to scientific models of cognition if "representation" is understood to mean that cognition is *represented as* an idealized behavior exhibited by a model. Model construction opens several epistemic possibilities in the cognitive sciences. Computational modeling allows scientists to simulate cognitive behavior, statistical modeling allows scientists to predict the behavior of a system, target-directed modeling allows scientists to formulate testable hypotheses and predictions, and mechanistic modeling allows scientists to control target systems of interest.

The rest of this chapter discusses how constructing cognitive target systems involves transdisciplinary use of epistemic resources. First, however, I present my account of transdisciplinarity in the cognitive sciences in discussion with the 'received view' of this discipline as an interdisciplinary field of study.

The Received View of Interdisciplinarity

In a popular handbook of cognitive science, Bermúdez introduces it as an interdisciplinary field of study:

> Many different disciplines study the mind. Neuroscientists study the mind's biological machinery. Psychologists directly study mental processes, such as perception and decision-making. Computer scientists explore how those processes can be simulated and modeled in computers. Evolutionary biologists and anthropologists speculate about how the mind evolved. In fact, very few academic areas are not relevant to the study of the mind in some way. The job of cognitive science is to provide a framework for bringing all these different perspectives together (2014, xxvii).

The assumption that cognitive science provides frameworks that integrate partial disciplinary knowledge into more complete explanations of the mind is representative of what I call the **received view of interdisciplinarity** in the field. This view holds that understanding complex mental phenomena requires multilevel explanations that integrate partial disciplinary knowledge. The idea that mental phenomena are complex stems from the vast amount of data and observations of intelligent behavior in animals, plants, and tiny organisms relevant to forming ecosystems. At its core, the idea is that cognition is a complex phenomenon. As such, it triggers interdisciplinary research (Kaiser, Kronfeldner, and Meunier 2014).

One problem with the received view is that it fails to acknowledge that interdisciplinary collaboration is uncommon in this field, as bibliometric evidence shows (Núñez et al. 2019). Furthermore, it incorrectly assumes that the main aim of cognitive sciences is explanation and that interdisciplinarity is a means to that end. Examining knowledge production dynamics in this field reveals that knowledge transfer plays a central role in the discipline. For example, hypothesis testing often involves transferring theoretical resources from various disciplines (Newen, De Bruin, and Gallagher 2018). Yet, hypothesis testing itself does not aim to integrate partial explanations into complete ones. It is a more modest epistemic activity.

As discussed earlier, modeling cognitive systems requires heuristics that transform complex problems into simpler, more solvable ones. In doing so, cognitive scientists are not seeking complete explanations, but rather the opposite. When framed as cognitive tasks, cognition becomes a simpler, controllable, predictable and simulatable target system, rather than a complex phenomenon.

The received view of interdisciplinarity gives an idealized picture of the research practices of cognitive scientists. I will argue in the following that, rather than explaining complex cognitive phenomena, cognitive scientists are interested in addressing specific theoretical and practical problems. To this end, cognitive scientists borrow epistemic resources from various disciplines. However, this practice is transdisciplinary, not interdisciplinary. Through knowledge transfer, the models lose their disciplinary origins, and the target systems become transdisciplinary research objects.

Transdisciplinary Knowledge Transfers and Model Templates

Hanne Andersen (2016, 2) defines scientific disciplines as epistemic units "consisting of a set of closely related cognitive resources, such as concepts, models and theories." In transdisciplinary fields such as the cognitive sciences, these epistemic resources cross disciplinary boundaries. This can be understood in two ways. First, phenomena such as perception, attention, and language are transdisciplinary research objects. Second, these research objects are studied using transdisciplinary epistemic resources. One example is the use of Bayesian statistical concepts, models, and methods to study cognitive, biological, or social problems.

Consider the perception of other intentions as a target system. It is a transdisciplinary research object that lies at the intersection of various disciplines. Cognitive scientists rely on philosophical theories, such as mind-reading or simulation theory, to conceptualize it. Cognitive psychologists create experimental conditions to examine it empirically.

Neuroscientists use brain imaging techniques to gather data and isolate patterned brain activity elicited by this process. This target system is not a philosophical concept, a neuroscientific research object, or a psychological phenomenon. Rather, it is a transdisciplinary target system constructed by drawing together epistemic resources from these disciplines.

Knowledge transfer is the process of drawing epistemic resources from one discipline to address the scientific questions and problems of another. Herfeld and Lisciandra (2019, 1) point out that this is a common practice in interdisciplinary collaborations:

> Exemplary cases of knowledge transfer are numerous. For instance, psychological concepts have been introduced into economic models to acknowledge various motives behind human decision-making and explain phenomena such as pro-social, habitual, or addictive behavior. The underlying idea is that such concepts could help economic theories to better explain the effect of tax policies, trading behavior, or the exchange of goods beyond the marketplace.

According to these authors, knowledge transfer has become pervasive in scientific research in recent decades, leading to the formation of new scientific domains at the intersection of different fields of study (Andersen 2016; Herfeld and Lisciandra 2019). This is indeed the case in the cognitive sciences. In PP, Scientists have employed epistemic resources, such as Bayesian statistical tools and mathematical structures like the Free Energy Principle, to model cognition as probabilistic inference. However, the concept of knowledge transfer is too general to account for the specific mathematical, computational, and conceptual resources of PP. Knuuttila and Loettgers' (2014, 2016, 2023) notion of a model template fits the bill. Model templates are conceptualizations intertwined with the mathematical and computational tools used in quantitative sciences.

Knuuttila and Loettgers developed the notion of a model template by building on Humphreys's notions of computational and theoretical templates:

> The notion of a model template is inspired by Paul Humphreys' work on computational science, and especially his discussion on what he calls computational templates (Humphreys 2002; 2004). While for a computational template its tractability is the most important feature, the notion of a model template stresses also the conceptual side of model transfer. (2016, 379).

Humphreys (2002, 2004) introduced the concept of a computational model to explain why computationally oriented sciences rely on a small set of equations to solve problems. These equations are used and reused across domains in computational modeling practices, becoming templates that describe diverse phenomena in the same mathematical terms.

Examples of computational templates include the Poisson distribution, the wave equation, harmonic oscillators, the Ising model, the Lotka–Volterra equations, and agent-based models (Knuuttila and Loettgers 2023, 200).

On the other hand, theoretical templates are representational devices within a theory that contain at least one schematic variable that can be replaced by a different variable to represent a new phenomenon. An example is Newton's second law, in which force is a schematic variable with gravitational, electrostatic, and other types of instances. Notably, Newton's second law states that "force must be represented by the vector sum of separate forces," which restricts the template's application to systems where vector addition is meaningful (Humphreys 2019, 114). Humphreys notes that theoretical templates have disciplinary origins but can be applied to new domains thanks to their schematic variables. However, as with Newton's second law, their use is limited by the intended interpretations of the theories. In other words, "a theoretical template is not a purely formal or mathematical object, and many of the alternative interpretations that are possible when the mathematical features of the theory are considered only as formal objects fall outside the domain of the theory" (2019, 114).

Knuuttila and Loettgers (2014, 2016, 2023) introduced the concept of a model template to account for the interplay between mathematical tools, computational techniques, and conceptualizations in scientific modeling. A model template is a transdisciplinary structure that propagates across different disciplines to investigate specific problems (Knuuttila and Loettgers 2017a). For our purposes, this concept clarifies why cognitive phenomena, brain networks, and biological target systems are modeled as probabilistic inference tasks in PP. As previously mentioned, during model construction, the properties of this epistemic resource are projected onto target systems. In this respect, PP is analogous to systems theory, which originated in mathematics and physics and has been applied to sociology (Luhmann, 1997). Systems theory enables sociologists to frame sociological problems in mathematical terms. Similarly, PP has drawn on statistical resources to model cognition in statistical terms.

Knuuttila and Loettgers (2023, 382) further define a model template as "a formal platform for minimal model construction coupled with a very general conceptualization without yet any subject-specific interpretation or adjustment." Interpretations and adjustments are necessary for constructing models and are relative to target systems and domains of interest. For example, artificial neural networks cannot have negative firing rates, weights, or activations when modeling brain networks. However, as templates, they could have them in principle. This adjustment is necessary to make them physiologically plausible.

According to Knuuttila and Loettgers (2014, 295), model templates capture "the intertwinement of a mathematical structure and associated computational tools with theoretical concepts that, together, depict a general mechanism potentially applicable to any subject or field displaying particular patterns of interaction." As I discuss in the next section, the concept of a model template helps explain how epistemic resources are drawn together in the construction of models and target systems in PP. PP research relies on a minimal conceptualization of cognition as probabilistic inference and mathematical structures, such as Bayes' rule and the Free Energy Principle, to construct models of various cognitive tasks. PP models can capture specific patterns of interaction in diverse cognitive tasks by relying on them. These patterns are captured because target systems are modeled using the same mathematical tools. By applying model templates to cognitive problems, PP research projects the properties of epistemic resources onto target systems, thereby fostering the production of structurally similar answers to similarly framed problems.

6.2. Epistemic Resources in Predictive Processing

This section explores Bayes' rule and the Free Energy Principle. These are two model templates used in PP to model cognition. This section also discusses their role in model and target construction. First, I briefly review the conceptualizations of cognition as probabilistic tasks with which these templates are intertwined. Then, I discuss how these conceptualizations and the mathematical and computational dimensions of model templates in PP are relevant to constructing target cognitive systems. To conclude, I address the problems associated with interpreting these conceptualizations as ontologies of cognition.

Conceptualizations of Cognition

Modeling frameworks are constellations of transdisciplinary epistemic resources used to construct scientific models and target systems. Those in the cognitive sciences provide distinctive conceptualizations of cognition that are complemented by other transdisciplinary epistemic resources when investigating target systems. As I have argued, research traditions in the cognitive sciences, such as Cognitivism, Connectionism, 4E Cognition, and PP, can be understood as modeling frameworks. These frameworks also exchange epistemic resources. For example, theories of 4E cognition have been used in PP research (e.g., Gallagher and Allen 2018; Hipólito and van Es 2022), and PP has been used

as a framework to study embodied cognition (Allen and Friston 2018; Kirchhoff and Robertson 2018).

The cognitivist framework conceptualizes cognition as the processing of information through semantic structures (Bickhard and Terveen 1995). Marr's cognitivist model of visual perception conceptualizes perception as "the process of discovering from images what is in the world and where it is" (2010, 3). Such conceptual work is necessary for constructing target systems. In cognitivist models, target systems are computational tasks. In Marr's model, for instance, visual perception is an output resulting from the processing of input images (Kitcher 1988). The model specifies the algorithms the system follows to produce outputs (Piccinini and Scarantino 2011) and describes a physical system that processes information (Bermúdez 2014).

The Atkinson-Shiffrin model of memory, as outlined in previous chapters, serves as an example of a cognitivist model. In this model, memory is conceptualized as the process of allocating sensory information into short-term and long-term stores:

> Incoming sensory information first enters the sensory register, where it resides for a very brief period of time, then decays and is lost. The short-term store is the subject's working memory; it receives selected inputs from the sensory register and also from the long-term store (Atkinson and Shiffrin 1968, 90-91).

The Atkinson-Shiffrin model is cognitivist because it conceptualizes cognition as an information-processing task. It also specifies that an algorithm determines the interactions of the model's components. Like other cognitivist models, however, it does not specify the physical realization of the process.[16]

In the connectionist framework, cognition is conceptualized as behavior that emerges from the patterned interaction of network components. Connectionist models instantiate this conceptualization using artificial neural networks as representational vehicles. These networks are computational models composed of layers of interconnected nodes that learn patterns from data (Kiang 2003, 303). Unlike cognitivist models, connectionist architectures do not require knowledge in the form of semantic content. They are governed by dynamic and soft rules (Clark and Lutz 1992). In artificial neural networks, nodes can represent neurons, and the weights denote the synapses that connect them (Lynch et al. 2019). PP models may be instantiated in artificial neural networks, but this is not necessary.

[16] Cognitivists often neglected the material realization of cognitive systems, leaving out any discussion about implementation. Those who embraced functionalism posited that cognition is multiply realizable. The physical realization of cognition was not a subject of interest to them (Churchland 1982).

Chapter 2 introduced generative models in PP as scientific models composed of predictions, precision estimates, error signals, priors, and hierarchical relationships. The construction of these models goes hand in hand with the construction of target cognitive systems, such as perception, learning, and action, as probabilistic inference tasks. Specifically, they are framed as systems that minimize prediction error through the interaction of their components. Scientific work employing generative models applies this conceptual framework to cognitive problems. As I will argue below, model templates play a crucial role in blending conceptualizations with representational devices such as artificial neural networks. These templates offer standardized methods of solving probabilistic inference problems computationally and mathematically.

Bayes' Rule as a Model Template

The term 'Bayesian' refers to "a set of interrelated principles, methods, tools, and problem-solving procedures" (Colombo and Hartmann 2017, footnote 1). These epistemic resources are useful for addressing problems involving non-fully observable, non-deterministic, or uncertain environments (Russell and Norvig 2022). When applied to cognitive problems, uncertainty becomes a property of the noisy or ambiguous sensory data that a cognitive system gathers from an environment. Cognitive systems must then make inferences under uncertainty to navigate their environments (Colombo, Elkin, and Hartmann 2021, 4). Within this framework, perception and action consist of inferring the most probable unobservable causes of sensations and the effects of an agent's interventions on the environment, respectively.

Bayesianism provides epistemic resources, particularly a model template, for modeling probabilistic inference under uncertainty. As previously discussed, model templates have conceptual, computational, and mathematical dimensions. The conceptual dimension frames target systems, such as perception and action, as probabilistic tasks. Through this conceptualization, the brain is *represented as* a statistical machine:

> If the brain is making inferences about the causes of its sensations then it must have a model of the causal relationships (connections) among (hidden) states of the world that cause sensory input. It follows that neuronal connections encode (model) causal connections that conspire to produce sensory information (Friston 2012c, 1230).

The conceptual dimension of the Bayesian template includes basic concepts and their relationships insofar as they facilitate the construction of target systems in mathematically tractable and computationally solvable ways. According to Friston (2012c), the brain is a

system that models how states in the world cause sensory inputs and make inferences about sensations. In my account, this simply means that the behavior of this organ is *represented as* a probabilistic inference task.

Bayesian inference is a statistical concept that involves using current information to update or reduce the degree of belief in hypotheses about the distribution of states of a given target system (Colombo and Hartmann 2017). It works as follows: An observer specifies a system and initial beliefs (priors) about the distribution of events. These beliefs are *represented as* a set of hypotheses, H $\{H_1, H_2, H_3, \ldots, Hn\}$ with assigned weights representing their probabilities (the sum of all weights equals one). The system adopts the hypothesis with the highest weight as a prior belief and estimates the distribution of events. The resulting estimate, called the posterior distribution, is then used to update the weights of the priors.

Bayes' rule is an algorithm that determines how Bayesian systems update their beliefs. More specifically, it "prescribes how the probability of a hypothesis should be updated based on new evidence" (Colombo and Hartmann 2017, footnote 1). Bayes' rule is expressed as follows:

$$P(H|E) = \frac{P(E|H)\,P(H)}{P(E)}$$

Figure 9: Bayes' Rule

Bayes' rule serves to calculate the posterior probability of a hypothesis, or its probability after being tested against evidence (E). According to the rule, the posterior probability of a hypothesis (P(H|E)) is the product of the likelihood (P(E|H)) and the prior (H*i*), divided by the marginal probability (P(E)) of the data. A prior probability, or H*i*, is the weight of a hypothesis and is independent of the evidence. Likelihood is the probability of new evidence given a hypothesis.

The Bayesian model template combines a generic conceptualization of problems as probabilistic inference tasks with a mathematical structure that solves such tasks. Bayes' rule is a transdisciplinary model template because its conceptualization of problems and mathematical structure have been used in diverse fields, including the social sciences, ecology, genetics, and medicine. However, traditional scientific approaches to Bayesianism, even when acknowledging its role in model construction, fail to articulate the intertwining of its conceptual and mathematical components. Consider Van de Schoot et al.'s (2017) distinction between the uses of Bayesianism in cognitive psychology:

i. Theoretical framework: Bayesian models are theoretical explanations of human reasoning and behavior, emphasizing their inferential nature.
ii. Modeling tools: Bayesian hierarchical models are detailed and concrete models of the psychological processes underlying human reasoning. These models can make predictions about human behavior that are comparable to data on actual behavior.
iii. Mathematical tools: Bayesian statistics are techniques for analyzing data.

In my account, Bayesianism is not a theoretical explanation of the inferential nature of human cognition. Rather, it is a template that conceptualizes target systems as problems involving inference under uncertainty. For example, Bayesian models of perception describe how prior knowledge should be combined with sensory data to make inferences about the world (Knill, Kersten, and Yuille 1996, 15).

The second and third uses of Bayesian approaches in Van de Schoot et al. (2017) correspond to the mathematical and computational apparatuses of Bayesianism. For instance, scientists provide "mathematical descriptions of the problems of perception" by constructing concrete models (Knill, Kersten, and Yuille 1996, 19). As discussed in Chapter 3, constructing these models enables the generation of testable hypotheses and predictions about target systems' behavior. For instance, one can predict attention-related responses in cortical area V4 (Rao 2005), or associate feedback connections with predictions and feedforward connections with prediction error signals (Shipp 2016; Harkness and Keshava 2017).

Although applying the Bayesian template to cognitive research enables the creation of target-oriented scientific models, such as the action-observation network model discussed in Chapter 3, these models are neither explanatory nor mechanistic. This is because these models do not specify the mapping between the scientific model and entities and processes in the brain. To become explanatory or mechanistic, PP models would need to specify a "mapping from informational contents to measurable neural activities to connect the variables in the model to signals in the brain" (Cao 2020, 524). However, even if this were the case,

> [a] mapping introduces an additional degree of freedom into any assessment of the model; with different mappings, the same model could make very different predictions about brain activity. This looseness in interpretation gives rise to [the] main worry: without pre-registration of which mappings are allowed, predictive models and traditional models alike could accommodate whatever brain data is found (Cao 2020, 524).

This looseness of interpretation is occasionally found in literature on the predictive mind. For instance, some researchers have proposed that predictive processing (PP) can account for delusions as false beliefs stemming from a systematic failure of the brain to process prediction errors, resulting in the formation of biased generative models (Adams et al. 2013; Fletcher and Frith 2009; Sterzer et al. 2018). However, these are only conceptualizations. The PP account of delusions does not explain the physical mechanisms underlying this phenomenon. As Cao (2020) rightly points out, this would be incorrect because accounts such as the PP account of delusion do not introduce any mapping between the model and the target system. The PP account of delusions is just one example of the flawed interpretations of the epistemic value of modeling cognition as probabilistic inference found in the literature.

As I have argued, templates are instrumental for constructing both models and target systems in science. By using templates, the mathematical and computational features of models are projected onto target systems. In doing so, target systems are *represented as* mathematical and computational problems. Knowledge of cognition derived from these practices does not represent real cognitive phenomena or mechanisms (Kaplan and Craver 2011; Harkness and Keshava 2017; Colombo, Elkin, and Hartmann 2021; Cao 2020). The mathematical or computational properties attributed to a target system are properties of cognitive tasks—research objects created in modeling practices.

In sum, using the Bayesian template to address questions about cognition provides a mathematical formalism associated with a conceptualization of problems involving probabilistic inference. However, it does not reveal much about the physical realization of cognition itself. As Shipp (2016, 17) points out, these models do not address "much of the richness of cortical microcircuitry." Therefore, these models cannot be considered *representations of* cognitive phenomena or mechanisms, nor should ontological theses concerning cognition be inferred based on the science using them. The kind of ontology they involve is operational, meaning it is relevant only within research practices. In this ontology, cognition is *represented as* probabilistic inference.

The Free Energy Principle as a Model Template

Before explaining why I consider the Free Energy Principle (FEP) to be a model template, I must first explain what it is. Consider the following passage:

> Many theories in the biological sciences are answers to the question: "what must things do, in order to exist?" The FEP turns this question on its head and asks: "if things exist, what must they do?" More formally, if we can define what it means to be

> something, can we identify the physics or dynamics that a thing must possess? To answer this question, the FEP calls on some mathematical truisms that follow from each other. Much like Hamilton's principle of least action, it is not a falsifiable theory about the way 'things' behave—it is a general description of 'things' that are defined in a particular way. As such, the FEP is not falsifiable as a mathematical statement, but its postulates may be falsifiable to the extent that they refer to a specific class of empirical phenomena that the principle aims to describe (Friston et al. 2023, 2).

The question "What must things do to exist?" rests on the premise that some things exist and some do not. The FEP rephrases the question as "If things exist, what must they do?" to distinguish between things that occupy space, such as water, and things that exist, such as a fish.

The FEP represents systems that exist or survive and persist over time as systems that minimize an informational quantity known as free energy. A corresponding equation describes how these systems minimize free energy. By referring to the FEP as a 'model template,' I am suggesting that it is a conceptualization intertwined with mathematical tools. Here, I focus on the conceptual dimension of this principle.

According to Parr, Pezzulo, and Friston (2022), the FEP conceptualizes living organisms as entities with internal states that exist within environments with external states and engage in reciprocal interactions with them. They mean that organisms observe what occurs in their environments via perception and change what happens in their environments via action. Through perception and action, organisms can adapt to their environment, maintaining their bodily integrity over time.

The FEP formally *represents* these systems *as* composed of sensory states (perception), active states (action), and internal states (generative model). These states are strictly separated from, yet interact with, external states (environment). This separation is represented by a Markov blanket, which defines the system's boundary in a statistical sense (see Kirchhoff et al. 2018, for an explanation). Sensory and active states enable interaction between internal and external states.

According to the FEP, living systems maintain their physical integrity by minimizing free energy (Colombo and Wright 2021). Free energy is defined as "the surprise of some data given a generative model" (Friston, 2009, 293). Let me unpack this definition. In the above context, a generative model is an internal representation of external states that a system uses to form expectations about upcoming sensory states. When sensory states are similar to what is expected, they are not surprising. In this case, the system does not gain much information and does not update the generative model. The function of a free energy system

is to minimize surprise. Thus, when surprise is high, the system either changes its expectations, gathers or samples information from the environment, or intervenes in the environment.

Maintaining physical integrity is a complex task for living systems because they must deal with a high number of variables. Environments are hazardous because they are in constant flux, and not all these changes are viable for living systems. Some fluctuations are caused by actions, while others are exogenous and result in non-deterministic change. Friston uses the example of "a fish out of water" (2010, 127). Like other sentient beings, fish can find themselves in states that threaten their existence. Some of their actions can cause these states. For instance, a fish can leap out of the water and eventually die. Fish in tanks may be sensitive to turbidity or extremely high-water temperatures.

The FEP posits that organisms can perceive environmental states that endanger their physical integrity because they model the relationship between their environment and internal states. Thus, they learn which environmental states can threaten their internal organization. Friston (2012a, 2101) argues that organisms can do this because they can "distil structural regularities from environmental fluctuations (like changing concentrations of chemical attractants or sensory signals) and embody them in their form and internal dynamics." In other words, by identifying regularities in the effects of environmental states on internal states, organisms learn to perceive which states are threatening (non-viable) and thus act to avoid them.

Perception and action are highly abstract concepts in the FEP. They are behaviors of generic self-organizing systems. The FEP does not address the specifics of, for instance, plant, human, or cat perception. Rather, it describes them functionally, that is, by the role they play in regulating a system’s interactions with an environment.

Perception provides indirect access to environmental states. It is indirect because it only indicates whether environmental states are viable or not, which is inferred from observed sensory states. Free energy-minimizing systems infer whether they are surprising, given a generative model. An action is a movement that aims to transform surprising states into non-surprising states given a model of how interventions change the environment or internal states (Sims and Pezzulo 2021). As Friston et al. (2010, 228) put it: "If we change the environment or our relationship to it, then sensory input changes. Therefore, action can reduce free energy by changing the sensory input predicted, while perception reduces free energy by changing predictions."

FEP systems minimize surprise by perceiving and acting. Surprise is an informational quantity relative to a generative model of environmental states. Since these systems

minimize surprise through perception and action, these two cognitive processes are also considered informational processes under the FEP. Badcock, Friston, and Ramstead (2019) argue that self-organizing systems avoid non-viable states by creating a model that distinguishes between viable and non-viable states. In this model, viable states are *represented as* non-surprising states, or states with a high probability of occurrence. Thus, under the FEP, "to be alive simply means to revisit a bounded set of states with a high probability" (107). By framing biological problems as informational ones, the FEP provides a mathematical framework for addressing phenomena such as homeostasis, survival, and allostasis, all of which are framed as the avoidance of surprising states.

Now that I have explained how the FEP conceptualizes living systems, I would like to discuss why I consider it a template model. In the literature, the FEP has been understood as a framework, a first principle, a scientific model, and a theory. In some cases, the FEP has been conflated with PP. Colombo and Wright (2021) have noted this, pointing out that the FEP is sometimes understood as a functional description of the brain's activity and structural connectivity. This is how PP is usually understood, not in the broader sense used in this dissertation. According to Friston (2009, 293), when the brain is considered in terms of free energy minimization, "nearly every aspect of its anatomy and physiology starts to make sense." For Friston, PP is a theory of brain function derived from the FEP, a first principle.

Colombo and Wright (2021) distinguish between the PP, the FEP, and active inference. Active inference is a theory that explains how biological entities, processes, and complex systems maintain their physical integrity in changing environments. In my view, these authors differentiate the conceptual dimension of a model template from its mathematical apparatus. In contrast, I argue that a strict separation between active inference and the FEP is unwarranted. As we have seen, the FEP's conceptual dimension frames the task of maintaining physical integrity as minimizing surprise, an informational quantity.

As a model template, the FEP primarily targets informational systems, which are more generic than cognitive and biological systems. Therefore, adjustments must be made when applying it to the latter. The FEP describes self-organizing systems that maintain their physical integrity. It accomplishes this with the aid of information theory concepts, such as surprise. This approach is not new. For instance, Schrödinger employed a similar strategy when he used physical concepts, such as the avoidance of states of thermodynamic equilibrium, to differentiate between organisms and matter. In his words:

> When a system that is not alive is isolated or placed in a uniform environment, all motion usually comes to a standstill very soon as a result of various kinds of friction; differences of electric or chemical potential are equalized, substances which tend to

> form a chemical compound do so, temperature becomes uniform by heat conduction. After that the whole system fades away into a dead, inert lump of matter. A permanent state is reached, in which no observable events occur. The physicist calls this the state of thermodynamical equilibrium, or of 'maximum entropy' (Schrödinger 1992, 69; originally published in 1944).

Schrödinger was not the first to use statistical physics and thermodynamics to the study of organisms. However, he was the first to observe that living systems defy the second law of thermodynamics, which states that "the entropy of an isolated system increases in the course of any spontaneous change" (Moberg 2019, 2571). The FEP transforms this physical notion of entropy into the informational notion of surprise. According to Badcock, Friston, and Maxwell (2019, footnote 4), living systems avoid information-theoretic entropy: "The free energy formulation is a mathematical description of the dynamics of systems at a nonequilibrium steady-state and should not be confused with the second law of thermodynamics. The FEP deals with information theoretic measures."

The fact that the FEP is an informational principle lends support to my interpretation of it as a model template. The FEP can be used to construct target systems, such as homeostatic or adaptive systems, or self-organizing systems that avoid deleterious phase transitions. According to Sajid, Ball, and Friston (2021, 679), an adaptive system avoids surprising states and "maintains homeostasis by residing in attractor states that minimize entropy (or surprising observations)." Thus, this template can be used to mathematically represent the behavior of living systems that maintain their physical, chemical, or other kinds of integrity. An example of homeostatic behavior is as follows: "If you mechanically dissociate a sponge into a random mixture of its cells, it is well known that these cells will re-aggregate and rearrange to recreate their normal histological arrangement within about three days" (Harris 2018, 67). Within the FEP, homeostasis is *represented as* the behavior of an information system that maintains its states within desired values, rather than that of a physical system.

In summary, the FEP represents organisms as systems that perceive and act to maintain their physical integrity. It also provides an equation to model to these systems. One reason that supports my interpretation of the FEP as a model template is its transfer of elements from physics and information theory to biological and cognitive problems. One such element is the notion of surprise. As an epistemic resource and model template with conceptual and mathematical dimensions, the FEP serves as a resource for constructing models and target systems. One example is the agent-environment coupled system model of niche construction discussed in Chapter 3. In the final section of the chapter, I discuss the epistemic value of the FEP and raise skepticism about metaphysical interpretations that neglect its role as an epistemic instrument.

A Critique to the Metaphysics of the Predictive Mind

I have proposed that the FEP belongs to a modeling framework and is a model template that represents biological and cognitive processes as probabilistic inference tasks. However, this is not how the FEP is commonly understood. Instead, its epistemic role is often associated with its alleged status as a scientific principle. To support my view of the FEP as a model template, I will discuss how the notion of a principle has been understood in the literature.

The FEP has been understood in the literature as a scientific, heuristic, normative, and first metaphysical principle. All these interpretations agree that the FEP is neither explanatory nor falsifiable. The FEP describes how certain systems behave to maintain their organization. However, it does not explain how they do so. Rather, the FEP states that if a system maintains its organization, then it must act as if it were minimizing free energy. Thus, at its core, the FEP asserts that any self-organizing system can be *represented as* a free energy-minimizing system. I will argue that the FEP is best understood as a scientific, heuristic, and normative principle for constructing scientific models and target systems rather than as a metaphysical first principle. I will now explore these interpretations of the notion of a principle in more detail.

Scientific principles are basic statements that articulate a scientific understanding of a process, system, or phenomenon. For instance, the FEP states that a class of systems should be represented mathematically as a free energy-minimizing system. Such a statement must play a role in scientific domains; otherwise, it would be irrelevant. This appears to be the case for the FEP, which has been used to model cortical networks and construct targetless models, as discussed in Chapter 3. In that chapter, I demonstrated how the FEP can serve as a formal model of niche construction. The FEP models this behavior as that of a system that minimizes an informational quantity.

By framing target systems as information processing systems, the FEP imposes constraints on them. Green (2015, 633) defines constraints as follows:

> Constraints refer to conditions that at the same time limit and afford a certain scope of possible structures and functions that can be instantiated in a system of a particular type. Constraints do not explain how a phenomenon is produced by a mechanism in virtue of the interactions between component parts but rather denote the boundaries of such processes through reasoning about spaces of possibilities. Biological systems operate within different types of constraints, such as physical (mechanical and thermodynamic), structural (or organizational), functional, and developmental.

The constraints that the FEP imposes on systems are relative to its mathematical nature. As such, the FEP imposes informational constraints, not physical, biological, or cognitive ones. Therefore, the FEP cannot make empirical claims, be falsified, or posit kinds or mechanisms, as Green (2015) suggests. This raises the question of whether the FEP is empirically informative and, if not, how it is useful in scientific research.

These questions are addressed in Andrews's (2021) account of the FEP as a model structure. She acknowledges that the FEP is empirically uninformative and does not posit mechanisms. However, the FEP can be viewed as a formal tool or framework for scientific theorizing. As a model structure, the FEP serves as the baseline for constructing scientific models. Unlike a model template, a model structure lacks a conceptual dimension, making it a purely mathematical formalism. Thus, she argues that the demand that the FEP must be falsifiable rests on a category error.

For Andrews, the FEP functions as a formal structure that can be applied to target systems, whether they are biological or cognitive. When applied to these targets, scientists must add construals; depending on these construals, diverse scientific models are constructed. Thus, for Andrews, "a plethora of models are based on the formal structure of the FEP" (2021, 1). Andrews is partially correct in claiming that the FEP is a formal structure that acquires content depending on the target system to which it is applied. However, as a model template, the FEP itself has a conceptual dimension that is projected onto the target systems. Specifically, the FEP imposes informational constraints on these target systems, framing them as informational systems, as discussed earlier.

In summary, according to Andrews (2021), the FEP is a scientific principle because it provides a formal structure for constructing scientific models. Therefore, it is not a metaphysical principle that captures the essential properties of specific domains of reality. This scientific principle has been treated as a heuristic principle that transforms cognitive and biological problems into informational ones. Since it is not a metaphysical principle, it does not posit that cognitive and biological phenomena are informational in nature. Importantly, using this heuristic invariably results in the loss of empirical referents. As Andrews has noted:

> When we speak of annealing a model in statistical mechanics, ratcheting the temperature of the system up and down in the hopes of bumping it out of local minima, this does not refer to an act of literally injecting energy into a physical system the speed of particle motion. It is a statistical analogue of a physical process. Likewise, the energy and entropy of the FEP are formal analogues of concepts defined in thermodynamics and statistical mechanics with a long history of use in information

> theory, statistics, and machine learning, in which they have lost their correspondence to any measurable properties of physical systems (2021, 9).

The transdisciplinary transfer of concepts from thermodynamics and statistical mechanics to cognition and life research suggests that the FEP is a model template. Since empirical targets are lost in translation, if the FEP involves any ontology, it must be generic and related to the informational constraints it imposes on systems. As Green (2015, 632) states, principles of this kind "define the generic features of a class of systems that operate under a similar set of constraints."

Proponents of the FEP have also noted its role as a heuristic principle. According to Badcock, Friston, and Ramstead (2019), the FEP transforms the physical problem of avoiding deleterious phase transitions into the informational problem of avoiding surprising states. In doing so, the FEP conceptualizes cognitive and biological behaviors as tasks of minimizing informational entropy, or surprise. The FEP posits that the behavior of these systems is characterized by the expectation and avoidance of surprising states.

Wimsatt offers the following characterization of the use of heuristics in science:

> Applying a heuristic to a problem transforms the problem into a non-equivalent but intuitively related problem [...]. But it is not the same problem, so beware: answers to the transformed problem may not be answers to the original problem. However cognitive biases operative in learning and science may lead us to ignore this possibility and assert or assume that we have answered the original problem [...], leading to inflated or premature claims about the power of an approach (2006, 465)

Thus, the use of heuristics in science consists of transforming a complex problem into a simpler one. Wimsatt points out that the solution to the new problem is not necessarily a solution to the original problem. Within PP, the FEP transforms cognitive and biological problems into informational problems and makes them mathematically tractable. However, it does not address their specificity as cognitive or biological phenomena. At best, this transformation provides a "biologically plausible mathematical formulation of the evolution, development, form, and function of the brain" (Badcock, Friston, and Maxwell 2019, 105).

As discussed in Chapter 3, this kind of heuristic can be useful for targetless modeling in PP, which creates new epistemic resources. More broadly, it is essential for target construction because it projects the properties of information systems onto targets. How else can cognition be studied scientifically besides assuming that it involves information processing? The FEP provides a means to study cognition under such constraints, but it does not entail any sort of metaphysical view of cognitive or biological systems as informational in nature.

In addition to being interpreted as a heuristic principle, the FEP is often referred to as a normative principle or a model for process theories (Allen and Friston 2018). Process theories describe cognitive behaviors as the result of the interaction between different parts of systems. Normative models delineate the rules that govern these interactions. Simply put, normative models specify how a system should behave if it belongs to a particular class of systems. However, normative models do not explain how systems work. Rather, they stipulate only that the system's behavior conforms to a certain rule. Schwartenbeck et al. (2013, 1) present the FEP as a normative model.

> The Free energy Principle [...] is normative in the sense that it provides a well-defined objective function (variational free energy) that is optimized both by action and perception. Having said this, the normative aspect of free energy minimization (and implicit active inference) is complemented by a neuronally plausible implementation scheme, in the form of predictive coding.

According to these authors, the FEP defines an objective function—minimization of variational free energy—and states that systems that perceive and act perform it. Process theories, such as Predictive Coding, specify how certain systems may implement this function. However, normative models and process theories are not necessarily related. Another normative model besides the FEP could provide a different objective function, and the same applies to process theories.

In my view, the scientific value of a normative model is not tied to its ability to capture an objective function in a particular system. An idealized model of biological systems that frames them as systems that minimize surprise does not become a good model simply by virtue of capturing something these systems actually do. In short, normativity has nothing to do with representing an objective function in nature. In my account, normativity is contingent on the achievements of modeling practices. As Colyvan (2013, 1348) states, "Not every idealization has an a priori normative justification." Therefore, merely framing cognitive behavior as an informational task does not make this heuristic strategy a good epistemic practice in itself. Justification comes *a posteriori* from what is achieved by using a model for a specific purpose.

Before concluding this chapter, I would like to revisit the metaphysical interpretation of the FEP as a foundational principle for cognitive and biological sciences. This interpretation mistakenly assumes that frameworks are explanatory (Kirchhoff 2018), implying that the FEP offers a universal explanation for cognition and life. Contesting this position, I have argued that the FEP is merely an epistemic resource for constructing models and target systems of biological and cognitive phenomena. According to the metaphysical interpretation, the FEP is reductionist because it explains diverse phenomena as governed by the same rule. This

interpretation treats free energy minimization as a natural law underlying certain classes of phenomena. In Chapter 4, I articulated an ontological agnosticism toward metaphysical interpretations such as this one. The argument was that one could believe models based on the FEP are good models because they capture something in nature. Nevertheless, one could remain agnostic about this possibility and still use these models if they prove useful in scientific research.

In my view, any metaphysical interpretation of the FEP must acknowledge that, from the perspective of scientific practice, positing an informational principle underlying natural cognitive and biological phenomena is unwarranted. I have suggested, instead, that informational and statistical tools are becoming standard epistemic resources for studying cognitive and biological systems. These resources are integral to how scientists think about and imagine these systems. This explains why these resources are sometimes reified as properties of phenomena.

The standard interpretation of PP is that it is a theory of the brain as a statistical machine. However, as I have suggested in this chapter, PP does not have to be interpreted as a reductionist program — that is, as a theory of "the underlying neural processes" (Wiese and Metzinger 2017, 2) — but rather as a reductionist heuristic (see Wimsatt 2006). One can treat PP models as potentially valuable epistemic resources for addressing scientific problems without assuming the existence of predictive mechanisms in the brain.

In this dissertation, I have defended ontological agnosticism—the view that scientific models of cognition are ontologically neutral. According to this view, the research objects of cognitive science—the target systems—are constructed within scientific modeling practices. In this chapter, I have argued that the epistemic resources of a modeling framework—its concepts, methods, and model templates—serve as the means by which target systems are constructed.

I have argued that the construction of target systems does not need to reflect the complex, multilevel nature of phenomena. This assumption is often found in predictive mind literature, such as when the FEP is presented as a fundamental law from which the PP, a theory of brain function, is derived. Likewise, Bayesianism does not theoretically depend on either the FEP or the PP. Any dependency among them exists only as a stipulation within scientific practice.

Consider the construct of adaptive systems. In the PP framework, brain networks are stipulated to be a subclass of adaptive systems. Since adaptive systems are modeled in the PP framework using the FEP, some assume that the FEP is more general than PP. However, this assumption mistakenly interprets a hierarchy in nature as implying a hierarchy in theories. Scientific principles, theories, and models are not hierarchically related because

they target phenomena at different levels. They do not reflect an objective hierarchy in the world. Theory is not a mirror of reality.

My skepticism toward such hierarchical dependencies also stems from the idea that epistemic resources are not tied to particular targets, whether they are concrete, such as brain networks, or abstract, such as adaptive systems. For instance, the FEP does not automatically apply to cognitive systems simply because it captures the mathematical structures of adaptive systems, from which cognitive systems are a subclass.

Ontological agnosticism responds to philosophical debates about the predictive mind, as discussed in Chapter 4, as well as to realist interpretations of PP. In particular, it challenges the assumption that framing cognition as probabilistic inference and analyzing it with statistical tools means cognition itself is a statistical phenomenon—if such a thing exists. Likewise, it resists metaphysical interpretations of the FEP as a first principle for cognition or life. Finally, it abandons the idea that cognitive science aims to explain complex phenomena at multiple ontological levels.

As an alternative, I have articulated a view in which the FEP and Bayesianism are components of a modeling framework—a constellation of epistemic resources for constructing models and defining target systems. In the PP modeling framework, these resources represent cognition as probabilistic inference tasks. These are models of heuristically transformed cognitive tasks, not models of mental phenomena or mechanisms.

Conclusions

In this dissertation, I explored two different ideas about the modeling mind. First, I concentrated on the view that minds create models of the world and of themselves to perceive and act. This is the central idea of Predictive Processing. This view, while arguably metaphysical, is consistent with several scientific theories in the cognitive sciences. Second, I explored the idea that scientific modeling is a cognitive and epistemic practice. I argued that modeling is essential to the cognitive sciences because cognition becomes amenable to scientific investigation only when it is conceptualized as the behavior of a target system within a modeling practice.

I expect to have convinced the reader that studying the cognitive sciences with the resources of the philosophy of scientific modeling offers a fresh perspective on several philosophical problems raised by that discipline. This perspective differs from the traditional answers offered by philosophers of mind. Rather than asking which ontology of the mind can be inferred from the best scientific theories or which theory of cognition is most realistic, I sought to understand how knowledge is produced in the cognitive sciences and what kind of knowledge is produced through modeling practices.

I argued that the kind of scientific knowledge produced by modeling cognition as probabilistic inference stems from particular research practices. I did not claim that Predictive Processing models are superior to other scientific models because they accurately depict cognition. Rather, I claimed that they are helpful epistemic instruments because of what they enable scientists to achieve. To explain knowledge production in this field, I developed three concepts: target construction, modeling framework, and ontological agnosticism. These concepts account for the creation of research objects in this field as an activity that relies on conventional epistemic resources, as well as the epistemic implications of using them.

In particular, I argued that scientists can use the conceptual, mathematical, and computational epistemic resources of Predictive Processing to construct target systems. These targets are cognitive tasks, which are heuristically transformed problems. Compared to cognitive phenomena, these targets are simpler research objects that can be investigated using the relevant epistemic resources.

Modeling cognition is not about representing cognitive phenomena or mechanisms in nature, but about projecting the properties of a model onto a target system. In Predictive Processing, this practice transforms cognitive behavior into tasks of probabilistically inferring the most likely cause of an observation given a generative model. The case of Predictive Processing illustrates how target construction works: by projecting the properties of the epistemic resources involved in constructing a model into a target system, the same resources become the appropriate means for investigating that target. In short, the model and the target system are co-created in scientific modeling within the cognitive sciences.

The target cognitive systems, conceptualized as probabilistic inference tasks, can thus be studied using the mathematical, computational, and other epistemic resources of the framework. I have argued in Chapter 6 that this is the role of Bayesian and free-energy model templates in Predictive Processing. On a conceptual level, the targets constructed using these resources, such as visual perception or emotion, become observable tasks that result from the activities of some systems that infer the external or internal causes of their sensations based on prior knowledge about their own and environmental states. These systems update such prior knowledge by computing the error or difference between their predictions and the sensations.

On my account, such conceptualizations are ontologically neutral. They make cognitive phenomena objects of inquiry, but they do not provide a substantive view of what cognition is beyond its meaning in science. In other words, just because modeling produces knowledge does not mean that it captures the intrinsic nature of cognition. For example, the fact that a computational model allows a robot to effectively simulate visual perception does not mean that this model tells us anything about the nature of visual perception itself. Conceptualizations do not aim to capture the essential properties of cognitive phenomena, but rather to make them amenable to investigation, intervention, or simulation. To this end, modelers construct them as solvable tasks. In Predictive Processing, it is only because cognition is conceptualized as a task of probabilistic inference that it can be investigated with statistical tools well-suited for dealing with this type of inference.

The construction of target systems and the choice of resources for scientific modeling are not arbitrary. They follow the conventions of the modeling framework involved. I have defined modeling frameworks as constellations or sets of conventionally related epistemic resources used for model construction. Scientists working within a framework share a *modus operandi* for investigating and answering uniformly formulated problems. On my account, the supposedly general theories of cognitive science, such as Cognitivism, Connectionism, 4E Cognition, or Predictive Processing, are better understood as modeling frameworks. Rather than providing substantive views of cognition, they provide scientists

with common strategies, concepts, models, methods, and other epistemic instruments for modeling target systems.

My concept of a modeling framework serves to challenge the common narrative that the cognitive sciences are an interdisciplinary field that seeks to integrate partial explanations of cognition into a unified theory. This narrative assumes that cognitive scientists seek to explain cognition by constructing scientific models that represent the complex mechanisms underlying cognitive processes. This narrative is popular in the Predictive Processing literature, where proponents often take literally the claim that the brain reduces prediction error or processes information.

Instead, I have argued that the epistemic achievements of the cognitive sciences cannot be reduced to mechanistic and/or other forms of explanation. Simulation, prediction, and intervention are examples of heterogeneous epistemic gains produced by modeling in the field. Rather than providing means for unified interdisciplinary explanation, frameworks provide transdisciplinary epistemic resources, including generic concepts, methods such as computational and mathematical procedures, and experiments. They are transdisciplinary because they are transferred from some disciplines to study cognitive target systems that are not disciplinarily bound.

This perspective has led me to reconceptualize cognition itself, not as a complex, real-world phenomenon, but as a research object constrained by modeling tools and analogies such as "information processing" or "probabilistic inference." Finally, my concept of a modeling framework suggests that the cognitive sciences are not unified by an ultimate theory of cognition, as if there were fundamental principles from which any model could be derived. Research in this area is opportunistic because it depends on specific problems and available epistemic resources. If there is unity, it comes from the standardization of ways of formulating and answering questions.

Through my analysis of scientific modeling in cognitive science, I have concluded that ontological agnosticism regarding mental modeling is the most reasonable stance. According to this view, scientific theories of cognition do not entail metaphysical commitments. Rather, they employ heuristics and analogies, such as viewing cognition as computation or probabilistic inference, to address scientific problems without asserting ontological truths.

In the cognitive sciences, these heuristic strategies generally rely on established analogies and/or metaphors to build workable research objects. These include the ideas that cognition is computation, the notion of cognitive mechanisms, and, in Predictive Processing, the analogy between cognition and probabilistic inference. In principle, these ideas can be

accepted both as scientific heuristics and as metaphysical claims about cognition. However, in scientific practices, they are only necessary for conceptualizing cognition, so they are neutral to metaphysical questions about whether cognition is intrinsically information processing or probabilistic inference. In short, a scientist can believe in the existence of mental mechanisms or be realistic about mental models, but this is completely irrelevant to gaining scientific knowledge under these assumptions.

My ontological agnosticism is consistent with the historical use of such metaphors, as discussed in Helmholtz's theory of perception or Ashby's systems theory. In Helmholtz's theory of perception, the analogy between inference and perception was proposed as a heuristic to mediate between psychological and physiological theories of the phenomenon, but it was not intended as an ontological description of the phenomenon.

Studying a second theoretical antecedent of Predictive Processing—Ashby's systems theory—allowed me to position the thesis of ontological agnosticism as a result of the scientific theorization of cognition. Systems theory provided early cognitive scientists with a framework to characterize their research objects. Not only did Ashby propose that intelligent systems process information, he also proposed that they are regulated by feedback and self-correction mechanisms. According to his theory, systems that maintain themselves in complex environments must create models of their environments.

Ontological agnosticism can be deduced from his theory because of the concept of a system. In cybernetics, systems are not entities in the world; rather, they are perspectives or ways of approaching reality. Systems produce observable behaviors through the interaction of their parts. When cognition is conceptualized as the behavior of a system, a heuristic transformation of a real phenomenon into a simpler one occurs. Thus, ontological agnosticism is grounded in the idea that cognitive science works with heuristically transformed problems.

My main question, which concerns the epistemic gain from modeling cognition in Predictive Processing, was answered directly in Chapter 3. I argue that this gain does not depend exclusively on some inherent quality of Predictive Processing models to capture the actual mechanisms underlying cognition. A conceptualization of cognition as inferential, predictive, and model-based is not in itself sufficient to solve any scientific problem or to capture anything in nature. As I have argued, epistemic gain is manifold and depends on how scientists construct and manipulate models.

I analyzed two case studies of the construction and manipulation of Predictive Processing models. The first case study was the construction of the Agent Observation Network model (Urgen and Saygin 2020), a scientific model of cortical activity in some areas of the brain that

activate during action perception. The type of modeling used in this case corresponds to what is known in the philosophy of science as target-oriented modeling. This model has as its target system a hypothetical system in the brain whose components interact in response to the behavior described above. Importantly, the assumption that this network functions as a probabilistic and inferential system allowed the modelers to make hypotheses about its behavior and scientific predictions about the activations of its parts.

Contrary to the scientists' own understanding of the epistemic gain of this scientific practice, I suggested that the AON model did not provide evidence for the existence of predictive mechanisms in the brain. This was not only because the model does not map structures or parts in the brain, but also because it predicted patterned behavior without specifying the causes responsible for it. Nevertheless, this case study illustrated the usefulness of Predictive Processing for scientific practices. For example, the assumption that the brain processes information in a hierarchical fashion helped the modelers (Urgen, Kutas, and Saygin 2018) in making accurate scientific predictions. Similarly, concepts such as prediction error helped Saygin et al. (2012) design an experiment in which they discovered a specific activation in the brain in response to a mismatch between perceptual anticipations and sensory evidence.

The second case study (Bruineberg et al. 2018) corresponded to the AECS model, a targetless model of niche construction under the Free Energy Principle. Unlike the AON model, this model did not help scientists formulate testable hypotheses or scientific predictions. The main point of the construction of the model was to show that niche construction, when defined in generic terms, can be described by free-energy equations. I characterized the construction of the AECS model as a two-step process of constructing a target behavior and attributing structure (the FEP). I argued that the model is targetless because it does not represent any real phenomena and was constructed to transfer a model template from biophysics to niche construction. The constructed target is an artificial agent interacting in an environment that changes in response to the effects of actions. The attribution of structure involves the application of the theoretical resources of the FEP to the coupled agent-environment system. Finally, I pointed out that this strategy transforms the complex problem of niche construction into a simpler informational problem.

The epistemic gain of such a practice is unclear since it does not provide information about real things. However, there are cases in which scientific practices consist of creating tools that scientists later use to study reality. In other words, a model must first be constructed before it can be used to study a target system. In the considered case, targetless modeling involves applying a model template to create a mathematical representation of niche

construction. In this way, a target system is created that can later become the subject of research in a target-directed modeling practice.

The AON and AECS model constructions are contrasting cases of scientific modeling under the Predictive Processing framework. They illustrate the variety of epistemic gains that can be achieved by relatively similar models, gains that range from the creation of a generic target system to the formulation of specific scientific predictions and hypotheses. In all cases, however, Predictive Processing models are not explanatory by virtue of their ability to represent any mathematical, computational, or informational property of phenomena. Importantly, I am not arguing here that explanations cannot be the result of modeling practices, only that the computational or mathematical structures of these models need not be designators of the computational or mathematical structures of cognition.

In summary, I have argued that models of cognition provide knowledge not because they represent cognitive processes or mechanisms in nature, but rather, because of how they combine diverse epistemic resources to solve specific problems. This epistemic gain is specific to modeling practices that create target systems. The forms of knowledge production, on the other hand, are general. Science conducted within a modeling framework uses the same conceptual, mathematical, and computational resources to investigate various cognitive phenomena. In doing so, science projects theory onto reality.

If this thesis is correct, it has two implications for philosophy and the sciences of the mind. For philosophy, it offers an alternative between the realist and anti-realist views of theses such as the existence of mental models, in which the means of explaining cognition should not be confused with cognitive phenomena. Philosophers must be careful not to reify the means of explaining reality with reality itself. To understand how knowledge of the mind is produced, they must take a closer look at the scientific practices. I argued in Chapter 4 that the philosophical debates about Predictive Processing, whether it is a representationalist or embodied theory of cognition, whether it is internalist or externalist, fail to recognize that it is ontologically neutral. We cannot say that the cognitive sciences favor a particular ontology of mind. That, I think, is an error in our judgment.

It also serves the philosophy of predictive mind to understand it not as a representationalist or anti-representationalist theory of cognition; I have argued that PP is neither. It is not a theory, but a modeling framework that is neutral about the realization of cognition and that tolerates multiple ontologies of mind. In scientific modeling, conceptualizations are operational descriptions that are useful in research practice and do not designate or represent any mental kind, mechanism, or structure.

As for the sciences of the mind, the thesis has explained that the research objects they work with are cognitive systems that are purposefully constructed to address specific problems. As such, it urges not to take seriously the idea that all cognitive behavior can be explained with a single theory (e.g., Friston et al. 2010; Hohwy 2018). On the positive side, the dissertation has provided an account of the scientific knowledge gained by using these models. It does not argue against the assumption that predictive mechanisms underlie cognition but warns that this is only one of many possible assumptions that scientists can make when constructing models. Unfortunately, this assumption has led scientists to believe that mechanistic explanations are the only valuable results of modeling practices. However, the examination of modeling practices that has been undertaken has shown that models need not be mechanistic to yield epistemic gains.

Finally, the dissertation has drawn attention to the crucial but largely neglected role of scientific conceptualizations in the sciences of mind. Such conceptualizations not only frame the objects of inquiry, but also make them amenable to, for example, exploration, manipulation, or simulation. It is the transformation of cognitive behavior into solvable tasks that makes it possible to gain control over target systems. Conceptualization is an integral part of this process.

In conclusion, model construction plays a crucial yet often overlooked role in the cognitive sciences. Through my examination of model construction in Predictive Processing, I have filled a gap in the literature and made a modest contribution to a prominent research program in the cognitive sciences. My contribution has been a critical and skeptical evaluation of its theoretical ambitions, assumptions, and philosophical interpretations. I have also provided an informed account of the heterogeneous epistemic gains afforded by modeling cognition as probabilistic inference. Regarding the philosophy of Predictive Processing, I expect to convince the reader that metaphysical views of cognition as predictive arise from conflating the properties of scientific models with those of phenomena. This conflation occurs when we are unaware of the heuristics involved in scientific modeling. Acknowledging this allows us to understand that Predictive Processing tolerates competing ontologies of cognition and mind.

Bibliography

Aaby, Bendik Hellem, and Hugh Desmond. 2021. “Niche Construction and Teleology: Organisms as Agents and Contributors in Ecology, Development, and Evolution.” *Biology and Philosophy* 36 (5). https://doi.org/10.1007/s10539-021-09821-2.

Adams, Rick A., Klaas Enno Stephan, Harriet R. Brown, Christopher D. Frith, and Karl J. Friston. 2013. “The Computational Anatomy of Psychosis.” *Frontiers in Psychiatry* 4. https://doi.org/10.3389/fpsyt.2013.00047.

Aizawa, Ken. 2017. “Cognition and Behavior.” *Synthese* 194 (11): 4269–88. https://doi.org/10.1007/s11229-014-0645-5.

Allen, Micah, and Karl J. Friston. 2018. “From Cognitivism to Autopoiesis: Towards a Computational Framework for the Embodied Mind.” *Synthese* 195 (6): 2459–82. https://doi.org/10.1007/s11229-016-1288-5.

Andersen, Hanne. 2012. “Conceptual Development in Interdisciplinary Research.” In *Scientific Concepts and Investigative Practice*, edited by Uljana Feest and Friedrich Steinle, 271–92. Berlin, Boston, Mass.: De Gruyter. https://doi.org/10.1515/9783110253610.271.

———. 2016. “Collaboration, Interdisciplinarity, and the Epistemology of Contemporary Science.” *Studies in History and Philosophy of Science Part A* 56 (April): 1–10. https://doi.org/10.1016/j.shpsa.2015.10.006.

Anderson, Michael L., and Tony Chemero. 2013. “The Problem with Brain GUTs: Conflation of Different Senses of ‘Prediction’ Threatens Metaphysical Disaster.” *Behavioral and Brain Sciences* 36 (3): 204–5. https://doi.org/10.1017/S0140525X1200221X.

Andrews, Mel. 2021. “The Math is not the Territory: Navigating the Free Energy Principle.” *Biology and Philosophy* 36: 30. https://doi.org/10.1007/s10539-021-09807-0.

Asaro, Peter. 2005. “On The Origins of Synthetic Mind: Working Models, Mechanisms, and Simulations.” PhD Thesis, Urbana: University of Illinois at Urbana-Champaign.

———. 2008. “From Mechanisms of Adaptation to Intelligence Amplifiers: The Philosophy of W. Ross Ashby.” In *The Mechanical Mind in History*, edited by Philip Husbands, Owen Holland, and Michael Wheeler, 149–84. Cambridge, Mass.: The MIT Press. https://doi.org/10.7551/mitpress/9780262083775.003.0007.

Ashby, W. Ross. 1940. “Adaptiveness and Equilibrium.” *The British Journal of Psychiatry 86*: 478–83.

———. 1947. “The Nervous System as Physical Machine: With Special Reference to the Origin of Adaptive Behavior.” *Mind* 56 (221): 44–59.

———. 1956. *An Introduction to Cybernetics*. London: Chapman and Hall.

———. 1960. *Design for a Brain. The Origin of Adaptive Behaviour*. New York, NY: John Wiley and Sons.

———. 1991. “Requisite Variety and Its Implications for the Control of Complex Systems.” In *Facets of Systems Science*, edited by George J. Klir, 405–17. Boston, Mass.: Springer US. https://doi.org/10.1007/978-1-4899-0718-9_28.

———. n.d. “Aphorisms.” http://www.rossashby.info/aphorisms.html.

Atkinson, R.C., and R.M. Shiffrin. 1968. “Human Memory: A Proposed System and Its Control Processes.” In *Psychology of Learning and Motivation*, vol. 2, edited by Kenneth W. Spence, and Janet Taylor Spence, 89–195. Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3.

Axtell, R., and Joshua Epstein. 1994. “Agent-Based Modeling: Understanding Our Creations.” *The Bulletin of the Santa Fe Institute* 9 (4): 28–32.

Badcock, Paul B., Karl J. Friston, Maxwell J. D. Ramstead, Annemie Ploeger, and Jakob Hohwy. 2019. “The Hierarchically Mechanistic Mind: An Evolutionary Systems Theory of the Human Brain, Cognition, and Behavior.” *Cognitive, Affective, and Behavioral Neuroscience* 19 (6): 1319–51. https://doi.org/10.3758/s13415-019-00721-3.

Badcock, Paul B., Karl J. Friston, and Maxwell J.D. Ramstead. 2019. “The Hierarchically Mechanistic Mind: A Free-Energy Formulation of the Human Psyche.” *Physics of Life Reviews* 31 (December): 104–21. https://doi.org/10.1016/j.plrev.2018.10.002.

Bailer-Jones, Daniela. 2009. *Scientific Models in the Philosophy of Science*. Pittsburgh, PA: University of Pittsburgh press.

Bear, Mark F., Barry W. Connors, and Michael A. Paradiso. 2007. *Neuroscience: Exploring the Brain*. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins.

Bechtel, William. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York, NY: Taylor and Francis.

———. 2016. “Investigating Neural Representations: The Tale of Place Cells.” *Synthese* 193 (5): 1287–1321. https://doi.org/10.1007/s11229-014-0480-8.

Beck, Lukas, and Marcel Jahn. 2021. “Normative Models and Their Success.” *Philosophy of the Social Sciences* 51 (2): 123–50. https://doi.org/10.1177/0048393120970908.

Beni, Majid D. 2021. “A Critical Analysis of Markovian Monism.” *Synthese* 199 (3–4). https://doi.org/10.1007/s11229-021-03075-x.

Bergmann, Till, Rick Dale, Negin Sattari, Evan Heit, and Harish S. Bhat. 2017. “The Interdisciplinarity of Collaborations in Cognitive Science.” *Cognitive Science* 41 (5): 1412–18. https://doi.org/10.1111/cogs.12352.

Bermúdez, José Luis. 2014. *Cognitive Science: An Introduction to the Science of the Mind*. 2nd ed. Cambridge, UK: Cambridge University Press.

Bickhard, Mark H., and Loren. Terveen. 1995. *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*. *Advances in Psychology*, vol. 109. Amsterdam: Elsevier.

Blake, Randolph, and Maggie Shiffrar. 2007. “Perception of Human Motion.” *Annual Review of Psychology* 58 (1): 47–73. https://doi.org/10.1146/annurev.psych.57.102904.190152.

Bonabeau, Eric. 2002. “Agent-Based Modeling: Methods and Techniques for Simulating Human Systems.” *Proceedings of the National Academy of Sciences* 99 (suppl. 3): 7280–87. https://doi.org/10.1073/pnas.082080899.

Brand, Jay L., Dennis H. Holding, and Paul D. Jones. 1987. “Conditioning and Blocking of the McCollough Effect.” *Perception and Psychophysics* 41 (4): 313–17. https://doi.org/10.3758/BF03208232.

Brette, Romain. 2022. “Brains as Computers: Metaphor, Analogy, Theory or Fact?” *Frontiers in Ecology and Evolution* 10 (April): 878729. https://doi.org/10.3389/fevo.2022.878729.

Brooks, Jessica X., and Kathleen E. Cullen. 2019. “Predictive Sensing: The Role of Motor Signals in Sensory Processing.” *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 4 (9): 842–50. https://doi.org/10.1016/j.bpsc.2019.06.003.

Bruin, Leon de, Albert Newen, and Shaun Gallagher, eds. 2018. *The Oxford Handbook of 4E Cognition*. Oxford, UK: Oxford University Press.

Bruineberg, Jelle, Julian Kiverstein, and Erik Rietveld. 2018. “The Anticipating Brain is Not a Scientist: The Free-Energy Principle from an Ecological-Enactive Perspective.” *Synthese* 195 (6): 2417–44. https://doi.org/10.1007/s11229-016-1239-1.

Bruineberg, Jelle, Erik Rietveld, Thomas Parr, Leendert van Maanen, and Karl J Friston. 2018. “Free-Energy Minimization in Joint Agent-Environment Systems: A Niche Construction Perspective.” *Journal of Theoretical Biology* 455 (October): 161–78. https://doi.org/10.1016/j.jtbi.2018.07.002.

Buchak, Lara Marie. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

Buckley, Christopher L., Chang Sub Kim, Simon McGregor, and Anil K. Seth. 2017. “The Free Energy Principle for Action and Perception: A Mathematical Review.” *Journal of Mathematical Psychology* 81 (December): 55–79. https://doi.org/10.1016/j.jmp.2017.09.004.

Cain, M. J. 2016. *The Philosophy of Cognitive Science*. Cambridge, UK ; Malden, MA: Polity.

Cao, Rosa. 2020. “New Labels for Old Ideas: Predictive Processing and the Interpretation of Neural Signals.” *Review of Philosophy and Psychology* 11 (3): 517–46. https://doi.org/10.1007/s13164-020-00481-x.

Cariani, Peter A. 2009. “The Homeostat as Embodiment of Adaptive Control.” *International Journal of General Systems* 38 (2): 139–54. https://doi.org/10.1080/03081070802633593.

Carney, James. 2020. “Thinking Avant La Lettre: A Review of 4E Cognition.” *Evolutionary Studies in Imaginative Culture* 4 (1): 77–90. https://doi.org/10.26613/esic.4.1.172.

Carrillo, Natalia, and Tarja Knuuttila. 2023. “Mechanisms and the Problem of Abstract Models.” *European Journal for Philosophy of Science* 13 (3): 1–19. https://doi.org/10.1007/s13194-023-00530-z.

Cermeño-Aínsa, Sergio. 2020. “The Cognitive Penetrability of Perception: A Blocked Debate and a Tentative Solution.” *Consciousness and Cognition* 77 (102838): 1–23 https://doi.org/10.1016/j.concog.2019.102838.

Chang, Hasok. 2022. *Realism for Realistic People: A New Pragmatist Philosophy of Science*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781108635738.

Chater, Nick, Joshua B. Tenenbaum, and Alan Yuille. 2006. “Probabilistic Models of Cognition: Conceptual Foundations.” *Trends in Cognitive Sciences* 10 (7): 287–91. https://doi.org/10.1016/j.tics.2006.05.007.

Chemero, Anthony. 2000. “Anti-Representationalism and the Dynamical Stance.” *Philosophy of Science* 67 (4): 625–47. https://doi.org/10.1086/392858.

———. 2001. “Dynamical Explanation and Mental Representations.” *Trends in Cognitive Sciences* 5 (4): 141–42. https://doi.org/10.1016/S1364-6613(00)01627-2.

———. 2009. *Radical Embodied Cognitive Science*. Cambridge, Mass.: The MIT Press. https://doi.org/10.7551/mitpress/8367.001.0001.

Churchland, Paul M. 1982. “Is Thinker a Natural Kind?” *Dialogue* 21 (2): 223–38. https://doi.org/10.1017/S001221730001636X.

Clark, Andy. 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge, Mass.: The MIT Press.

———. 2013. “Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science.” *Behavioral and Brain Sciences* 36 (3): 181–204. https://doi.org/10.1017/S0140525X12000477.

———. 2014. “Perceiving as Predicting.” In *Perception and Its Modalities*, edited by Dustin Stokes, Mohan Matthen, and Stephen Biggs, 23–43. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199832798.003.0002.

———. 2015. “Predicting Peace: The End of the Representation Wars-A Reply to Michael Madary.” In *Open MIND*, edited by T. Metzinger and J. M. Windt. Frankfurt: MIND Group. https://doi.org/10.15502/9783958570979.

———. 2015b. “Radical Predictive Processing.” *The Southern Journal of Philosophy* 53 (S1): 3–27. https://doi.org/10.1111/sjp.12120.

———. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217013.001.0001.

———. 2023. *The Experience Machine: How Our Minds Predict and Shape Reality*. New York, NY: Pantheon Books.

Clark, Andy, and David Chalmers. 1998. “The Extended Mind.” *Analysis* 58 (1): 7–19. https://doi.org/10.4324/9781315150420-2.

Clark, Andy, and Rudi Lutz. 1992. *Connectionism in Context*. London: Springer.

Clavel Vázquez, María Jimena. 2020. “A Match made in Heaven: Predictive Approaches to (an Unorthodox) Sensorimotor Enactivism.” *Phenomenology and the Cognitive Sciences* 19 (4): 653–84. https://doi.org/10.1007/s11097-019-09647-0.

Cobb, Matthew. 2020. *The Idea of the Brain: The Past and Future of Neuroscience*. New York, NY: Basic Books.

Colding, Johan, and Stephan Barthel. 2019. “Exploring the Social-Ecological Systems Discourse 20 Years Later.” *Ecology and Society* 24 (1): art2. https://doi.org/10.5751/ES-10598-240102.

Colombo, Matteo, Lee Elkin, and Stephan Hartmann. 2021. “Being Realist about Bayes, and the Predictive Processing Theory of Mind.” *The British Journal for the Philosophy of Science* 72 (1): 185–220. https://doi.org/10.1093/bjps/axy059.

Colombo, Matteo, and Stephan Hartmann. 2017. “Bayesian Cognitive Science, Unification, and Explanation.” *British Journal for the Philosophy of Science* 68 (2): 451–84. https://doi.org/10.1093/bjps/axv036.

Colombo, Matteo, and Peggy Seriès. 2012. “Bayes in the Brain -On Bayesian Modelling in Neuroscience.” *British Journal for the Philosophy of Science* 63 (3): 697–723. https://doi.org/10.1093/bjps/axr043.

Colombo, Matteo, and Cory Wright. 2021. “First Principles in the Life Sciences: The Free-Energy Principle, Organicism, and Mechanism.” *Synthese* 198 (S14): 3463–88. https://doi.org/10.1007/s11229-018-01932-w.

Colyvan, Mark. 2013. “Idealisations in Normative Models.” *Synthese* 190 (8): 1337–50. https://doi.org/10.1007/s11229-012-0166-z.

Conant, Roger C., and W. Ross Ashby. 1970. “Every Good Regulator of a System must be a Model of that System.” *International Journal of Systems Science* 1 (2): 89–97. https://doi.org/10.1080/00207727008920220.

Cooper, Rachel. 2005. “Thought Experiments.” *Metaphilosophy* 36 (3): 328–47. https://doi.org/10.1111/j.1467-9973.2005.00372.x.

Cooper, Richard P. 2019. “Multidisciplinary Flux and Multiple Research Traditions Within Cognitive Science.” *Topics in Cognitive Science* 11 (4): 869–79. https://doi.org/10.1111/tops.12460.

Corcoran, Andrew W., Giovanni Pezzulo, and Jakob Hohwy. 2020. “From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and the Origins of Cognition.” *Biology and Philosophy* 35 (3): 32. https://doi.org/10.1007/s10539-020-09746-2.

Corlett, Philip R., Aprajita Mohanty, and Angus W. MacDonald. 2020. “What We think About When We think About Predictive Processing.” *Journal of Abnormal Psychology* 129 (6): 529–33. https://doi.org/10.1037/abn0000632.

Courville, Aaron C., Nathaniel D. Daw, and David S. Touretzky. 2006. “Bayesian Theories of Conditioning in a Changing World.” *Trends in Cognitive Sciences* 10 (7): 294–300. https://doi.org/10.1016/j.tics.2006.05.004.

Crane, Tim. 2003. *The Mechanical Mind. A Philosophical Introduction to Minds, Machines, and Mental Representations*. London, New York, NY: Routledge.

Craver, Carl F. 2006. “When Mechanistic Models explain.” *Synthese* 153 (3): 355–76. https://doi.org/10.1007/s11229-006-9097-x.

———. 2007. *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press. https://doi.org/10.1136/jnnp.40.1.101.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Dale, Rick, Eric Dietrich, and Anthony Chemero. 2009. “Explanatory Pluralism in Cognitive Science.” *Cognitive Science* 33 (5): 739–42. https://doi.org/10.1111/j.1551-6709.2009.01042.x.

De Haan, Sanneke. 2020. *Enactive Psychiatry*. Cambridge University Press. https://doi.org/10.1017/9781108685214.

DeFelipe, Javier. 2002. “Microstructure of the Neocortex: Comparative Aspects.” *Journal of Neurocytology* 31 (3/5): 299–316. https://doi.org/10.1023/A:1024130211265.

Demeter, Tamás. 2009. “Two Kinds of Mental Realism.” *Journal for General Philosophy of Science* 40 (1): 59–71. https://doi.org/10.1007/s10838-009-9090-4.

Den Ouden, Hanneke E. M., Peter Kok, and Floris P. De Lange. 2012. “How Prediction Errors shape Perception, Attention, and Motivation.” *Frontiers in Psychology* 3. https://doi.org/10.3389/fpsyg.2012.00548.

Dewhurst, Joe. 2019. “British Cybernetics.” In *The Routledge Handbook of the Computational Mind*, edited by Mark Sprevak and Matteo Colombo, 38–51. London, New York, NY: Routledge, Taylor and Francis Group.

Dreyfus, Hubert L. 1971. “Phenomenology and Mechanism.” *Noûs* 5 (1): 81. https://doi.org/10.2307/2214453.

Egan, Frances. 2014. “How to think about Mental Content.” *Philosophical Studies* 170 (1): 115–35. https://doi.org/10.1007/s11098-013-0172-0.

Elkin, Lee, and Karolina Wiśniowska. 2022. “Too Rational: How Predictive Coding’s Success risks Harming the Mentally Disordered and Ill.” *Journal of NeuroPhilosophy* 1(1), 31-36. https://doi.org/10.5281/ZENODO.6637734.

Embree, Lester. 2010. “Disciplinarity in Phenomenological Perspective.” *Indo-Pacific Journal of Phenomenology* 10 (2): 1–5. https://doi.org/10.2989/IPJP.2010.10.2.2.1083.

Encyclopaedia Britannica. 2019. “Equilibrium.” In *Encyclopaedia Britannica*. https://www.britannica.com/science/equilibrium-physics.

Engel, Andreas K., Pascal Fries, and Wolf Singer. 2001. “Dynamic Predictions: Oscillations and Synchrony in Top-Down Processing.” *Nature Reviews Neuroscience* 2 (10): 704–16. https://doi.org/10.1038/35094565.

Epstein, Joshua M., and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, Mass.: The MIT Press.

Ewuzie, Ugochukwu, Oladotun Paul Bolade, and Abisola Opeyemi Egbedina. 2022. “Application of Deep Learning and Machine Learning Methods in Water Quality Modeling and Prediction: A Review.” In *Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering*, edited by Gonçalo Marques, and Joshua O. Ighalo, 185–218. Academic Press. https://doi.org/10.1016/B978-0-323-85597-6.00020-3.

Fabry, Regina E. 2018. “Betwixt and between: The Enculturated Predictive Processing Approach to Cognition.” *Synthese* 195 (6): 2483–2518. https://doi.org/10.1007/s11229-017-1334-y.

Facchin, Marco. 2021. “Are Generative Models Structural Representations?” *Minds and Machines* 31 (2): 277–303. https://doi.org/10.1007/s11023-021-09559-6.

Farrell, Simon, and Stephan Lewandowsky. 2018. *Computational Modeling of Cognition and Behavior*. New York, NY: Cambridge University Press.

Fernandez-Duque, Diego, and Mark L. Johnson. 1999. “Attention Metaphors: How Metaphors guide the Cognitive Psychology of Attention.” *Cognitive Science* 23 (1): 83–116. https://doi.org/10.1207/s15516709cog2301_4.

Fingerhut, Joerg. 2020. “Habits and the Enculturated Mind: Pervasive Artifacts, Predictive Processing, and Expansive Habits.” In *Habits*, edited by Fausto Caruana and Italo Testa, 352–75. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781108682312.018.

———. 2021. “Enacting Media. An Embodied Account of Enculturation between Neuromediality and New Cognitive Media Theory.” *Frontiers in Psychology* 12: 1151. https://doi.org/10.3389/fpsyg.2021.635993.

Fintz, Matan, Margarita Osadchy, and Uri Hertz. 2022. “Using Deep Learning to Predict Human Decisions and Using Cognitive Models to Explain Deep Learning Models.” *Scientific Reports* 12 (1): 4736. https://doi.org/10.1038/s41598-022-08863-0.

Fiser, József, Pietro Berkes, Gergő Orbán, and Máté Lengyel. 2010. “Statistically Optimal Perception and Learning: From Behavior to Neural Representations.” *Trends in Cognitive Sciences* 14 (3): 119–30. https://doi.org/10.1016/j.tics.2010.01.003.

Fletcher, Paul C., and Chris D. Frith. 2009. “Perceiving is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia.” *Nature Reviews Neuroscience* 10 (1): 48–58. https://doi.org/10.1038/nrn2536.

Fountas, Zafeirios, Anastasia Sylaidi, Kyriacos Nikiforou, Anil K. Seth, Murray Shanahan, and Warrick Roseboom. 2022. “A Predictive Processing Model of Episodic Memory and Time Perception.” *Neural Computation* 34 (7): 1501–44. https://doi.org/10.1162/neco_a_01514.

French, Steven, and James Ladyman. 1999. “Reinflating the Semantic Approach.” *International Studies in the Philosophy of Science* 13 (2): 103–21. https://doi.org/10.1080/02698599908573612.

Frigg, Roman. 2010. “Models and Fiction.” *Synthese* 172 (2): 251–68. https://doi.org/10.1007/s11229-009-9505-0.

———. 2022. *Models and Theories: A Philosophical Inquiry*. London ; New York, NY: Routledge/Taylor & Francis Group.

Friston, Karl J. 2005. “A Theory of Cortical Responses.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 360 (1456): 815–36. https://doi.org/10.1098/rstb.2005.1622.

———. 2009. “The Free-Energy Principle: A Rough Guide to the Brain?” *Trends in Cognitive Sciences* 13 (7): 293–301. https://doi.org/10.1016/j.tics.2009.04.005.

———. 2010. “The Free-Energy Principle: A Unified Brain Theory?” *Nature Reviews Neuroscience* 11 (2): 127–38. https://doi.org/10.1038/nrn2787.

———. 2012a. “A Free Energy Principle for Biological Systems.” *Entropy* 14 (11): 2100–2121. https://doi.org/10.3390/e14112100.

———. 2012b. “Policies and Priors.” In *Computational Neuroscience of Drug Addiction*, edited by Boris Gutkin, and Serge H. Ahmed, 237–83. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-0751-5_9.

———. 2012c. “The History of the Future of the Bayesian Brain.” *NeuroImage* 62: 1230–33. https://doi.org/10.1016/j.neuroimage.2011.10.004.

———. 2012d. “Prediction, Perception and Agency.” *International Journal of Psychophysiology* 83 (2): 248–52. https://doi.org/10.1016/j.ijpsycho.2011.11.014.

———. 2013. “Life as we Know It.” *Journal of the Royal Society Interface* 10 (86). https://doi.org/10.1098/rsif.2013.0475.

———. 2017. “The Mathematics of Mind-Time.” May 18, 2017. https://aeon.co/essays/consciousness-is-not-a-thing-but-a-process-of-inference.

———. 2019. “Waves of Prediction.” *PLOS Biology* 17 (10): e3000426. https://doi.org/10.1371/journal.pbio.3000426.

Friston, Karl J., Rick Adams, and Read Montague. 2012. “What is Value—Accumulated Reward or Evidence?” *Frontiers in Neurorobotics* 6. https://doi.org/10.3389/fnbot.2012.00011.

Friston, Karl J., Lancelot Da Costa, Noor Sajid, Conor Heins, Kai Ueltzhöffer, Grigorios A. Pavliotis, and Thomas Parr. 2023. “The Free Energy Principle Made Simpler but Not Too Simple.” *Physics Reports* 1024 (June): 1–29. https://doi.org/10.1016/j.physrep.2023.07.001.

Friston, Karl J., Jean Daunizeau, James Kilner, and Stefan J. Kiebel. 2010. “Action and Behavior: A Free-Energy Formulation.” *Biological Cybernetics* 102 (3): 227–60. https://doi.org/10.1007/s00422-010-0364-z.

Friston, Karl J., Klaas Enno Stephan, Read Montague, and Raymond J Dolan. 2014. “Computational Psychiatry: The Brain as a Phantastic Organ.” *The Lancet Psychiatry* 1 (2): 148–58. https://doi.org/10.1016/S2215-0366(14)70275-5.

Friston, Karl J., L. Harrison, and W. Penny. 2003. “Dynamic Causal Modelling.” *NeuroImage* 19 (4): 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7.

Froese, Tom, and John Stewart. 2010. “Life After Ashby: Ultrastability and the Autopoietic Foundations of Biological Autonomy.” *Cybernetics and Human Knowing* 17 (4): 7–50.

Fuchs, Thomas. 2018. *Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind*. Oxford, UK: Oxford University Press.

Gallagher, Shaun. 2017. *Enactivist Interventions*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198794325.001.0001.

Gallagher, Shaun, and Micah Allen. 2018. “Active Inference, Enactivism and the Hermeneutics of Social Cognition.” *Synthese* 195 (6): 2627–48. https://doi.org/10.1007/s11229-016-1269-8.

Gallagher, Shaun, Daniel Hutto, and Inês Hipólito. 2022. “Predictive Processing and Some Disillusions about Illusions.” *Review of Philosophy and Psychology* 13 (4): 999–1017. https://doi.org/10.1007/s13164-021-00588-9.

Gallagher, Shaun, and Dan Zahavi. 2012. *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*. London: Routledge.

———. 2014. “Primal Impression and Enactive Perception.” In *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*, edited by Valtteri Arstila and Dan Edward Lloyd, 83–100. Cambridge, Mass.: The MIT press.

Gärdenfors, Peter. 2004. “Conceptual Spaces as a Framework for Knowledge Representation.” *Mind and Matter* 2 (2): 9–27.

Gardner, Howard. 1985. *The Mind’s New Science: A History of the Cognitive Revolution*. New York, NY: Basic Books.

Gelfert, Axel. 2011. “Model-Based Representation in Scientific Practice: New Perspectives.” *Studies in History and Philosophy of Science Part A* 42 (2): 251–52. https://doi.org/10.1016/j.shpsa.2010.11.032.

———. 2016. *How to do Science with Models. A Philosophical Primer*. Singapore: Springer.

Gibson, James J. 2015. *The Ecological Approach to Visual Perception*. Hove, East Sussex: Psychology Press.

Giere, Ronald. 1999. “Using Models to Represent Reality.” In *Model-Based Reasoning in Scientific Discovery*, edited by Lorenzo Magnani, Nancy J. Nersessian, and Paul Thagard, 41–57. New York, NY: Kluwer Academic Publishers.

———. 2010. “An Agent-Based Conception of Models and Scientific Representation.” *Synthese* 172 (2): 269–81. https://doi.org/10.1007/s11229-009-9506-z.

Gigerenzer, Gerd, and Daniel G. Goldstein. 1996. “Reasoning the Fast and Frugal Way: Models of Bounded Rationality.” *Psychological Review* 103 (4): 650–69. https://doi.org/10.1037/0033-295X.103.4.650.

Gilaie-Dotan, Sharon, Assaf Harel, Shlomo Bentin, Ryota Kanai, and Geraint Rees. 2012. “Neuroanatomical Correlates of Visual Car Expertise.” *NeuroImage 62* (1). https://doi.org/10.1016/j.neuroimage.2012.05.017.

Gilaie-Dotan, Sharon, Ryota Kanai, Bahador Bahrami, Geraint Rees, and Ayse P. Saygin. 2013. “Neuroanatomical Correlates of Biological Motion Detection.” *Neuropsychologia* 51 (3): 457–63. https://doi.org/10.1016/j.neuropsychologia.2012.11.027.

Gładziejewski, Paweł. 2016. “Predictive Coding and Representationalism.” *Synthese* 193 (2): 559–82. https://doi.org/10.1007/s11229-015-0762-9.

———. 2019. “Mechanistic Unity of the Predictive Mind.” *Theory and Psychology* 29 (5): 657–75. https://doi.org/10.1177/0959354319866258.

Godfrey-Smith, Peter. 2006. “The Strategy of Model-Based Science.” *Biology and Philosophy* 21 (5): 725–40. https://doi.org/10.1007/s10539-006-9054-6.

Goodey, Chris F. 2011. *A History of Intelligence and “Intellectual Disability”: The Shaping of Psychology in Early Modern Europe*. New York, NY: Routledge.

Graham, George. 2023. “Behaviorism.” In T*he Stanford Encyclopedia of Philosophy*, edited by Edward Zalta and Uri Nodelman, Spring. https://plato.stanford.edu/archives/spr2023/entries/behaviorism/.

Grazzini, Jakob, and Matteo Richiardi. 2015. “Estimation of Ergodic Agent-Based Models by Simulated Minimum Distance.” *Journal of Economic Dynamics and Control* 51 (February): 148–65. https://doi.org/10.1016/j.jedc.2014.10.006.

Green, Sara. 2015. “Revisiting Generality in Biology: Systems Biology and the Quest for Design Principles.” *Biology and Philosophy* 30 (5): 629–52. https://doi.org/10.1007/s10539-015-9496-9.

Grüne-Yanoff, Till. 2013. “Appraising Models Nonrepresentationally.” *Philosophy of Science* 80 (5): 850–61. https://doi.org/10.1086/673893.

Guest, Olivia, and Andrea E. Martin. 2021. “How Computational Modeling can Force Theory Building in Psychological Science.” *Perspectives on Psychological Science* 16 (4): 789–802. https://doi.org/10.1177/1745691620970585.

Hacking, Ian. 1983. *Representing and Intervening. Introductory Topics in the Philosophy of Natural Science*. New York, NY: Cambridge University Press.

Harkness, Dominic L, and Ashima Keshava. 2017. “Moving from the What to the How and Where – Bayesian Models and Predictive Processing.” In *Philosophy and Predictive Processing*, edited by Wanja Wiese and Thomas Metzinger, 1–10. Frankfurt: MIND Group. https://doi.org/10.15502/9783958573178.

Harris, Albert K. 2018. “The Need for a Concept of Shape Homeostasis.” *Computational, Theoretical, and Experimental Approaches to Morphogenesis* 173 (November): 65–72. https://doi.org/10.1016/j.biosystems.2018.09.012.

Haugeland, John. 1978. “The Nature and Plausibility of Cognitivism.” *The Behavioral and Brain Sciences* 2: 215–60. https://doi.org/10.1017/S0140525X00074148.

Heath, B., R. Hill, and F. Ciarallo. n.d. “A Survey of Agent Based Modelling Practices (January 1998 to July 2008).” *The Journal of Artificial Societies and Social Simulation* 12 (4). http://jasss.soc.surrey .ac.uk/12/4/9.html.

Helmholtz, Hermann von. 1867. *Handbuch Der Physiologischen Optik*. Leipzig: Leopold Voss.

———. 1925. *Treatise on Physiological Optics*. *Vol. III. The Perception of Vision*. Edited by James P. C. Southall. Wisconsin: The Optical Society of America.

———. 1977. *Epistemological Writings*. *The Paul Hertz/ Moritz Schlick Edition of 1921*. Dordrecht, Boston: D. Reidel Publishing Company. https://doi.org/10.2307/2105836.

Herfeld, Catherine, and Chiara Lisciandra. 2019. “Knowledge Transfer and Its Contexts.” *Studies in History and Philosophy of Science* Part A 77 (October): 1–10. https://doi.org/10.1016/j.shpsa.2019.06.002.

Heylighen, Francis, and Cliff Joslyn. 2003. “Cybernetics and Second-Order Cybernetics.” *Encyclopedia of Physical Science and Technology* 4: 155–69. https://doi.org/10.1016/B0-12-227410-5/00161-7.

Hilgetag, Claus C., and Alexandros Goulas. 2020. “‘Hierarchy’ in the Organization of Brain Networks.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1796): 20190319. https://doi.org/10.1098/rstb.2019.0319.

Hipólito, Inês. 2018. “Perception is Not Always and Everywhere Inferential.” *Australasian Philosophical Review* 2 (2): 184–88. https://doi.org/10.1080/24740500.2018.1552093.

Hipólito, Inês, and Thomas van Es. 2022. “Enactive-Dynamic Social Cognition and Active Inference.” *Frontiers in Psychology* 13 (April): 855074. https://doi.org/10.3389/fpsyg.2022.855074.

Hodson, Rowan, Marishka Mehta, and Ryan Smith. 2024. “The Empirical Status of Predictive Coding and Active Inference.” *Neuroscience & Biobehavioral Reviews* 157 (February):105473. https://doi.org/10.1016/j.neubiorev.2023.105473.

Hoehl, Stefanie, Kahl Hellmer, Maria Johansson, and Gustaf Gredebäck. 2017. “Itsy Bitsy Spider…: Infants react with Increased Arousal to Spiders and Snakes.” *Frontiers in Psychology* 8 (October): 1710. https://doi.org/10.3389/FPSYG.2017.01710.

Hohwy, Jakob. 2012. “Attention and Conscious Perception in the Hypothesis Testing Brain.” *Frontiers in Psychology* 3. https://doi.org/10.3389/fpsyg.2012.00096.

———. 2013. *The Predictive Mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001.

———. 2018. “The Predictive Processing Hypothesis.” In *The Oxford Handbook of 4E Cognition*, by Jakob Hohwy, edited by Albert Newen, Leon De Bruin, and Shaun Gallagher, 128–46. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.7.

———. 2020. “New Directions in Predictive Processing.” *Mind and Language* 35 (2): 209–23. https://doi.org/10.1111/mila.12281.

———. 2021. “Self-Supervision, Normativity and the Free Energy Principle.” *Synthese* 199 (1–2): 29–53. https://doi.org/10.1007/s11229-020-02622-2.

Humphreys, Paul. 2002. “Computational Models.” *Philosophy of Science* 69 (S3): S1–11. https://doi.org/10.1086/341763.

———. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford, New York, NY: Oxford University Press.

———. 2019. “Knowledge Transfer across Scientific Disciplines.” *Studies in History and Philosophy of Science Part A* 77 (October): 112–19. https://doi.org/10.1016/j.shpsa.2017.11.001.

Husserl, Edmund. 2001. *Logical Investigations*, vol. II. Edited by Dermot Moran. London, New York, NY: Routledge.

Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, Mass.: The MIT Press.

Hutto, Daniel D. 2018. “Getting into Predictive Processing’s Great Guessing Game: Bootstrap Heaven or Hell?” *Synthese* 195 (6): 2445–58. https://doi.org/10.1007/s11229-017-1385-0.

Hutto, Daniel D., Shaun Gallagher, Jesús Ilundáin-Agurruza, and Inês Hipólito. 2020. “Culture in Mind – An Enactivist Account.” In *Culture, Mind, and Brain*, 163–87. Cambridge University Press. https://doi.org/10.1017/9781108695374.009.

Hutto, Daniel D, and Erik Myin. 2014. “Neural Representations not Needed - No More Pleas, Please.” *Phenomenology and the Cognitive Sciences* 13 (2): 241–56. https://doi.org/10.1007/s11097-013-9331-1.

Hutto, Daniel D., and Erik Myin. 2017. *Evolving Enactivism: Basic Minds meet Content*. Cambridge, Mass., London: The MIT Press.

Issad, Tarik, and Christophe Malaterre. 2015. “Are Dynamic Mechanistic Explanations Still Mechanistic?” In *History, Philosophy and Theory of the Life Sciences*, vol. 11. https://doi.org/10.1007/978-94-017-9822-8_12.

Itti, Laurent, and Pierre Baldi. 2009. “Bayesian Surprise attracts Human Attention.” *Vision Research* 49 (10): 1295–1306. https://doi.org/10.1016/j.visres.2008.09.007.

Jacquette, Dale. 2009. *The Philosophy of Mind: The Metaphysics of Consciousness*. 2nd ed. London: Bloomsbury Publishing Plc.

Kaiser, Marie I, Maria Kronfeldner, and Robert Meunier. 2014. “Interdisciplinarity in Philosophy of Science.” *Journal for General Philosophy of Science* 45 (1): 59–70. https://doi.org/10.1007/s10838-014-9269-1.

Kaplan, David Michael, and Carl F. Craver. 2011. “The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective.” *Philosophy of Science* 78 (4): 601–27. https://doi.org/10.1086/661755.

Keller, Georg B., and Thomas D. Mrsic-Flogel. 2018. “Predictive Processing: A Canonical Cortical Computation.” *Neuron* 100 (2): 424–35. https://doi.org/10.1016/j.neuron.2018.10.003.

Kiang, Melody Y. 2003. “Neural Networks.” In *Encyclopedia of Information Systems*, 3: 303–15. Academic Press Inc. https://doi.org/10.1016/B0-12-227240-4/00121-0.

Kim, Chang Sub. 2023. “Free Energy and Inference in Living Systems.” *Interface Focus* 13 (3): 20220041. https://doi.org/10.1098/rsfs.2022.0041.

Kirchhoff, Michael. 2018. “Predictive Brains and Embodied, Enactive Cognition: An Introduction to the Special Issue.” *Synthese* 195 (6): 2355–66. https://doi.org/10.1007/s11229-017-1534-5.

Kirchhoff, Michael D., and Ian Robertson. 2018. “Enactivism and Predictive Processing: A Non-Representational View.” *Philosophical Explorations* 21 (2): 264–81. https://doi.org/10.1080/13869795.2018.1477983.

Kirsch, David. 2010. “Thinking with External Representations.” *AI and Society* 25 (4): 441–54. https://doi.org/10.1007/s00146-010-0272-8.

Kitcher, Patricia. 1988. “Marr’s Computational Theory of Vision.” *Philosophy of Science* 55 (1): 1–24. https://doi.org/10.1086/289413.

Kiverstein, Julian. 2020. “Free Energy and the Self: An Ecological–Enactive Interpretation.” *Topoi* 39 (3): 559–74. https://doi.org/10.1007/s11245-018-9561-5.

Kleckner, Ian R., Jiahe Zhang, Alexandra Touroutoglou, Lorena Chanes, Chenjie Xia, W. Kyle Simmons, Karen S. Quigley, Bradford C. Dickerson, and Lisa Feldman Barrett. 2017. “Evidence for a Large-Scale Brain System Supporting Allostasis and Interoception in Humans.” *Nature Human Behaviour* 1 (5): 0069. https://doi.org/10.1038/s41562-017-0069.

Knill, David C., Daniel Kersten, and Alan Yuille. 1996. “Introduction.” In *Perception as Bayesian Inference*, edited by David C. Knill and Whitman Richards, 1–22. Cambridge, Mass., New York, NY: Cambridge University Press. https://doi.org/10.1017/cbo9780511984037.

Knill, David C, and Alexandre Pouget. 2004. “The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation.” *Trends in Neurosciences* 27 (12): 712–19. https://doi.org/10.1016/j.tins.2004.10.007.

Knuuttila, Tarja. 2005a. “Models as Epistemic Artefacts: Toward a Non-Representationalist Account of Scientific Representation.” PhD Thesis, Helsinki: University of Helsinki.

———. 2005b. “Models, Representation, and Mediation.” *Philosophy of Science* 72 (5): 1260–71. https://doi.org/10.1086/508124.

———. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42 (2): 262–71. https://doi.org/10.1016/j.shpsa.2010.11.034.https://doi.org/10.1007/S13194-021-00374-5/FIGURES/2.

———. 2017. “Imagination Extended and Embedded: Artifactual Versus Fictional Accounts of Models.” *Synthese* 198 (S21): 5077–97. https://doi.org/10.1007/s11229-017-1545-2.

———. 2021. “Epistemic Artifacts and the Modal Dimension of Modeling.” *European Journal for Philosophy of Science* 11 (3): 1–18.

Knuuttila, Tarja, and Mieke Boon. 2011. “How do Models give Us Knowledge? The Case of Carnot’s Ideal Heat Engine.” *European Journal for Philosophy of Science* 1 (2): 309–34.

Knuuttila, Tarja, and Mary S. Morgan. 2019. “Deidealization: No Easy Reversals.” *Philosophy of Science* 86 (4): 641–61. https://doi.org/10.1086/704975.

Knuuttila, Tarja, and Andrea Loettgers. 2014. “Magnets, Spins, and Neurons: The Dissemination of Model Templates across Disciplines.” *The Monist* 97 (3): 280–300. https://doi.org/10.5840/monist201497319.

———. 2016. “Model Templates within and between Disciplines: From Magnets to Gases – and Socio-Economic Systems.” *European Journal for Philosophy of Science* 6 (3): 377–400. https://doi.org/10.1007/s13194-016-0145-1.

———. 2017a. “Modelling as Indirect Representation? The Lotka–Volterra Model Revisited.” *The British Journal for the Philosophy of Science* 68 (4): 1007–36. https://doi.org/10.1093/bjps/axv055.

———. 2017b. “What are Definitions of Life Good for? Transdisciplinary and Other Definitions in Astrobiology.” *Biology and Philosophy* 32 (6): 1185–1203. https://doi.org/10.1007/s10539-017-9600-4.

———. 2023. “Model Templates: Transdisciplinary Application and Entanglement.” *Synthese* 201 (6): 200. https://doi.org/10.1007/s11229-023-04178-3.

Koskinen, Rami. 2017. “Synthetic Biology and the Search for Alternative Genetic Systems: Taking How-Possibly Models Seriously.” *European Journal for Philosophy of Science* 7 (3). https://doi.org/10.1007/s13194-017-0176-2.

Krippendorff, Klaus. 2009. “Ross Ashby’s Information Theory: A Bit of History, Some Solutions to Problems, and What We face Today.” *International Journal of General Systems* 38 (2): 189–212. https://doi.org/10.1080/03081070802621846.

Kruschke, John K. 2015. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed. Boston: Academic Press.

Kutas, Marta, and Kara D. Federmeier. 2011. “Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP).” *Annual Review of Psychology* 62 (1): 621–47. https://doi.org/10.1146/annurev.psych.093008.131123.

Laland, Kevin, Blake Matthews, and Marcus W. Feldman. 2016. “An Introduction to Niche Construction Theory.” *Evolutionary Ecology* 30 (2). https://doi.org/10.1007/s10682-016-9821-z.

Laland, Kevin N., and Kim Sterelny. 2006. “Seven Reasons (not) to neglect Niche Construction.” *Evolution* 60 (9): 1751. https://doi.org/10.1554/05-570.1.

Latour, Bruno. 1986. “Visualization and Cognition: Thinking with Eyes and Hands.” *Knowledge and Society* 6: 1–40.

Lecas, Jean Claude. 2006. “Behaviourism and the Mechanization of the Mind.” *Comptes Rendus - Biologies* 329 (5–6): 386–97. https://doi.org/10.1016/j.crvi.2006.03.009.

Lewandowsky, Stephan, Toby D. Pilditch, Jens K. Madsen, Naomi Oreskes, and James S. Risbey. 2019. “Influence and Seepage: An Evidence-Resistant Minority can affect Public Opinion and Scientific Belief Formation.” *Cognition* 188 (July): 124–39. https://doi.org/10.1016/j.cognition.2019.01.011.

Lochmann, Timm, and Sophie Deneve. 2011. “Neural Processing as Causal Inference.” *Current Opinion in Neurobiology* 21 (5): 774–81. https://doi.org/10.1016/j.conb.2011.05.018.

Lorentz, Hendrik Antoon. 1927. *Theoretical Physics*. London: MacMillan.

Lowe, Robert, Gordana Dodig-Crnkovic, and Alexander Almer. 2018. “Predictive Regulation in Affective and Adaptive Behaviour: An Allostatic-Cybernetics Perspective.” In

Advanced Research on Biologically Inspired Cognitive Architectures, edited by Manuel Mazzara, Jordi Vallverdu, Max Talanov, Salvatore Distefano, and Robert Lowe, 149–76. Pennsylvania: Igi-global. https://doi.org/10.4018/978-1-5225-1947-8.ch008.

Luhmann, Niklas. 1997. *Die Gesellschaft Der Gesellschaft*. Vol. 1. Frankfurt: Suhrkamp.

Luisi, Pier Luigi. 2016. *The Emergence of Life*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9781316135990.

Lynch, Scott M. 2015. "Bayesian Theory, History, Applications, and Contemporary Directions." In *International Encyclopedia of the Social and Behavioral Sciences*, 378–82. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.43013-8.

Machery, Edouard. 2005. "Concepts are not a Natural Kind." *Philosophy of Science* 72 (3): 444–67. https://doi.org/10.1086/498473.

Madsen, Jens Koed, Richard Bailey, Ernesto Carrella, and Philipp Koralus. 2019. "Analytic Versus Computational Cognitive Models: Agent-Based Modeling as a Tool in Cognitive Sciences." *Current Directions in Psychological Science* 28 (3). https://doi.org/10.1177/0963721419834547.

Markman, Arthur B., and Eric Dietrich. 2000. "Extending the Classical View of Representation." *Trends in Cognitive Sciences* 4 (12): 470–75. https://doi.org/10.1016/S1364-6613(00)01559-X.

Marr, David. 2010. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, Mass.: The MIT Press.

Martínez, Sergio. 2010. "Science as Evolution of Technologies of Cognition." In *The Hereditary Hourglass. Genetics and Epigenetics*, 1968-2000, edited by A. Barahona and H. Rheinberger, 97–109. Berlin: Max Planck Institute for the History of Science.

Massimi, Michela. 2022. *Perspectival Realism*. New York, NY: Oxford University Press.

McClelland, James L. 2009. "The Place of Modeling in Cognitive Science." *Topics in Cognitive Science* 1 (1): 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x.

McCollough, Celeste. 1965. "Color Adaptation of Edge-Detectors in the Human Visual System." *Science* 149 (3688): 1115–16. https://doi.org/10.1126/science.149.3688.1115.

McLane, Adam J., Christina Semeniuk, Gregory J. McDermid, and Danielle J. Marceau. 2011. "The Role of Agent-Based Models in Wildlife Ecology and Management." *Ecological Modelling* 222 (8): 1544–56. https://doi.org/10.1016/j.ecolmodel.2011.01.020.

Millikan, Ruth Garrett. 2009. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, Mass.: The MIT Press.

Miłkowski, Marcin, and Mateusz Hohol. 2021. "Explanations in Cognitive Science: Unification Versus Pluralism." *Synthese* 199 (S1): 1–17. https://doi.org/10.1007/s11229-020-02777-y.

Miller, George A. 2003. "The Cognitive Revolution: A Historical Perspective." *Trends in Cognitive Sciences* 7 (3): 141–44. https://doi.org/10.1016/S1364-6613(03)00029-9.

Moberg, Christina. 2020. "Schrödinger's What is Life? —The 75th Anniversary of a Book that Inspired Biology." *Angewandte Chemie* 132 (7): 2570–73. https://doi.org/10.1002/ange.201911112.

Moeller, Korbinian, Ursula Fischer, Tanja Link, Mirjam Wasner, Stefan Huber, Ulrike Cress, and Hans-Christoph Nuerk. 2012. "Learning and Development of Embodied Numerosity." *Cognitive Processing* 13 (S1): 271–74. https://doi.org/10.1007/s10339-012-0457-9.

Moran, R. J., P. Campo, M. Symmonds, K. E. Stephan, R. J. Dolan, and K. J. Friston. 2013. "Free Energy, Precision and Learning: The Role of Cholinergic Neuromodulation." *Journal of Neuroscience* 33 (19): 8227–36. https://doi.org/10.1523/JNEUROSCI.4255-12.2013.

Morgan, Alex, and Gualtiero Piccinini. 2018. "Towards a Cognitive Neuroscience of Intentionality." *Minds and Machines* 28 (1): 119–39. https://doi.org/10.1007/s11023-017-9437-2.

Morgan, Mary, and Margaret Morrison. 1999. *Models As Mediators*. Cambridge, UK: Cambridge University Press.

Mumford, D. 1991. "On the Computational Architecture of the Neocortex: I. The Role of the Thalamo-Cortical Loop." *Biological Cybernetics* 65 (2): 135–45. https://doi.org/10.1007/BF00202389.

Myin, Erik, and Thomas van Es. 2020. "Predictive Processing and Representation: How Less can be More." In The *Philosophy and Science of Predictive Processing,* edited by Dina Mendonça, Manuel Curado, and Steven S. Gouveia, 7–24. New York, NY: Bloomsbury Academic.

Nath, Rajakishore. 2020. "Alan Turing's Concept of Mind." *Journal of Indian Council of Philosophical Research* 37 (1): 31–50. https://doi.org/10.1007/s40961-019-00188-0.

Neisser, Ulric. 2014. *Cognitive Psychology*. New York, NY: Psychology Press. https://doi.org/10.4324/9781315736174.

Nersessian, Nancy J. 2009. "How do Engineering Scientists think? Model-Based Simulation in Biomedical Engineering Research Laboratories." *Topics in Cognitive Science* 1 (4): 730–57. https://doi.org/10.1111/j.1756-8765.2009.01032.x.

Newen, Albert, Shaun Gallagher, and Leon De Bruin. 2018. "4E Cognition: Historical Roots, Key Concepts, and Central Issues." In *The Oxford Handbook of 4E Cognition*, edited by Albert Newen, Leon De Bruin, and Shaun Gallagher, 2–16. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.1.

Niazi, M., and A. Hussain. 2009. "Agent-Based Tools for Modeling and Simulation of Self-Organization in Peer-to-Peer, Ad Hoc, and Other Complex Networks." *IEEE Communications Magazine* 47 (3): 166–73. https://doi.org/10.1109/MCOM.2009.4804403.

Noë, Alva. 2004. *Action in Perception*. *Representation and Mind*. Cambridge, Mass.: The MIT Press.

Núñez, Rafael, Michael Allen, Richard Gao, Carson Miller Rigoli, Josephine Relaford-Doyle, and Arturs Semenuks. 2019. "What Happened to Cognitive Science?" *Nature Human Behaviour* 3 (8): 782–91. https://doi.org/10.1038/s41562-019-0626-2.

Odling-Smee, F. J. 1988. "Niche-Constructing Phenotypes." In *The Role of Behavior in Evolution*, edited by H. C. Plotkin, 73–132. Cambridge, Mass.: The MIT Press.

O'Regan, J. Kevin, and Alva Noë. 2001. "A Sensorimotor Account of Vision and Visual Consciousness." *Behavioral and Brain Sciences* 24 (5): 939–73. https://doi.org/10.1017/S0140525X01000115.

Orlandi, Nico. 2016. "Bayesian Perception Is Ecological Perception:" *Philosophical Topics* 44 (2): 327–51. https://doi.org/10.5840/philtopics201644226.

———. 2018. "Predictive Perceptual Systems." *Synthese* 195 (6): 2367–86. https://doi.org/10.1007/s11229-017-1373-4.

Parr, Thomas, David A. Benrimoh, Peter Vincent, and Karl J. Friston. 2018. "Precision and False Perceptual Inference." *Frontiers in Integrative Neuroscience* 12 (September): 39. https://doi.org/10.3389/fnint.2018.00039.

Parr, Thomas, Andrew W Corcoran, Karl J. Friston, and Jakob Hohwy. 2019. "Perceptual Awareness and Active Inference." *Neuroscience of Consciousness* 1. https://doi.org/10.1093/nc/niz012.

Parr, Thomas, Giovanni Pezzulo, and K. J. Friston. 2022. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, Mass.: The MIT Press.

Partelow, Stefan. 2023. "What is a Framework? Understanding Their Purpose, Value, Development and Use." *Journal of Environmental Studies and Sciences* 13 (3): 510–19. https://doi.org/10.1007/s13412-023-00833-w.

Pavlov, Ivan P. 2010. "Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex." *Annals of Neurosciences* 17 (3). https://doi.org/10.5214/ans.0972-7531.1017309.

Pernu, Tuomas K. 2017. "The Five Marks of the Mental." *Frontiers in Psychology* 8 (July): 1084. https://doi.org/10.3389/fpsyg.2017.01084.

Peschard, Isabelle. 2011. "Making Sense of Modeling: Beyond Representation." *European Journal for Philosophy of Science* 1(3): 335–52. https://doi.org/10.1007/s13194-011-0032-8.

Pezzulo, Giovanni, Francesco Donnarumma, Pierpaolo Iodice, Domenico Maisto, and Ivilin Stoianov. 2017. "Model-Based Approaches to Active Perception and Control." *Entropy* 19 (6): 266. https://doi.org/10.3390/e19060266.

Pezzulo, Giovanni, Francesco Rigoli, and Karl J. Friston. 2015. "Active Inference, homeostatic Regulation and Adaptive Behavioural Control." *Progress in Neurobiology* 134: 17–35. https://doi.org/10.1016/j.pneurobio.2015.09.001.

Piccinini, Gualtiero, and Andrea Scarantino. 2011. "Information Processing, Computation, and Cognition." *Journal of Biological Physics* 37 (1): 1–38. https://doi.org/10.1007/s10867-010-9195-3.

Pickering, Andrew. 2009. "Psychiatry, Synthetic Brains and Cybernetics in the Work of W. Ross Ashby." *International Journal of General Systems* 38 (2): 213–30. https://doi.org/10.1080/03081070802712025.

Piekarski, Michał. 2021. "Understanding Predictive Processing. A Review." *Avant* 12 (1). https://doi.org/10.26913/avant.2021.01.04.

Pinker, Steven. 2005. "So How does the Mind Work?" *Mind and Language* 20 (1): 1–24. https://doi.org/10.1111/j.0268-1064.2005.00274.x.

Poldrack, Russell A. 2020. "The Physics of Representation." *Synthese* (July): 1–19. https://doi.org/10.1007/s11229-020-02793-y.

Pouyan, Nasser. 2014. “Helmholtz Who Revolutionized Ophthalmology.” *Research in Ophthalmology* 3 (1): 1–4. https://doi.org/10.5923/j.ophthal.20140301.01.

Putnam, Hilary. 1960. “Minds and Machines.” In *Dimensions of Mind: A Symposium*, edited by Sidney Hook, 138–64. New York, NY: New York University Press.

———. 1980. “17. The Nature of Mental States.” In *The Language and Thought Series*, edited by Ned Block. Cambridge, Mass., London: Harvard University Press. https://doi.org/10.4159/harvard.9780674594623.c26.

Ramsey, William. 2024. “Eliminative Materialism.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta and Uri Nodelman, Winter. https://plato.stanford.edu/archives/win2024/entries/materialism-eliminative/.

Ramstead, Maxwell J., Michael D. Kirchhoff, and Karl J. Friston. 2020. “A Tale of Two Densities: Active Inference is Enactive Inference.” *Adaptive Behavior* 28 (4): 225–39. https://doi.org/10.1177/1059712319862774.

Rao, Rajesh P.N. 2005. “Bayesian Inference and Attentional Modulation in the Visual Cortex.” *NeuroReport* 16 (16): 1843–48. https://doi.org/10.1097/01.wnr.0000183900.92901.fc.

Rao, Rajesh P.N., and Dana H. Ballard. 1999. “Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects.” *Nature Neuroscience* 2 (1): 79–87. https://doi.org/10.1038/4580.

Rauss, Karsten, and Gilles Pourtois. 2013. “What is Bottom-Up and What is Top-Down in Predictive Coding?” *Frontiers in Psychology* 4. https://doi.org/10.3389/fpsyg.2013.00276.

Rey, Georges. 1997. *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Oxford, UK: Blackwell.

Rheinberger, Hans-Jörg. 1997. *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford, CA: Stanford University Press.

Rescorla, Michael. 2020. “The Computational Theory of Mind.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta, Fall 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/computational-mind/.

Rice, Collin. 2018. “Idealized Models, Holistic Distortions, and Universality.” *Synthese* 195 (6): 2795–2819. https://doi.org/10.1007/s11229-017-1357-4.

Rizzolatti, G., L. Fadiga, M. Matelli, V. Bettinardi, E. Paulesu, D. Perani, and F. Fazio. 1996. “Localization of Grasp Representations in Humans by PET: 1. Observation Versus Execution.” *Experimental Brain Research* 111 (2). https://doi.org/10.1007/BF00227301.

Rosenblueth, Arturo, and Norbert Wiener. 1945. “The Role of Models in Science.” *Philosophy of Science* 12 (4): 316–21.

Rusanen, A-M., and O. Lappi. 2016. “The Limits of Mechanistic Explanation in Neurocognitive Sciences.” In *Proceedings of EuroCogSci07: The European Cognitive Science Conference*, edited by S. Vosniadou, D. Kayser, and A. Protopapas, 284–89. Taylor and Francis.

Russell, Stuart J., and Peter Norvig. 2022. A*rtificial Intelligence: A Modern Approach*. 4th ed. Harlow: Pearson.

Sajid, Noor, Philip J Ball, and Karl J. Friston. 2021. “Active Inference: Demystified and Compared.” *Neural Computation* 33 (3): 674–712. https://doi.org/10.1162/neco_a_01357.

Salis, Fiora. 2021. “The New Fiction View of Models.” *The British Journal for the Philosophy of Science* 72 (3): 717–42. https://doi.org/10.1093/bjps/axz015.

Sanches De Oliveira, Guilherme. 2021. “Representationalism Is a Dead End.” *Synthese* 198 (1): 209–35. https://doi.org/10.1007/s11229-018-01995-9.

Sanches De Oliveira, Guilherme, Thomas van Es, and Inês Hipólito. 2023. “Scientific Practice as Ecological-Enactive Co-Construction.” *Synthese* 202 (1): 4. https://doi.org/10.1007/s11229-023-04215-1.

Sánchez-Dorado, Julia. 2024. In *The Routledge Handbook of Philosophy of Scientific Modeling*, edited by Tarja Knuuttila, Natalia Carrillo, and Rami Koskinen, 59–73. London, New York, NY: Routledge.

Sato, Daisuke, Yudai Yamazaki, Koya Yamashiro, Hideaki Onishi, Yasuhiro Baba, Koyuki Ikarashi, and Atsuo Maruyama. 2020. “Elite Competitive Swimmers Exhibit Higher Motor Cortical Inhibition and Superior Sensorimotor Skills in a Water Environment.” *Behavioural Brain Research* 395. https://doi.org/10.1016/j.bbr.2020.112835.

Saygin, A. P. 2007. “Superior Temporal and Premotor Brain Areas Necessary for Biological Motion Perception.” *Brain* 130 (9): 2452–61. https://doi.org/10.1093/brain/awm162.

Saygin, Ayse Pinar, Thierry Chaminade, Hiroshi Ishiguro, Jon Driver, and Chris Frith. 2012. “The Thing that should not ne: Predictive Coding and the Uncanny Valley in Perceiving Human and Humanoid Robot Actions.” *Social Cognitive and Affective Neuroscience* 7 (4): 413–22. https://doi.org/10.1093/scan/nsr025.

Schelling, Thomas C. 1971. “Dynamic Models of Segregation.” *The Journal of Mathematical Sociology* 1 (2): 143–86. https://doi.org/10.1080/0022250X.1971.9989794.

———. 1978. *Micromotives and Macrobehavior. Fels Lectures on Public Policy Analysis*. New York, NY: Norton.

Schoot, Rens van de, Sonja D. Winter, Oisín Ryan, Mariëlle Zondervan-Zwijnenburg, and Sarah Depaoli. 2017. “A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years.” *Psychological Methods* 22 (2): 217–39. https://doi.org/10.1037/met0000100.

Schrödinger, Erwin. 1992. *What Is Life? The Physical Aspect of the Living Cell ; with, Mind and Matter ; and Autobiographical Sketches*. Cambridge, UK, New York, NY: Cambridge University Press.

Schroeder, Charles E., Donald A. Wilson, Thomas Radman, Helen Scharfman, and Peter Lakatos. 2010. “Dynamics of Active Sensing and Perceptual Selection.” *Current Opinion in Neurobiology* 20 (2): 172–76. https://doi.org/10.1016/j.conb.2010.02.010.

Schwartenbeck, Philipp, Thomas FitzGerald, Raymond J. Dolan, and Karl J. Friston. 2013. “Exploration, Novelty, Surprise, and Free Energy Minimization.” *Frontiers in Psychology* 4 (October): 710. https://doi.org/10.3389/fpsyg.2013.00710.

Searle, John R. 1980. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3 (3): 417–24. https://doi.org/10.1017/S0140525X00005756.

Seth, Anil K. 2013. “Interoceptive Inference, Emotion, and the Embodied Self.” *Trends in Cognitive Sciences* 17 (11): 565–73. https://doi.org/10.1016/j.tics.2013.09.007.

———. 2014. “A Predictive Processing Theory of Sensorimotor Contingencies: Explaining the Puzzle of Perceptual Presence and its Absence in Synesthesia.” *Cognitive Neuroscience* 5 (2): 97–118. https://doi.org/10.1080/17588928.2013.877880.

———. 2015. “The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies.” In *Open MIND*, edited by T. Metzinger and J. M. Windt. Frankfurt: MIND Group. https://doi.org/10.15502/9783958570108.

Seth, Anil K., and Karl J. Friston. 2016. “Active Interoceptive Inference and the Emotional Brain.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1708): 20160007. https://doi.org/10.1098/rstb.2016.0007.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shapiro, Larry. 2007. “The Embodied Cognition Research Programme.” *Philosophy Compass* 2 (2): 338–46. https://doi.org/10.1111/j.1747-9991.2007.00064.x.

Shipp, Stewart. 2016. “Neural Elements for Predictive Coding.” *Frontiers in Psychology* 7 (1792). https://doi.org/10.3389/fpsyg.2016.01792.

Shvarts, Anna, Rosa Alberto, Arthur Bakker, Michiel Doorman, and Paul Drijvers. 2021. “Embodied Instrumentation in Learning Mathematics as the Genesis of a Body-Artifact Functional System.” *Educational Studies in Mathematics* 107 (3): 447–69. https://doi.org/10.1007/s10649-021-10053-0.

Sims, Matt, and Giovanni Pezzulo. 2021. “Modelling Ourselves: What the Free Energy Principle Reveals about Our Implicit Notions of Representation.” *Synthese* (April). https://doi.org/10.1007/s11229-021-03140-5.

Sireteanu, Ruxandra, and Regina Rettenbach. 2000. “Perceptual Learning in Visual Search Generalizes over Tasks, Locations, and Eyes.” *Vision Research* 40 (21): 2925–49. https://doi.org/10.1016/S0042-6989(00)00145-0.

Sloane, Michael E., John W. P. Ost, Deborah B. Etheriedge, and Stephen E. Henderlite. 1989. “Overprediction and Blocking in the McCollough Aftereffect.” *Perception and Psychophysics* 45 (2): 110–20. https://doi.org/10.3758/BF03208045.

Slors, Marc, Leon de Bruin, and Derek Strijbos. 2015. *Philosophy of Mind, Brain and Behaviour*. Amsterdam: Boom.

Sprenger, Jan, and Stephan Hartmann. 2019. *Bayesian Philosophy of Science*. Oxford, UK: Oxford University Press.

Sprevak, Mark. 2013. “Fictionalism about Neural Representations.” *The Monist* 96 (4): 539–60. https://doi.org/10.5840/monist201396425.

Stadler, Friedrich. 2001. *The Vienna Circle: Studies in the Origins, Development, and Influence of Logical Empiricism*. Vienna: Springer.

Steinbock, Anthony J. 1995. “Generativity and Generative Phenomenology.” *Husserl Studies* 12 (1): 55–79. https://doi.org/10.1007/BF01324160.

Steinle, Friedrich. 2009. “Concepts, Facts, and Sedimentation in Experimental Science.” In *Science and the Life-World: Essays on Husserl's Crisis of European Sciences*, edited by David Hyder and Hans-Jörg Rheinberger, 199–216. Redwood City, CA: Stanford University Press. https://doi.org/10.1515/9780804772945-015.

Sterzer, Philipp, Rick A. Adams, Paul Fletcher, Chris Frith, Stephen M. Lawrie, Lars Muckli, Predrag Petrovic, Peter Uhlhaas, Martin Voss, and Philip R. Corlett. 2018. “The

Predictive Coding Account of Psychosis." *Biological Psychiatry* 84 (9): 634–43. https://doi.org/10.1016/j.biopsych.2018.05.015.

Strevens, Michael. 2011. *Depth: An Account of Scientific Explanation*. Cambridge, Mass.: Harvard University Press.

Suárez, Mauricio. 2010. "Scientific Representation." *Philosophy Compass* 5 (1): 91–101. https://doi.org/10.1111/j.1747-9991.2009.00261.x.

Sugden, Robert. 2000. "Credible Worlds: The Status of Theoretical Models in Economics." *Journal of Economic Methodology* 7 (1): 1–31. https://doi.org/10.1080/135017800362220.

Sussman, Alan. 1975. "Mental Entities as Theoretical Entities." *American Philosophical Quarterly* 12 (4): 277–88. http://www.jstor.org/stable/20009586.

Swanson, Link R. 2016. "The Predictive Processing Paradigm Has Roots in Kant." *Frontiers in Systems Neuroscience* 10 (October). https://doi.org/10.3389/fnsys.2016.00079.

Tee, Sim Hui. 2020. "Generative Models." *Erkenntnis* 88 (1). https://doi.org/10.1007/s10670-020-00338-w.

Teller, Paul. 2001. "Twilight Of The Perfect Model Model." *Erkenntnis* 55 (3): 393–415. https://doi.org/10.1023/A:1013349314515.

Thagard, Paul. 2005. *Mind: Introduction to Cognitive Science*. 2nd ed. Cambridge, Mass.: The MIT Press.

Thoma, Johanna. 2019. "Decision Theory." In *The Open Handbook of Formal Epistemology*, edited by R. Pettigrew and J. Weisberg, 57–106. PhilPapers Foundation. https://books.google.at/books?id=p6j4xwEACAAJ.

Thompson, Evan. 2010. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, Mass., London: The Belknap Press.

Thomson, Eric, and Gualtiero Piccinini. 2018. "Neural Representations Observed." *Minds and Machines* 28. https://doi.org/10.1007/s11023-018-9459-4.

Toon, Adam. 2012. *Models as Make-Believe: Imagination, Fiction, and Scientific Representation*. New York, NY: Palgrave Macmillan.

———. 2023. *Mind as Metaphor: A Defence of Mental Fictionalism*. Oxford: Oxford University Press.

Treisman, Anne M., and Garry Gelade. 1980. "A Feature-Integration Theory of Attention." *Cognitive Psychology* 12 (1): 97–136. https://doi.org/10.1016/0010-0285(80)90005-5.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–60.

Urgen, Burcu A., Marta Kutas, and Ayse P. Saygin. 2018. "Uncanny Valley as a Window into Predictive Processing in the Social Brain." *Neuropsychologia* 114 (June): 181–85. https://doi.org/10.1016/j.neuropsychologia.2018.04.027.

Urgen, Burcu A., and Ayse P. Saygin. 2020. "Predictive Processing Account of Action Perception: Evidence from Effective Connectivity in the Action Observation Network." *Cortex* 128 (July): 132–42. https://doi.org/10.1016/j.cortex.2020.03.014.

Uzgalis, William. 2000. "John Locke." In *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta, Spring 2020 Edition. <https://plato.stanford.edu/archives/spr2020/entries/locke/>.

Van de Cruys, Sander, and Pieter van Dessel. 2021. “Mental Distress through the Prism of Predictive Processing Theory.” *Current Opinion in Psychology* 41 (October): 107–12. https://doi.org/10.1016/j.copsyc.2021.07.006.

Vance, Jona, and Dustin Stokes. 2017. “Noise, Uncertainty, and Interest: Predictive Coding and Cognitive Penetration.” *Consciousness and Cognition* 47 (January): 86–98. https://doi.org/10.1016/j.concog.2016.06.007.

Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 2016. *The Embodied Mind: Cognitive Science and Human Experience*. Revised ed. Cambridge, Mass., London: The MIT Press.

Venter, Elmarie. 2021. “Toward an Embodied, Embedded Predictive Processing Account.” *Frontiers in Psychology* 12 (January): 543076. https://doi.org/10.3389/fpsyg.2021.543076.

Vincent, Benjamin T. 2014. “A Tutorial on Bayesian Models of Perception.” *Journal of Mathematical Psychology* 66: 103–14. https://doi.org/10.1016/j.jmp.2015.02.001.

Walsh, Kevin S., David P. McGovern, Andy Clark, and Redmond G. O’Connell. 2020. “Evaluating the Neurophysiological Evidence for Predictive Processing as a Model of Perception.” *Annals of the New York Academy of Sciences* 1464 (1): 242–68. https://doi.org/10.1111/nyas.14321.

Weilnhammer, Veith A., Heiner Stuke, Philipp Sterzer, and Katharina Schmack. 2018. “The Neural Correlates of Hierarchical Predictions for Perceptual Decisions.” *The Journal of Neuroscience* 38 (21): 5008–21. https://doi.org/10.1523/JNEUROSCI.2901-17.2018.

Weisberg, Michael. 2007. “Three Kinds of Idealization:” *Journal of Philosophy* 104 (12): 639–59. https://doi.org/10.5840/jphil20071041240.

———. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Westbrook, R. F., and W. Harrison. 1984. “Associative Blocking of the Mccollough Effect.” *The Quarterly Journal of Experimental Psychology Section A* 36 (2): 309–18. https://doi.org/10.1080/14640748408402161.

Wiener, Norbert. 1914. “Relativism.” *The Journal of Philosophy, Psychology and Scientific Methods* 11 (21): 561–77. https://doi.org/10.2307/2012619.

———. 2019. *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, Mass.: The MIT Press.

Wiese, Wanja. 2017. “What Are the Contents of Representations in Predictive Processing?” *Phenomenology and the Cognitive Sciences* 16 (4): 715–36. https://doi.org/10.1007/s11097-016-9472-0.

———. 2018. *Experienced Wholeness: Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*. Cambridge, Mass.: The MIT Press.

Wiese, Wanja, and Thomas K. Metzinger. 2017. “Vanilla PP for Philosophers: A Primer on Predictive Processing.” In *Philosophy and Predictive Processing*, edited by Wanja Wiese and Thomas Metzinger. Frankfurt: MIND Group. https://doi.org/10.15502/9783958573024.

Williams, Daniel. 2018. “Predictive Processing and the Representation Wars.” *Minds and Machines* 28 (1): 141–72. https://doi.org/10.1007/s11023-017-9441-6.

Wimsatt, William C. 2006. “Reductionism and its Heuristics: Making Methodological Reductionism Honest.” *Synthese* 151 (3): 445–75. https://doi.org/10.1007/s11229-006-9017-0.

Windt, Jennifer M., Dominic L. Harkness, and Bigna Lenggenhager. 2014. “Tickle Me, I think I might be Dreaming! Sensory Attenuation, Self-Other Distinction, and Predictive Processing in Lucid Dreams.” *Frontiers in Human Neuroscience* 8 (September). https://doi.org/10.3389/fnhum.2014.00717.

Wittgenstein, Ludwig. 2009. *Philosophische Untersuchungen = Philosophical investigations*. Edited by G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte. 4th ed. Chichester, West Sussex, Malden: Wiley-Blackwell.

Worrall, John. 2009. “Normal Science and Dogmatism, Paradigms and Progress: Kuhn ‘Versus’ Popper and Lakatos.” In *Thomas Kuhn*, edited by T. Nickles, 65–100. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511613975.005.

Wright, Georg Henrik von. 1994. “Introduction: Analytische Philosophie — eine historisch-kritische Betrachtung.” In *Analyomen / Analyomen: Proceedings of the 1st Conference Perspectives in Analytical Philosophy*, edited by Georg Meggle and Ulla Wessels, 1–30. Berlin, Boston, Mass.: De Gruyter. https://doi.org/10.1515/9783110885675-004

Van Es, Thomas van. 2020. “Minimizing Prediction Errors in Predictive Processing: From Inconsistency to Non-Representationalism.” *Phenomenology and the Cognitive Sciences* 19 (5): 997–1017. https://doi.org/10.1007/s11097-019-09649-y.

Yarritu, Ion, and Helena Matute. 2015. “Previous Knowledge can induce an Illusion of Causality through Actively Biasing Behavior.” *Frontiers in Psychology* 6 (March): 389. https://doi.org/10.3389/fpsyg.2015.00389.

Yon, Daniel, and Chris D. Frith. 2021. “Precision and the Bayesian Brain.” *Current Biology* 31 (17): 1026–32. https://doi.org/10.1016/j.cub.2021.07.044.

Zahavi, Dan. 2018. “Brain, Mind, World: Predictive Coding, Neo-Kantianism, and Transcendental Idealism.” *Husserl Studies* 34 (1): 47–61. https://doi.org/10.1007/s10743-017-9218-z.

———. 2021. “Applied Phenomenology: Why It Is Safe to Ignore the Epoché.” *Continental Philosophy Review* 54 (2): 259–73. https://doi.org/10.1007/s11007-019-09463-y.

Zhang, Jiajie, and Donald A. Norman. 1995. “A Representational Analysis of Numeration Systems.” *Cognition* 57 (3): 271–95. https://doi.org/10.1016/0010-0277(95)00674-3.

Zhang, Jiajie, and Vimla L. Patel. 2006. “Distributed Cognition, Representation, and Affordance.” *Pragmatics & Cognition* 14 (2): 333–41. https://doi.org/10.1075/pc.14.2.12zha.