ENACTIVISM AND THE COGNITIVIST TRIAD: FUNCTIONAL ROLES, REPRESENTATION, AND COMPUTATION

by
Henrique Mendes Gonçalves

Submitted to
Central European University
Department of Philosophy

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Supervisor: Timothy Martin Crane

Vienna, Austria
2025

Copyright Notice

Henrique Mendes Gonçalves

Vienna, August 6, 2025

To the myriad of lifeforms that had to try, fail, thrive, and fail again
in this strange game of existing;
so that I could have these hands, eyes, body, and brain
—and, on occasion, find a measure of success.
I stand on the shoulders of those tiny giants.

Abstract

This dissertation critically examines the widely held assumption that enactivism, a prominent 4E approach to cognition, is fundamentally incompatible with the three core tenets of classical cognitive science: functionalism, representationalism, and computationalism. Challenging the prevailing narrative of a revolutionary paradigm shift, it argues for a more synthetic and integrative understanding. It proceeds by first clarifying the unique status of enactivism among other 4E theories (Chapter I) and then analyzing its internal diversity (Chapters II and III), positing that autopoietic enactivism offers a more robust framework than its sensorimotor counterpart by grounding cognition in biological principles of autonomy and sense-making. The dissertation then systematically tests the compatibility of this refined enactivist framework with each cognitivist tenet. Far from being antagonistic, autopoietic enactivism is shown to be a radical, biologically-grounded form of functionalism (Chapter III). The challenge of representationalism is met by adopting a deflationary account of subpersonal representation, which treats content not as an intrinsic property but as a pragmatic, heuristic tool for scientific modeling, thereby dissolving what the literature has called the "hard problem of content" (Chapter IV). Finally, compatibility with computationalism is established through a non-semantic, mechanistic framework, reframed through a perspectival lens where computational ascriptions are constrained by a system's causal structure (Chapter V). Ultimately, the dissertation demonstrates that with careful conceptual refinement, a coherent version of enactivism can be reconciled with modified versions of the cognitivist triad, thus bridging a significant theoretical divide in the sciences of the mind.

Acknowledgements

There are many people to whom I owe a great debt of gratitude. This dissertation would not have been possible without their help.

I would like to start by thanking my supervisor, Tim, for encouraging me to hold my work to the highest standards. Your guidance has shaped the way I think and write. I am grateful not only for the scholarly training I received under your supervision, but also for the patience, kindness, and understanding with which you provided it.

For the precious feedback given on my project proposal before the first-year comprehensive examination, I would like to thank Mike, who was also very generous in offering me the opportunity to serve as his teaching assistant the following year, and Kati, whose constructive insights were crucial in helping me develop a strong defensible proposal.

I was quite fortunate to meet many friendly and very bright people at CEU. I would like to extend my gratitude to them and, particularly, to my friends from the 2021–2022 MA and PhD cohorts. I would also like to thank my colleagues from the work-in-progress seminars for the very helpful feedback received during the sessions and for the stimulating discussions that helped me refine my arguments.

I am also very grateful to my parents, Luiz and Rita, as well as to my sisters, Letícia and Viviane, for all the support and encouragement they offered me during these four years. Being far from them was the hardest part of this journey, but every moment we spent together during visits and reunions was deeply cherished and gave me renewed strength to keep going.

Finally, I would like to thank my wife, Bianca, for her love and encouragement. Her support has made the challenges lighter and the successes brighter, and I am deeply grateful to have shared this journey with her.

Table of contents

List of tables and figure:

Tables:

Figure:

INTRODUCTION

1. CONTEXT

This dissertation critically investigates the intricate relationship between the enactivist theory of mind and cognition and three central doctrines that have structured the conceptual and explanatory landscape of classical cognitive science: representationalism, computationalism, and functionalism. These commitments have profoundly shaped, sometimes explicitly and sometimes by default, how theorists delineate the objects and methods of the scientific study of mind. Yet the story of their emergence, dominance, mutual interplay, and, ultimately, their contestation from within cognitive science itself is highly complex, both historically and philosophically. Each approach brings distinct philosophical presuppositions and historical legacies, and understanding their evolution and contemporary status is essential for any critical appraisal of present-day debates—especially when seeking to clarify what enactivism rejects, what it inherits, and what it seeks to transform. Thus, I will start by briefly describing the core ideas and the historical genesis of these doctrines, attempting to offer a more solid and contextually informed grounding for the discussions that will follow in the next five chapters.

1.1. REPRESENTATIONALISM

Among the three tenets, representationalism is the oldest and most deeply entrenched. Over subsequent centuries, this motif was iteratively adapted and transformed by medieval Scholastics, Early Modern philosophy, Kantian transcendental epistemology, as well as early experimental psychology in the 19th century. Thus, the notion of mental representation became a centerpiece in the history of philosophical and psychological theories of cognition and perception. However, the rise of the behaviorist paradigm in the early 20th century dramatically altered the theoretical terrain. In seeking

to establish psychology as a rigorously scientific and independent discipline, especially in the United States, scholars such as John Watson and later B. F. Skinner developed a framework that explicitly rejected all appeals to inner mental representations, insisting instead that explanation be confined to observable patterns of stimulus and response (Watson 1913; Skinner 1953).

For several decades—roughly from the 1920s through the 1950s—this approach was hegemonic, relegating topics such as perception, memory, imagination, and representation itself to the periphery of the discipline (Miller 2003). Topics requiring an appeal to introspection or inner mental states—including perception, memory, planning, problem-solving, imagination, and especially representation—were marginalized or addressed only in strictly operationalist terms. While certain topics such as animal navigation, learning, and language use occasionally forced some covert reintroduction of mediational hypotheses, these were constrained to behaviorally tractable proxies.

The so-called cognitive revolution of the 1950s and 1960s thus marks not the arrival of representationalism for the first time, but its *renaissance* as a central scientific and philosophical principle—one now transformed by advances in computer science, logic, and formal modeling. Noam Chomsky's influential critique of Skinner's behaviorist account of language acquisition (Chomsky 1959), together with experimental and conceptual work by figures like George Miller (Miller, Galanter, and Pribram 1960), Jerome Bruner (1966), Donald Broadbent (1958), Ulric Neisser (1967), as well as Allen Newell and Herbert A. Simon (1972) paved the way for internal, representational "information-processing" accounts to regain respectability. Representationalism's return did not involve a simple regression to philosophical intuition, but in fact advanced a new, highly formalized vision: cognition as the processing and transformation of structured, rule-governed representations, closely modeled on the computations of digital computers.

This scientific form of representationalism was more explicit, more formally articulated, and—crucially—more closely bound to methodologically controlled experimental paradigms than anything available in nineteenth-century philosophy or even early Wundtian or Jamesian

psychology. For this reason, the classical "cognitivist" period, from the late 1950s through the mid-1980s, is often described as seeing an explosion of representational explanation in the sciences of the mind, even as debates about what counts as a representation—formal, mental, neural, digital, analog—raged (Fodor 1975; Haugeland 1981; Dennett 1982). In summary, representationalism is the oldest of the cognitivist triad, conceptually and historically, and it was the recovery and scientific reformulation of this doctrine that lay at the heart of the cognitive revolution's critique of behaviorism and the birth of modern cognitive science (Marr 1982; Miller 2003; Gardner 1985).

1.2. COMPUTATIONALISM

Computationalism, in turn, originated in the flourishing mid-century disciplines of mathematical logic, cybernetics, and computer science. Its central claim is that cognitive processes are computational processes: operations defined over formal states—symbols, strings, tokens—following explicit rules, closely analogous to those performed by a Turing machine (Turing 1936; see also Minsky 1967). Alan Turing's demonstration that discrete, rule-based operations on symbolic configurations could realize any effectively calculable function inspired many to see the mind or brain as a special kind of "biological computer."

The formal properties of digital computation offered both a philosophical and a technical template for modeling intelligence. Key early figures included McCulloch and Pitts (1943), who developed an early connectionist model of neural networks using Boolean logic; Frank Rosenblatt (1958), who improved McCulloch-Pitts's model by making it capable of learning, thus creating the model called the *perceptron*; and Newell, Shaw, and Simon (1958), who developed artificial intelligence programs capable of problem-solving and heuristic search. As the theory developed, scientists increasingly saw perception, reasoning, language, and even imagination as computations in the sense of transformations upon internal representations in accordance with explicit rules, a view that was elegantly synthesized in Jerry Fodor mid-1970s work (1975).

Philosophically, computationalism provided a new way to naturalize and constrain the ontology of representation. Rather than being mysterious "ghostly" entities, mental representations could be formally specified, manipulated, and ultimately implemented, at least in principle, on universal hardware. Fodor's language of thought hypothesis offered perhaps the clearest theoretical articulation of this view, identifying thought with computation over structured, syntactically and semantically rich symbols (1975, 1981). Cognitive modeling—algorithms, automata, semantic networks, and so forth—became central to both theoretical and experimental research. David Marr's (1982) influential analysis of vision, which distinguished different levels of computational explanation—computational, algorithmic, and implementational—further entrenched the computationalist paradigm. Marr's hierarchy enabled researchers to clarify which kinds of representations and computations were relevant at which explanatory level, with the computational paradigm embracing levels from abstract, mathematical specification down to physical neural instantiation.

1.3. FUNCTIONALISM

Functionalism is the view that what makes something a mental state—a belief, a desire, a pain—is not its internal structure or physical makeup, but rather the role it plays in the overall functioning of a cognitive system. On this view, mental states are identified by their characteristic patterns of causation: their relations to sensory experiences, to behavioral outputs, and to each other. This way of understanding the mind emphasizes the importance of the roles or functions that mental states perform, rather than the materials out of which they are made. Thus, an important consequence of functionalism is the possibility of *multiple realization*, even if this is not a requirement of the basic theory: the same kind of mental state could in principle be instantiated in very different kinds of physical systems, as long as the relevant functional organization is preserved.

Some functionalists, most notably Hilary Putnam (Putnam 1967a, 1967b), have suggested that the mind can be usefully compared to the software of a computer. Just as a particular program

can run on many kinds of hardware, so too (according to this analogy) might mental processes be "implemented" in different biological or artificial substrates. However, it is important to note that this computational analogy, while intuitive for some functionalist views, is also not a necessary feature of functionalism as such, and many functionalists do not appeal to it. By focusing on causal roles rather than physical details, functionalism offers a flexible approach to understanding the mind. It is compatible with a wide range of empirical research programs, and has been influential in both philosophy and cognitive science.

1.4. POSTCOGNITIVISM AND ENACTIVISM

By the 1980s, the classical cognitivist paradigm—defined, again, by its commitments with representationalism, computationalism, and functionalism—was put under pressure by a series of innovations in the field. Most dramatically, neoconnectionism, having now overcome the limitations of the earlier connectionist models, such as McCulloch and Pitts's and Rosenblatt's perceptron, offered neural network models whose operation was neither explicitly symbolic nor necessarily functionally decomposable into discrete, localized "states" or "modules" (Rumelhart and McClelland 1986a, 1986b; Smolensky 1988). These proposals challenged both the architectural commitments of classicism and the centrality of explicit, compositional representations (Churchland 1986).

Simultaneously, alternative theoretical currents started to gain momentum, prompted by a renewed interest in James J. Gibson's (1979) ecological approach to the psychology of perception; the development of new applications of dynamical systems theory to psychology (Kelso 1995; Thelen and Smith 1993, 1994), novel studies in the field of distributed cognition (Hutchins 1995); an emerging, decentralized, non-representationalist, and non-computational approach in robotics (Brooks 1991); as well as the publication of the enactive approach's manifesto book *The Embodied Mind* by Francisco Varela, Eleanor Rosch, and Evan Thompson (1991). Collectively, this variegated group of research programs has been labelled "postcognitivist," due to their perceived

abandonment of some of the central motifs of the cognitivist framework. Beyond this self-definition in negative terms, however, these approaches also tended to emphasize real-time interaction, bodily constraints, as well as material, cultural, and social scaffolding in their methodology and topics—i.e., they tended to emphasize context in the broadest sense.

Among these, enactivism stands out both for the radicality of its critique and its ambition to offer new foundations. Drawing on cybernetics, phenomenology, and theoretical biology, enactivism proposes that cognition is fundamentally about autonomous, self-maintaining, sensorimotor engagement with the world, not about the manipulation of internal models of a pre-given reality (Varela, Thompson, and Rosch 1991; Thompson 2007; Noë 2004; Di Paolo, Buhrmann, and Barandiaran 2017). According to widespread narratives, what sets the enactive approach apart, historically and philosophically, is its direct and comprehensive challenge to the very presuppositions of the cognitivist triad.

Broadly characterized, enactivism claims, therefore, that the world is not "represented" internally but enacted in the ongoing dynamic coupling between organism and environment; that computation is not a sufficient—or even strictly necessary—theoretical posit for explaining living cognition; and that functional decomposition is limited unless it takes into account the organic, lived unity of acting agents. Such a stance may seem, at first glance, to be flatly incompatible with the commitments of classical cognitive science.

This dissertation goes against this narrative and proposes that, albeit indeed transformative in many aspects, this strongly oppositional characterization of the enactive framework may be too impetuous. Thus, the chapters that follow aim to clarify what is genuinely at stake in these alleged incompatibilities. The task is not only to map lines of conflict but also to identify points of continuity and transformation—cases where enactivism appropriates, reworks, or even inadvertently depends on elements of the very doctrines it critiques. In this way, the project seeks to move beyond the conventional oppositions that dominate the literature and to provide a more nuanced

account of enactivism's place in the evolving landscape of cognitive science. I will now present a short survey of the way my argument will be developed by briefly considering the contents of each chapter and section.

2. TEXT

I have divided this dissertation into two parts. The first, entitled "The Varieties of Enactivism," provides an expository and critical analysis of enactive cognition's distinctiveness among 4E theories and explores its internal diversity, assessing which variant of enactivism offers a more robust framework. Thus, it attempts to answer the following questions: How is enactivism distinct from embodied, embedded, and extended approaches in the cognitive sciences and philosophy of mind, with which it is frequently grouped together? What types of enactivism are there, or is it a monolithic, unified theory? Is one of these varieties of the theory more promising than the others, and why?

In turn, Part II, "Enactivism, Representation, and Computation," shifts its focus to more granular and specific inquiries into compatibility, specifically scrutinizing the potential for reconciliation between the enactive approach, on the one side, and two of the central tenets of mainstream cognitive science, on the other: representationalism and computationalism. It attempts to answer three main questions: Is there any significant sense in which enactivism could be made compatible with representationalism and computationalism? How would such compatibility work? What sort of compromises would be needed from each side? Notice that the last chapter of the first part, Chapter III, also serves a transitional role: beyond discussing the autopoietic variety of enactivism, it also presents that theory as a more specific and sophisticated form of functionalism, thus filling a potential gap left by the lack of a specific chapter dedicated to functionalism in the second part of the dissertation.

I will now present a general outline of the arguments and topics discussed in each one of the chapters. The first, "Embodied, Embedded, Extended, and Enactive Cognition," sets the stage by analyzing the 4E approaches. It highlights the frequent ambiguity in distinguishing these notions and argues that most 4E frameworks remain compatible with classical cognitive science, with enactivism being the sole exception. A central analytical tool used in that chapter is the distinction between causation and constitution—i.e., the idea that an external factor is causally relevant for cognition if it merely influences an internal cognitive process, but constitutively relevant if it's an indispensable part of the process itself. This distinction gives rise to "weak" and "strong" varieties of E theories.

For example, weak embodied cognition (§1.1) posits that the body's influence is merely causally relevant to cognition, as exemplified by Alvin Goldman and Frédérique de Vignemont's use of the notion of "B-formatted representations." This view, which sanitizes bodily data for central processing, remains fully compatible with computationalism, representationalism, and functionalism, as it simply expands input sources without challenging neurocentrism. Conversely, strong embodied cognition (§1.2) asserts that the body's structures and dynamics are constitutive of cognitive activity. While this perspective might initially seem to contradict classical cognitivist tenets, the chapter argues that a broader interpretation of the notions of computation—including, for instance, the notions of morphological or wide computation—, representation—as exemplified in Lawrence Barsalou's (1999) perceptual symbol systems—and functional role—e.g., allowing for functions anchored in richly textured living systems—can accommodate strong embodiment without abandoning these frameworks altogether.

Similarly, the chapter distinguishes *embedded cognition*, which I also call *weak extended cognition,* from *strong extended cognition* (§2). Embedded cognition (§2.1) posits that environmental contexts play a significant causal role in cognitive processes, acting as external scaffolds that offload mental work. Clearly, this view is modest enough to be compatible with mainstream cognitivism, as

environmental elements are seen as influencing internal information processing without constituting it. Conversely, strong extended cognition (§2.2), as advocated by Andy Clark and David Chalmers (1998), advances the more radical claim that external elements can be constitutive parts of cognitive processes themselves, given certain functional criteria. This position is encapsulated by the *parity principle*, which states that if a part of the world functions in a way that, were it located in the head, we would deem it cognitive, then it must be considered a constitutive part of the cognitive process itself. Strong extended cognition relies both on the notion of *wide computationalism*, which asserts that computational processes can be realized across brain, body, and environment, and on what is often called *extended functionalism*, in order to justify the inclusion of extracorporeal elements within the cognitive system. Moreover, it preserves a certain type of representationalism, treating external artifacts as analogous to internal representations. Thus, while strong extended cognition stretches the physical boundaries of the cognitive system, it generally retains the core explanatory principles of traditional cognitive science.

Enactivism (§3), however, departs from this pattern. Unlike the other 4E approaches, it fundamentally rejects the possibility of a weak variant, as it insists that cognition is always constituted by the dynamic coupling of brain, body, and environment. In this view, the cognitive system is not an internal core that occasionally extends outward, but is inherently a relational process defined by its continuous, recursive interaction with the world. This holistic stance means that enactivism resists the analytical distinctions of causation versus constitution, or weak versus strong theoretical variants, as applied to other E-theories. Therefore, the compatibility between enactivism and the core tenets of cognitivism is significantly more challenging to establish. This is what motivates a more detailed investigation, in the second half of the dissertation, of its compatibility with functionalism, representationalism, and the computational theory of mind.

Before that, however, the second and third chapters explore enactivism's own internal diversity, as the term encompasses some relatively distinct research programs. While often

subdivided into *sensorimotor*, *autopoietic*, and *radical enactivism*, a closer examination reveals important differences in the scope and theoretical presuppositions of each one of these variants. Chapter II, "Sensorimotor Enactivism," focuses on the most minimalist version of the enactive framework, primarily concerned with perception, action, and phenomenal consciousness in its sensory dimensions. Its central concept is that of sensorimotor contingencies, which are defined as the law-like dependencies between motor activity and sensory change.

The chapter reconstructs Susan Hurley's (1998) influential critique of what she called the "input-output model" or "classical sandwich model" of the mind, which posits perception and action as peripheral modules separate from a central cognitive core (§1). Hurley argues against this linear, unidirectional flow of information, emphasizing instead a complex network of sensorimotor feedback loops that constitute mental processes. Noë and O'Regan's sensorimotor theory builds upon these insights, arguing that the phenomenal qualities of perception, including intermodal and intramodal variations, emerge from the lawful coordination of afferent and efferent channels through active exploration, rather than from neural correlates or specific nerve pathways (§2). The chapter provides examples such as the vestibulo-ocular reflex and Bach-y-Rita's sensory substitution experiments to illustrate how bodily movements and environmental interaction are constitutive of perception.

However, despite its ingenuity and success in reimagining basic cognition, sensorimotor enactivism is shown to have significant shortcomings (§3). Most critically, sensorimotor enactivism is criticized for its limited explanatory scope, failing to provide substantive accounts for higher-level cognitive capacities such as language, memory, or abstract reasoning, and remains silent on the crucial aspects of normativity, valence, and the intrinsic meaningfulness of interactions for the organism. While it captures the structure of perceptual engagement, it does not explain why certain interactions matter to the system, thereby lacking the depth required for a comprehensive theory of mind.

The third chapter, "Autopoietic Enactivism," discusses an arguably more satisfactory alternative which grounds cognition in the biological organization of living systems. Historically predating the sensorimotor variant, this framework builds upon the *theory of autopoietic systems* developed by Humberto Maturana and Francisco Varela (1980). The theory defines living systems as self-producing and self-maintaining unities, characterized by a network of reactions that recursively produce their own components and define their boundary (§1). The chapter delves into the philosophy of biology, exploring various strategies for defining life (§1.1) and positioning autopoiesis as a functionalist theory thereof (§1.2), abstracting from contingent material implementations to focus on an invariant organizational pattern. It also discusses the extension of autopoiesis from first-order systems—i.e., cells—to second-order systems—multicellular organisms—, acknowledging the theoretical difficulties in defining clear boundaries for the latter and leading to the more general, functionally defined concept of autonomy.

A crucial contribution of autopoietic enactivism is its account of the teleological dimension of living systems (§1.3), moving beyond purely mechanistic or teleonomic explanations. This is what leads Varela to extend autopoietic theory from the domain of life, in his earlier works with Maturana, to that of cognition (§2), introducing the notion of sense-making (§2.1) to capture the organism's purposive and norm-sensitive engagement with its environment. This concept directly addresses the shortcomings of sensorimotor enactivism by explaining how intrinsic normativity emerges from the organism's need to sustain its own viability through adaptive interactions. The chapter also introduces neurophenomenology (§2.2) as a methodological tool to bridge the third-person scientific perspective with the first-person lived experience of sense-making, thereby vindicating a biologically grounded form of teleology. This, in turn, leads to the life-mind continuity thesis (§2.3), which posits that all living systems, by virtue of their very autopoietic and autonomous organization, exhibit a minimal form of cognition, and conversely, that all minded systems are necessarily living. This radically decenters the brain, positioning the nervous system as a specialized organ that facilitates sense-making within an already living, autonomous organism.

A key contribution of Chapter III is its re-evaluation of the relationship between autopoietic enactivism and functionalism (§3). While often portrayed as antagonistic, the chapter argues that autopoietic enactivism can be understood as a more specific and sophisticated, biologically-grounded, and organizationally-specific type of functionalism. This resolution leverages the conceptual shift from strict autopoiesis to the more flexible concept of autonomy, demonstrating how extended cognitive systems can be understood as higher-order manifestations of this self-constituting identity, thereby accommodating insights from extended cognition without sacrificing a commitment to organizational closure.

After having discussed the internal structure of enactivism in detail, we move on to Part II of this study, which is dedicated to compatibility issues. Thus, the fourth chapter, "Enaction and Representation," transitions to the more constructive part of the dissertation, directly assessing the compatibility of enactivism with representationalism, a cornerstone of classical cognitivism. This chapter can also be seen as a response to some of the main concerns of radical enactivism, which, despite sometimes being treated as a variant of enactivism, is argued here to function primarily as a critical endeavor. It begins by exploring the various theoretical and historical motivations for antirepresentationalism within 4E cognition (§1), then reconstructing Daniel Hutto and Erik Myin's (2013) central argument against representationalism, which they call the "hard problem of content" (§2). This problem highlights the alleged impossibility of naturalizing what they call *information-as-content*—i.e., information as semantic content with satisfaction conditions—from the more basic, naturalistically acceptable notion of *information-as-covariance*—i.e., reliable co-occurrence.

As a solution, Hutto and Myin propose denying any role for mental representation in "basic minds," defined as the prelinguistic domains of action and perception, while still conceding its presence in "linguistic minds." This puzzling proposal, however, has been characterized as unsatisfactory in the literature for its lack of coherence and *ad hoc* character, a view I share. Thus, to effectively overcome the hard problem of content and bridge this apparent divide, Chapter IV

resorts to a different strategy: a pragmatic account of mental representations (§4), specifically of the subpersonal type (§3), that treats them as useful modelling tools without ontologizing them.

The main inspiration for this proposal is Frances Egan's (2010, 2014, 2018) deflationary view of mental representations (§4). Egan proposes that representational content is not an intrinsic metaphysical property but rather an "intentional gloss" used by cognitive scientists to make sense of computational theories and map them onto cognitive tasks. I argue that, by treating representational content as a useful construct for scientific modeling rather than a substantive claim about the nature of cognition, Egan's perspective effectively dissolves the so-called hard problem of content. This approach allows for the continued, pragmatic use of representational vocabulary without requiring the metaphysical commitments that enactivists critique. This has the significant advantage of reducing the perceived gap between traditional representational research—which forms a substantial part of the established *corpus* of cognitive science—and the still-developing proposals of enactivism, thereby fostering a more integrated understanding of cognitive phenomena. The chapter concludes by suggesting that similar deflationary strategies, such as mental fictionalism, could be applied to personal-level representations, further harmonizing enactivist and main-stream approaches (§5).

The fifth chapter, "Enaction and Computation," confronts the compatibility of enactivism with computationalism. After defining computation and summarizing the main tenets of the computational theory of mind (§1), it addresses the prevalent view that computation relies *inherently* on representation—a thesis that would be incompatible with the deflationary strategy employed in the previous chapter. With that in mind, the chapter explores the idea of non-semantic computation, particularly through Gualtiero Piccinini's (2015) mechanistic account (§2).

Piccinini's framework defines physical computation objectively as the processing of "medium-independent vehicles" by a functional mechanism according to rules sensitive solely to differences between portions of the vehicles (§2.1). This approach, motivated by the perceived

redundancy of semantic notions in defining computation and the difficulty of preserving objectivity in semantically individuated accounts (§2.2), aims to avoid pancomputationalism—i.e., the idea that all physical systems compute—by insisting that only systems with a *proper function* of computing genuinely do so (§2.3). The chapter then examines the project of applying this non-semantic, mechanistic view of computation to enactive approaches, as proposed by philosophers like Joe Dewhurst and Mario Villalobos (§3). Their proposal, broadly dependent upon wide computationalism, aims to resolve tensions within 4E cognition by accommodating embodied, extended, and enactive insights within a computational framework where systems are not confined by skull or skin (§3.1).

However, the chapter proceeds to identify a significant obstacle in thus instrumentalizing Piccinini's account: his reliance on the notion of "proper functions" as objective teleological ends, which introduces a potential circularity between function and structure (§3.2). To address this, the chapter advocates for an alternative perspectival and deflationary approach to computational function (§4). This view acknowledges the observer-relativity of computational attributions but insists that they are robustly constrained by the intrinsic causal organization and structural regularities of physical systems (§4.1 and §4.2).

Drawing on Maturana and Varela's distinction between "interactions" (actual physical engagements) and "relations" (conceptual distinctions made by observers), the chapter argues that computational attributions arise from the interplay between observer-defined relations and the system's intrinsic, stable interactions, thereby avoiding pancomputationalism without resorting to rigid teleological frameworks. Ultimately, the last chapter concludes that while strict objectivist computationalism may clash with enactivism, a perspectival and deflationary view of computation can preserve explanatory power and scientific utility while aligning seamlessly with enactivism's emphasis on relational, embodied cognition (§5). This unified perspective demonstrates that computational descriptions, rather than being metaphysical truths, function as indispensable interpretive

tools for understanding the complex causal architectures underlying cognition and information processing.

Finally, in this dissertation's Conclusion, I present some of the main ideas that can be distilled from the preceding analyses. I argue that the relationship between enactivism and classical cognitive science is best understood not in terms of an outright opposition, but rather in terms of conceptual realignment. While enactivism challenges the core assumptions of computationalism, representationalism, and functionalism, it also reveals that these frameworks are more malleable than is often assumed. By deflating the metaphysical weight of representational content, reframing computationalism in perspectival rather than objectivist terms, and recognizing a biologically anchored form of functionalism, the supposed gulf between enactivism and classical cognitivism narrows considerably.

PART I: THE VARIETIES OF ENACTIVISM

I. EMBODIED, EMBEDDED, EXTENDED, AND ENACTIVE COGNITION

Over the past few decades, the so-called *4E approaches*—those that emphasize cognition as *embodied*, *embedded*, *extended*, and *enactive*—have gained prominence in the cognitive sciences and philosophy of mind. Yet the label remains ambiguous: it is often unclear how these notions are to be individuated, whether they are mutually compatible, and how they relate to more traditional theories in the field. Although frequently presented as heralding a paradigm shift in the philosophy and sciences of the mind, I will argue that most 4E frameworks remain largely compatible with the core premises of mainstream cognitive science—provided these are properly qualified—, with the sole exception being enactivism.

The idea of four central Es was consolidated by a number of influential publications over the last fifteen years (Menary 2010; Rowlands 2010; Ward and Stapleton 2012; Gallagher 2017, 2023; Newen, de Bruin, and Gallagher 2018), and is typically understood along the following lines:

> EMBODIED COGNITION: Cognition is shaped by the body's anatomy, morphology, sensorimotor capacities, and biological constitution.
> EMBEDDED COGNITION: Cognition is embedded or situated within and dependent on specific environmental contexts.
> EXTENDED COGNITION: Cognition often includes the environment, external artifacts, and tools as its functional elements.
> ENACTIVE COGNITION: Cognition arises through dynamic interactions between an organism and its environment, involving sense-making and autonomous activity.

While some authors have proposed additional "Es" and other letters beyond these four—e.g., "emotive" (Damasio 1994; Barrett 2017), "encultured" (Tomasello 1999; Sterelny 2010),

"situated" (Suchman 1987; Greeno 1998), "technologically-mediated" (González, Bach-y-Rita and Haase 2005), etc.—, these often overlap with, or can be subsumed under the more canonical four categories. For conceptual clarity, I adopt the 4E taxonomy as the heuristic foundation for the analysis that follows. This framework offers a principled way of individuating theories not by thematic concerns, but by their underlying ontological commitments.

Generally speaking, 4E-cognition theories tend to be quite polarizing. Perhaps the most contentious issue concerns their exact role and position within the scientific study of the mind. Should they be understood as a replacement for traditional frameworks, challenging the core assumptions of classical cognitive science? Or are they better viewed as a complement, offering new tools and perspectives without displacing foundational commitments? This *replacement–complement dichotomy* is central to contemporary debates, and clarifying it is crucial in order to understand the stakes of 4E theorizing. A natural way to approach this problem is by asking what 4E theories define themselves against—what they present as their main point of contrast. At first sight, a common target emerges in what is frequently referred to as the *classic* or *cognitivist approach* to the mind. Importantly, within 4E theorizing itself, the existence of weak and strong varieties of E theories, a topic that will be discussed in the next section, is often taken to indicate how far each position diverges from this classical picture. Weaker versions are, at least *prima facie*, more compatible with it, whereas stronger versions push further away from its basic assumptions. As we will see, this variation can be systematically analyzed in terms of the extent to which different E theories reject or accept three core theses commonly associated with classical cognitivism:

> COMPUTATIONALISM: Cognitive activity is constituted by computational processes.
>
> REPRESENTATIONALISM: Cognitive activity is constituted by the manipulation of internal representations.

FUNCTIONALISM: Different types of cognitive activity are individuated by the functional roles cognitive (i.e., mental) states play within a system of inputs, internal processes, and outputs.

Very often, what is taken to be the most distinctive feature of 4E-cognition theories is their shared rejection of these core theses. In what follows, however, I will argue that this characterization is mistaken. All types of 4E-cognition theories—again, with the exception of enactivism—are compatible with these premises. The distinctive status of enactivism will be addressed in greater detail in §4, where we will see why the case for incompatibility and replacement there is more robust.

1. EMBODIED COGNITION

1.1. WEAK EMBODIED COGNITION

At first glance, the claim that the mind is embodied may seem obvious: mental activity depends on the brain—an organ of the body—and is shaped by physiological processes contingent on the body's integrity. Even René Descartes explicitly acknowledges our embodiment in his sixth meditation. There, he draws a contrast between the way a pilot relates to a ship and the way the mind relates to the body, highlighting a deeper kind of union:

> Nature likewise teaches me, through these very feelings of pain, hunger, thirst, and so forth, that I am not present in my body only as a pilot is present in a ship, but that I am very closely conjoined to it and, so to speak, fused with it, so as to form a single entity with it. For otherwise, when the body is injured, I, who am nothing other than a thinking thing, would not feel pain as a result, but would perceive the injury purely intellectually, as the pilot perceives by sight any damage occurring to his ship; and when the body lacks food or drink, I would understand this explicitly, instead of having confused feelings of hunger and thirst. For certainly, these feelings of thirst, hunger, pain, and so

> forth are nothing other than certain confused modes of thinking, arising from the union and, so to speak, fusion of the mind with the body.
>
> (Descartes, AT VII, 81/CSM II, 56)

Descartes' notion of "confused modes of thinking" aligns, in current discussions on embodied cognition, with what is called *interoception*—a term often used to describe the perception of "physiological conditions of the body, such as pain, temperature, itching, muscular and visceral sensations, vasomotor activity, hunger and thirst" (Goldman and de Vignemont 2009, 156). Interoception contrasts with *exteroception*, which refers to the perception of stimuli originating outside the body. It encompasses a series of more specific perceptual phenomena, namely: proprioception or kinesthesia, which are the awareness of the body's position and movement; nociception, the sensation of pain; thermoception, temperature; and equilibrioception, which concerns the perception of balance and spatial orientation relative to gravity. Translating Descartes' disanalogy into this terminology, we have the following: the pilot represents damage to his vessel exclusively through exteroception; an embodied mind, on the other hand, represents it both exteroceptively and through interoception.

The challenges posed by *interoceptive representations* are not confined to Descartes' philosophy; they also trouble cognitivist theories, which view cognition as computation over symbolic representations. Pain, for instance, is often taken as a paradigmatic example of interoception, and it illustrates the complexities that arise when trying to situate interoceptive phenomena within existing theoretical frameworks. Unlike exteroceptive modalities, which typically involve the detection of external objects or events, pain is deeply bound up with affective tone, bodily salience, and motivational force. There is ongoing debate about whether pain should be treated as a representational state, and if so, what exactly it represents (see Tye 1995; Bain 2003, 2007; Klein 2015; Aydede 2020; Martínez 2011). Regardless of one's position in this debate, it is widely acknowledged that interoceptive states like pain differ markedly from more modular, object-directed perceptual states.

Similar considerations apply to other types of interoception. As a representation, what would be, for example, the object of the sensation of hunger? The empty stomach? Or the food desired? The fact that the stomach is empty? Or the fact that the organism needs nourishment?

One may propose to exclude all interoceptive sensations from the scope of the concept of "representation," reserving the term only for states that involve unambiguous intentional objects. But such a move seems *ad hoc*. This becomes clear when we consider recent evidence that interoceptive representations are not confined to tracking the internal state of the body; they also play roles in emotional regulation, decision-making, and linguistic processing. As Shaun Gallagher (2017, 44–45) notes, these extended functions align with the growing acceptance of ideas such as *neuronal reuse* (Dehaene 2007), the *massive redeployment hypothesis* (Anderson 2010), or the *shared circuits models* (Hurley 2007). Despite differences in scope and emphasis, these accounts converge on the idea that neural circuits originally evolved for specific cognitive functions can be repurposed for new tasks through neuroplasticity. The main argument in support of these hypotheses is typically evolutionary in nature: it seems implausible that natural selection could have shaped the human brain specifically to support reading or advanced mathematics. This line of thought appeals to the biological concept of exaptation, the idea that a trait may have originally evolved for one purpose, only to be later co-opted for a different function (Gould and Vrba 1982). A classic example is insect wings, which originally evolved for thermal regulation and balance, but were subsequently adapted for flight, a function that conferred significant evolutionary advantages.

To address the challenge that interoception poses for traditional computationalist theories—particularly those that treat cognitive processes as independent from the body—Alvin Goldman and Frédérique de Vignemont introduce the notion of *B-formatted representations*, where "B" stands for "body" (Goldman and de Vignemont 2009, 155). Their aim is to account for how bodily data—specifically interoceptive representations, which are directly dependent on embodiment—can be integrated into a functionally neutral cognitive architecture. According to Goldman, B-

formatted representations are sanitized: their bodily origin is abstracted away so that they can be processed by the mind's software, which, in the human case, happens to run on an embodied brain. In principle, interoception could just as well operate in a brain-in-a-vat system, so long as the system were somehow fed the relevant B-formatted representations. From a functionalist standpoint, the contingent manner in which such representations enter the system is irrelevant. As noticed by Gallagher (2019, 30–31), this view implies that anatomical features of our perceptual or motor systems have no bearing on the central processing of representations.

To illustrate, let us consider pain one more time. According to some studies, the raw interoceptive mechanism of physical pain—along with its neural correlates—is evolutionarily more primitive and may have formed the basis for the human conceptualization of emotional pain—e.g., grief, dysphoria, heartbreak, or social rejection—and much of the language used to describe it (Eisenberger and Lieberman 2005; Kross *et al.* 2011). If this is indeed the case, however, emotional pain—understood as a B-formatted representation—no longer bears a significant connection to its interoceptive origins. This disconnect allows strictly functionalist accounts to treat it as fully independent of bodily input. In other words, while physical pain may have played a crucial causal role in the phylogenetic emergence of emotional pain, it is not constitutive of it. In its interoceptive "phase," the representation remains tied to the peripheral bodily system and is not yet assimilable into the cognitive architecture. Only when this connection is severed—when the representation is transformed into an amodal, manipulable symbol—does it become accessible to central cognitive processing.

At this point, it may be elucidating to resort to a distinction drawn from analytic philosophy of mind, particularly from debates over the mind-body relation: the difference between causal and constitutive relevance. I will use this distinction in my subsequent attempt at individuating the various types of E-theories in their weaker and stronger varieties. An external factor—e.g., body or environment—is causally relevant for cognitive activity when it merely influences or modulates

a process whose realization is internal—for example, by triggering or shaping neural activity. This is the case, of course, of the type of embodiment proposed by Goldman and de Vignemont. In contrast, it is constitutively relevant when it is part of what makes the process properly cognitive in the first place. On this view, cognition is not merely influenced by body or world; it is partially realized or constituted by them. This distinction helps clarify the internal fault lines among 4E theories: different versions of embodiment, embedment, and extension can be categorized by whether they treat body and environment as merely causally supportive or ontologically indispensable.

Fred Adams and Kenneth Aizawa (2008, Ch. 6) were arguably the first philosophers to apply the causation-constitution distinction to 4E theories, using it to critique the idea of extended cognition. Building on this opposition between causality and constitution, Gallagher (2017, 2023) introduced the terminology of weak and strong 4E theories, distinguishing them by whether bodily and environmental factors are seen as merely causally relevant—in the case of weak theories—or constitutively involved in cognitive processes—in the case of strong ones. I adopt this framework here, and formulate the distinction in the following manner:

> WEAK E THEORY: An E theory that is not committed to either body or environment playing a constitutive role in cognitive activity, but only a causal one.
>
> STRONG E THEORY: An E theory that is committed to either body or environment playing a constitutive role in cognitive activity, and not a merely causal one.

In rejecting the idea that anatomy, physiology, and bodily activity—such as actions and postures—play a constitutive role in cognition, Goldman and de Vignemont (2009) treat these as merely causally relevant factors. What remains are sanitized bodily representations: B-formatted representations derived from interoceptive states, abstracted from their peripheral origins, and centrally processed within a functionally neutral cognitive system. This dissociation perfectly

exemplifies what Gallagher (2017, 28–35) calls weak embodied cognition: a view that preserves the core tenets of the classic cognitivist tradition—computationalism, representationalism, and functionalism—while granting that some representations may originate within the body itself. Yet in this framework, embodiment remains structurally peripheral; the body contributes inputs but does not shape cognition from within. Unsurprisingly, then, weak embodied cognition is fully compatible with the premises outlined in §1, as it was designed to integrate findings from embodiment research without abandoning the foundations of the cognitivist paradigm. A more provocative and philosophically challenging alternative, however, is found in strong embodied cognition.

1.2. STRONG EMBODIED COGNITION

If embodied cognition were taken to mean, more generally, that the body plays a causal role in the way we think—as is claimed by proponents of weak embodied cognition—it is unlikely that many would object to this view. The claim at the heart of the strong variety of this theory, however, is not merely that the body provides the causal scaffolding for cognition, but that it is, in a quite literal and ontologically robust sense, part of the very mechanism that constitutes cognitive activity in the first place. Thus, what distinguishes this position from its weaker counterpart is not simply the insistence on the importance of bodily factors, but the assertion that the very architecture of cognitive processes is inextricably intertwined with the anatomical, physiological, and affective properties of the organism. The body, on this view, is not a passive vessel or a set of peripheral constraints, but an active shaping force whose structures and dynamics are indispensable to the realization of perception, judgment, reasoning, and action.

It is in this sense that strong embodiment is often portrayed as a radical departure from the computationalist, representationalist, and functionalist orthodoxy that has dominated cognitive science during its first few decades of existence. This alleged incompatibility stems from the fact that, according to many, strong embodied cognition denies these core cognitivist tenets. *Grosso modo*, the

reasoning goes along the following lines: if extracranial bodily structures are constitutive of cognitive processes, then representational content cannot be confined to internal symbols, realized through patterns of neuronal connectivity or activation, and computational routines cannot be abstracted from the organism's specific anatomy and physiology. This, it is claimed, also challenges the idea that mental states are individuated solely by their functional roles, regardless of how they are physically realized. Yet, as I shall argue in this section, this portrayal is both overstated and misleading, for it rests on a narrow construal of what is entailed by these three theses—one that fails to appreciate the possibility of understanding them in a more expansive and inclusive manner.

A key insight of strong embodied cognition lies in how the body contributes both to the "pre-processing" and "post-processing" phases of cognition (Gallagher 2023, 10–11). Before information even reaches the brain, the structure and dynamics of the body already modulate what is perceived—for instance, the distance between the eyes determines the nature of binocular vision, the body's orientation or posture affects spatial judgment, and cortisol levels in the blood determine the perception of a stimulus as threatening or neutral. If this is indeed the case, this bodily formatting of sensory input would mean that perception is never entirely neutral or raw; it is always already structured by the body's anatomy and situatedness. Since perception is usually taken to be our cognitive system's main source of information—debates between nativists and empiricists notwithstanding—, its dependence on anatomical and physiological contingencies should also be reflected in more paradigmatic, internal cases of cognitive activity, such as categorization, conceptualization, judgment, and reasoning.

Similarly, after neural processing occurs and a behavioral response is initiated, the body plays a crucial role in shaping action: the specific properties of muscles, joints, and tendons constrain and enable different motor patterns. Strong embodiment theorists claim that these are not merely mechanical outcomes but are, instead, integral to how agents interact with the world in intelligent, adaptive ways, an idea that is very well encapsulated by Andy Clark's concept of "soft

assembly:" the flexible, real-time coordination of bodily and neural resources in response to specific tasks and environmental conditions, without relying on fixed, preprogrammed routines (Clark 1997, 42–45). In a soft-assembled system, components—muscles, joints, perceptual systems, postural adjustments—are dynamically recruited and configured on the fly, depending on context and goals, rather than being governed by a centralized control architecture. This stands in contrast to traditional models of cognition that presuppose a stable internal plan executed by the brain; instead, behavior emerges from the fluid interplay between brain, body, and world.

Strong embodied cognition seems to be supported by a growing body of experimental research demonstrating how bodily states directly influence cognitive processes. Studies show, for instance, that congruence between bodily movements and sentence meaning can facilitate language comprehension—i.e., subjects process sentences more easily when their physical actions align with the spatial direction implied by the sentence's meaning (Glenberg and Kaschak 2002). Similarly, simple changes in posture or proprioceptive feedback can alter spatial perception or resolve ambiguous visual stimuli (Rock and Harris 1967; Roll and Roll 1988). These effects occur without any changes in the sensory input itself, underscoring the body's constitutive role in shaping perception and meaning. Further experiments reveal that affective and physiological states modulate higher-order judgments. Hunger and fatigue, for example, have been shown to influence decision-making, including legal judgments, in ways that are robust and measurable (Danziger *et al.* 2011). Cardiac and respiratory rhythms have been found to bias emotional perception—for example, the heartbeat phase can affect recognition of fearful faces (Garfinkel *et al.* 2014). Such findings challenge the notion that cognition is insulated from the body's physiological conditions and instead suggest that it is, in fact, inherently embedded in the ongoing dynamics of bodily systems.

These results, however, do not necessarily undermine representational, computational, or functional accounts; they may instead prompt revisions to those frameworks, inviting them to incorporate bodily parameters as essential computational variables or functionally significant inputs,

rather than treating them as peripheral modulators of brain-based cognition. Let me illustrate this point, starting with computationalism. If, as is sometimes assumed, this framework is wedded to the idea that cognition is nothing more than the manipulation of amodal, syntactically individuated symbols within the confines of a brain-bound processor, then it is indeed difficult to see how strong embodiment could be anything other than a direct repudiation of this view. There is a growing recognition, however, both in philosophy and in the empirical sciences, that computation need not be conceived exclusively in terms of digital symbol processing, nor need it be localized solely within neural tissue (see Clark 1997; Clark and Chalmers 1998; Wheeler 2005; Piccinini 2015).

The notion of "morphological computation," for instance, has gained increasing traction in recent years, highlighting the ways in which the physical properties of the body—its mass, elasticity, geometry, and so forth—can offload, distribute, or even supplant computational operations that would otherwise have to be performed by the brain (Pfeifer and Iida 2006; Pfeifer, Iida and Gómez 2006; Pfeifer and Bongard 2006). The classic example of passive dynamic walking, in which the mechanical structure of the legs and joints enables stable locomotion with minimal neural control, serves as a paradigmatic illustration of how bodily morphology can be harnessed as an integral component of the cognitive system's computational architecture.

In this light, strong embodied cognition does not so much reject computationalism as it demands its reconceptualization: computation must be understood as a distributed, multi-level phenomenon, realized not only in neural circuits but also in the dynamic interplay of muscles, tendons, bones, and even the organism's ongoing engagement with its environment. A more detailed investigation of these new, more inclusive conceptions of computation will be carried out in Chapter V of this study, together with the assessment of the potential risks in terms of rigor or significance loss when one expands the notion in this manner. For the time being, it suffices to note that embodiment and computationalism are not *a priori* incompatible notions.

A similar point can be made with respect to representationalism. The traditional picture, inherited from the early days of cognitive science, posits that cognition consists in the manipulation of internal representations—mental tokens that stand in for, or "mirror," features of the external world. This view has often been criticized for its tendency to reify representations as static, context-free entities, insulated from the bodily and environmental conditions in which they are embedded. A significant amount of literature on embodied cognition, by contrast, suggests that representations are rarely entirely disembodied (e.g., Barsalou 1999; Lakoff and Johnson 1980, 1999, 2002). Rather, they are deeply shaped by the sensorimotor capacities, affective dispositions, and pragmatic concerns of the cognizing agent. Within the framework of Lawrence Barsalou's perceptual symbol systems, for example, concepts like "hammer" are not abstract, amodal tokens, but partial reactivations of multimodal experiences—including visual, tactile, and proprioceptive states—gained through bodily interaction with objects and situations. To think about a hammer is, on this view, to covertly reenact aspects of the perceptual and motor patterns involved in seeing, gripping, and using one.

Within such frameworks, these simulations are not decorative embellishments to a central symbolic core; they are part of the very substance of conceptual content. Moreover, converging evidence from neuroscience supports this idea: research on neural reuse and shared circuits shows that the same brain regions involved in motor control and perception are frequently recruited for higher-order cognition, such as reasoning and planning (Anderson 2010). In this sense, representations may not be internal stand-ins for external states of affairs, passively mirrored and manipulated in a cognitive black box, but rather dynamically enacted patterns of sensorimotor activity, embedded in the agent's ongoing bodily engagement with the world. To the extent that representationalism is willing to accommodate this richer, embodied and action-oriented conception of representation, there is no principled reason why it must be incompatible with the core commitments of strong embodied cognition.

Functionalism, for its part, due to its close connection to the thesis of multiple realizability—i.e., the idea that the same cognitive function can be realized in a variety of different media—, is often treated as particularly incompatible with embodied approaches. The reasoning goes along the following lines: while the relevant notion of function is, in principle, substrate-neutral, it has often been interpreted in practice as privileging a certain kind of abstract, input-output characterization of cognitive processes—one that abstracts away from the messy, contingent details of extracranial anatomy. Strong embodied cognition, by contrast, insists that the realization of cognitive functions cannot be fully understood without reference to the specific ways in which anatomy and physiology shape, constrain, and enable those functions. The body, in other words, is not merely a vehicle for the implementation of abstract functions, but is itself an essential part of the functional architecture that constitutes cognition.

This conclusion, however, depends on a quite restrictive interpretation of the functionalist thesis—one that treats functional roles as abstract causal mappings between inputs and outputs, independent of the specific material or dynamical properties of the system realizing them. But functionalism need not be committed to such an austere view. A more generous and biologically informed interpretation would allow us to see functional roles as inherently shaped by the organism's morphology, physiology, and mode of interaction with the world. On this broader view, functional descriptions are not fixed at a single level of abstraction but can be anchored in the richly textured organization of living systems, including their bodily structures and dynamic sensorimotor loops. When functional roles are understood as context-sensitive patterns of organism-environment coupling, rather than as isolated causal relations abstracted from embodiment, the tension between functionalism and strong embodiment largely dissolves.

It is important to recognize, however, that this *rapprochement* between strong embodiment and the classical cognitivist theses is not achieved by *fiat*, nor is it a matter of mere semantic accommodation. Rather, it requires a genuine transformation in how we conceive of the basic

categories of cognitive science. Computation, representation, and function must all be reinterpreted in light of the constitutive role played by the body and its dynamic engagement with the environment. This is not a trivial or cosmetic revision, but a substantive reorientation of the field—one that opens up new avenues for empirical investigation and theoretical reflection. The promise of strong embodied cognition, then, is not that it offers a wholesale replacement for the computational-representational paradigm, but that it provides the resources for a more adequate and comprehensive account of the mind—one that does justice to the richness, complexity, and situatedness of cognitive life.

2. EMBEDDED AND EXTENDED COGNITION

2.1. WEAK EXTENDED COGNITION, OR EMBEDDED COGNITION

Let us recall the conclusions reached thus far: if interoceptive representations cannot be abstracted into sanitized B-formatted inputs and processed independently of the body's anatomical structures, then a good case can be made that body parts beyond the brain are constitutive components of human cognition. This, in essence, defines the position of strong embodied cognition. Yet the constitution–causation distinction we have drawn upon also opens the door to a broader question: if parts of the body beyond the brain can also be constitutive of cognition, what are the outer boundaries of the cognitive system? Could elements in the environment, similarly to body parts, also figure as genuine components of cognitive activity? It is from this line of inquiry that the theoretical programs of embedded and extended cognition arise—shifting the focus from brain-centered models to more distributed, world-involving accounts of cognition.

The second and third Es—i.e., *embedded* and *extended cognition*—explore how far beyond the skull—and even beyond the skin—cognitive processes may extend. Although the literature often treats embedded and extended cognition as distinct positions (Clark 2008b; Newen, Gallagher and

de Bruin 2018; Gallagher 2017, 2023), I suggest that they are better understood as *weaker* and *stronger variants* of the *same underlying theory*, a choice whose motivations will become clear over the next few paragraphs. Both concern the role of environmental structures in cognition, but differ in the type of explanatory commitment they make: embedded cognition treats environmental elements as causally relevant, while extended cognition claims that, in some cases, they are constitutively involved. Thus, I will use here the terms "embedded cognition" and "weak extended cognition" synonymously, and reserve "strong extended cognition" or "extended cognition" *simpliciter* for theories that endorse the constitution thesis.

Discussions of embedded or extended cognition often begin with examples like maps, smartphones, calculators, or calendars—tools that offload significant portions of cognitive work onto external structures (Clark and Chalmers 1998). These artifacts function as cognitive scaffolds, reducing the mental load required to perform tasks that would otherwise be demanding or infeasible. Yet such examples, drawn mostly from technological or literate contexts, risk giving the impression that environmental extension is a rare or culturally contingent phenomenon. In fact, most proponents of these views argue the opposite: cognitive processes routinely rely on environmental structures in ways that are neither exceptional nor dependent on advanced cultural artifacts. Rather than being limited to devices of modernity, the cognitive role of the environment often emerges in ordinary, low-level interactions with the world.

Consider a simple example. Suppose a hunter-gatherer is fashioning tools from stones collected along a riverbank. As they examine each stone, they assess its properties: some are better suited for cutting, others for scraping, hammering, or drilling. Rather than retaining this classification internally, they begin to sort the stones into distinct piles according to their potential use. In doing so, they offload the cognitive labor of memory and categorization onto the environment itself. The spatial configuration now mirrors the cognitive structure of the task: future decisions are guided not by an internal map of what each stone was for, but by the layout the agent has

created through direct bodily interaction with the world. This is a clear case of embedded cognition: the environment is not merely a passive backdrop to internal computation but plays an active role in sustaining and organizing the task. Crucially, even this minimal example does not rely on language, symbolic processing, or culturally complex artifacts. The material structure of the world, as shaped through embodied engagement, becomes a *scaffold* for cognition.

At first glance, one might even be tempted to claim that the piles of stones are not merely supportive, but constitutive of the cognitive process itself. Yet this stronger reading has been the target of sustained criticism. An objection originally directed at strong embodiment can be applied to the domain of extended cognition: external structures and artifacts, no matter how intimately integrated into the task, do not genuinely constitute cognitive activity but merely influence it causally. This view holds that, however narrow the gap between external scaffolds and neurally instantiated processes, the former remains outside the cognitive system proper. Within the context of the debate on embedded and extended cognition, this line of critique is known as the *coupling–constitution fallacy* (Adams and Aizawa 2008; Rupert 2010). Accepting this objection draws a firm boundary between the mind and its environment—one that weak E theories are willing to preserve, while stronger theories explicitly seek to challenge. As Julian Kiverstein puts it:

> The embedded theory (henceforth EMT) and the family of extended theories (EXT) disagree about what it is for a process to count as cognitive. EMT holds that cognitive processes are deeply dependent on bodily interactions with the environment [...] [yet] claims that cognitive processes are nevertheless wholly realized by systems and mechanisms located inside the brain. Thus, advocates of EMT continue to interpret cognition along more or less traditional lines […] as being constituted by computational, rule-based operations carried out on internal representational structures that carry information about the world.
>
> (Kiverstein 2018, 21)

Indeed, as a member of the extended cognition family, weak extended cognition—or embedded cognition, in the sense defined above—presents a curious case. Its core claim, namely that

elements in the environment play a significant causal role in cognitive processes, is so modest that even the most resolute cognitivist rarely contests it. What distinguishes this approach is not a metaphysical stance on the constitutive boundaries of cognition, but rather a methodological shift: instead of asking what is part of the cognitive system, it focuses on how cognition is shaped, scaffolded, or facilitated by environmental structures. In this sense, weak extended cognition functions less as a unified philosophical theory than as a background framework for more domain-specific research. It provides a conceptual lens through which to study particular cognitive phenomena, without requiring commitment to any particular ontology. Examples of theories operating within this framework include Clark's (1997) work on *scaffolded minds*, Edwin Hutchins's (1995) account of navigation practices in *distributed cognition*, and perhaps certain forms of *ecological psychology* that emphasize task-dependent coupling without committing to constitutional claims (Gibson 1979; Chemero 2009).

It is also worth noting that, within the landscape of weak extended cognition, Rodney Brooks's work in *robotics* presents a peculiar case (Brooks 1991). While his approach challenges the ubiquity of internal representations in intelligent behavior, he does not appear committed to a more radical rejection of representational explanation altogether. Rather, his *subsumption architecture* demonstrates that many functions traditionally implemented through internal symbolic processing—especially in robotic engineering—can, in practice, be explained through the dynamic coupling between agent and environment. Moreover, Brooks is primarily concerned with practical problems in robotics and their consequences for cognitive architecture, not with questions in the metaphysics of mind, and therefore it would not be inaccurate to say that his work remains agnostic about the bounds of cognition.

Weak extended cognition fits squarely within the mainstream framework of classical cognitive science. It does not challenge the core tenets of computationalism, representationalism, or functionalism. Environmental structures are treated as part of the causal context in which cognition

unfolds, but not as constituents of cognitive processes themselves. Computationalism is preserved, since cognitive activity is still understood as internal information processing over symbolic or sub-symbolic states. Representationalism remains unthreatened, as external elements are not treated as representations *per se*, but rather as triggers, guides, or supports for internal representations. Functionalism, too, remains intact, because the environment's contribution is mediated through standard input–output channels—chiefly perception and action—even when feedback loops and dynamic couplings are involved. What weak extended cognition offers is not a metaphysical redefinition of the cognitive system, but a pragmatic reorientation: it emphasizes how cognition is facilitated, shaped, and constrained by environmental regularities without rethinking where cognition resides or what constitutes it.

2.2. STRONG EXTENDED COGNITION

While Clark's work in the mid-1990s can be reasonably characterized as instantiating weak extended cognition, his position later evolved toward a stronger thesis. This shift culminated in his 1998 paper with David Chalmers, "The extended mind," which laid the groundwork for what is now known as *strong extended cognition* (Clark and Chalmers 1998). Unlike embedded cognition, which focuses on how environmental structures shape or support cognitive activity, strong extended cognition claims that under certain conditions, external elements can be constitutive parts of cognitive processes themselves. This is the essence of the constitution thesis, which holds that the cognitive system is not bounded by the skull or skin, but may extend into the world when certain functional criteria are met.

This argument is structured around the *parity principle*, thus formulated by Clark and Chalmers: "If as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting it as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process" (Clark and Chalmers

1998, 8). The now-canonical illustration of this idea is the case of Otto, an Alzheimer's patient who uses a notebook to record information he can no longer store reliably in biological memory. Otto's notebook, they argue, serves the same role for him that biological memory serves for a neurotypical individual like Inga. If we are prepared to treat Inga's internal belief that the museum she wants to visit is on 53rd Street as a cognitive state, we should, by parity of reasoning, also treat Otto's reliance on his notebook as part of his cognitive system. The criterion for inclusion, in this framework, is not ontological location but functional equivalence. This reframing shifts the focus of cognitive science from where cognition occurs to how it is functionally organized, opening the door to viewing external artifacts not merely as tools or supports, but as genuine constituents of thought when they are tightly integrated into an agent's cognitive routines.

The move from embedded to extended cognition required not just a conceptual shift in how the boundaries of the mind were drawn, but also a revision of what counted as cognitive processing in the first place. This was made possible by the adoption of the framework of *wide computationalism*—the idea that computational processes can be realized across brain, body, and environment, so long as the appropriate functional organization is preserved. One of the earliest proponents of this view was Robert Wilson, who argued that cognitive systems could be partially realized in external structures without ceasing to qualify as computational (Wilson 1994; Wheeler 2005). In contrast to traditional computationalism, which typically locates information processing within a bounded physical substrate—i.e., usually the brain—, wide computationalism treats the *locus* of computation as abstract and substrate-independent. What matters is not where the computation occurs, but whether a given process plays the right causal-functional role within the larger cognitive architecture. This move allows the theorist to treat external artifacts—such as Otto's notebook—not merely as causal aids, but as physical realizers of computational states in a distributed cognitive system.

Crucially, wide computationalism enables strong extended cognition to remain fully compatible with the computationalist framework of mainstream cognitive science. The thesis is not that cognition is somehow non-computational once it extends into the world, but that computational systems need not be spatially confined. As long as the external component is reliably integrated into the agent's cognitive routines and participates in the same kind of information processing one would expect from internal systems, it qualifies as part of the cognitive system. This position maintains the central explanatory tools of cognitive science—computation and functional role—but applies them to a larger physical system. In this way, wide computationalism serves as the theoretical linchpin of strong extended cognition: it stretches the boundaries of the mind without altering its internal logic.

Although strong extended cognition expands the spatial boundaries of the cognitive system, it does not thereby abandon its foundational representational and functionalist commitments. On the contrary, Clark and Chalmers's own justification for including external resources like Otto's notebook relies heavily on the notion of functional equivalence, via the parity principle (Clark and Chalmers 1998, 8; Menary 2010, 5–7). This argument clearly presupposes a functionalist framework, in which mental states are individuated not by their physical substrate, but by their causal role in a system. Otto's notebook is cognitively relevant because it figures in the same kind of information-processing loop—prompting action, storing and retrieving content, guiding navigation—as Inga's biological memory. Far from challenging functionalism, strong extended cognition theory depends on its liberal application to justify the inclusion of non-biological elements within the cognitive system.

Likewise, the representational dimension of classical cognitive science is preserved rather than rejected. Otto's notebook is not just a behavioral trigger—it is assumed to carry content: a representation of where the museum is located (Clark and Chalmers 1998, 12–14). What makes it eligible for cognitive status is precisely its role in encoding, storing, and accessing information that

would otherwise be held in internal memory. Thus, strong extended cognition treats the environment as an external memory store, but one that functions analogously to internal representations. This continuity is evident throughout Clark's early and later writings, where even when emphasizing action and dynamic coupling, he continues to describe cognition in terms of internal models, contentful states, and information processing (Clark 2008a). The result is that strong extended cognition, for all its boundary-stretching rhetoric, retains the computational-representational machinery of traditional cognitive science—only now applied to a hybrid system that spans brain, body, and world.

In the past decade, Clark has shifted toward predictive processing, recasting cognition as a hierarchical Bayesian process in which the brain continuously updates generative models to minimize prediction error (Clark 2013, 2016, 2023). Although many interpret the predictive processing framework as fundamentally representational and computational, there is ongoing controversy in the literature over whether it truly retains all the standard assumptions of representationalism, computationalism, and functionalism. Although Clark often presents it as compatible with the extended mind thesis, especially when discussing active inference and sensorimotor loops, the framework itself does not fundamentally rely on the constitution thesis that defines strong extended cognition. Instead, what survives from Clark's earlier work is a sensitivity to context and embodiment, reframed within a model that is internally driven and representationally structured. Predictive processing illustrates a broader trend: the language and concerns of 4E cognition persist at the margins, but they are absorbed into an explanatory architecture that continues to prioritize internal computation.

Despite its apparent radicalism, strong extended cognition does not break with the core explanatory framework of mainstream cognitive science. Its central claim—that external structures can, under certain conditions, be constitutive of cognitive processes—rests on a liberal application of functionalism, made possible by the adoption of wide computationalism, and it continues to

treat cognition as a matter of information processing over content-bearing representations. Even in Clark's later work on predictive processing, where the vocabulary of action, embodiment, and environmental coupling reappears, the explanatory architecture remains internally grounded and computationally structured. What strong extended cognition offers, then, is not a revision of the metaphysics of mind, but a redefinition of its physical boundaries within an unchanged theoretical framework. It stretches the domain of cognition outward without challenging the foundational premises—computationalism, representationalism, and functionalism—that have long defined the cognitivist tradition.

3. ENACTIVE COGNITION

The theories discussed so far—embodied, embedded, and extended cognition—differ in emphasis and scope, but none of them decisively break with the central explanatory framework of cognitive science. Each, in its own way, stretches the boundaries of cognition beyond the brain, whether by highlighting the causal relevance of the body, the scaffolding role of the environment, or the potential inclusion of external tools and structures as cognitive components. Yet they remain, to a significant extent, compatible with the foundational assumptions of cognitivism: computationalism, representationalism, and functionalism. Enactivism, at first sight, appears to depart sharply from this pattern. While embodied, embedded, and extended cognition all admit versions that treat the body or environment as causally relevant without insisting on their constitutive status, enactivism does not allow for such gradation. From the outset, enactivism holds that cognition is always constituted by the dynamic coupling of brain, body, and environment (Varela, Thompson, and Rosch 1991, Chs. 8 and 9; Di Paolo, Buhrmann, and Barandiaran 2017). In this view, the cognitive system is not an internal core that sometimes extends outward—it is from the beginning a relational process, defined by its continuous, recursive interaction with the world.

This *holistic* attitude explains why enactivism resists the analytical distinctions we have employed thus far, such as causation versus constitution or weak versus strong theoretical variants. In extended cognition, for instance, Clark and Chalmers (1998, 8) emphasize that cognition only sometimes extends into the environment, under specific functional criteria; the implication is that it can also remain internal, and that the boundary between internal and external components can shift. Enactivism rejects this framework entirely: cognition is never "internal" in isolation, nor does it occasionally reach outward. It is always a process of world-involving activity, realized through embodied interaction in a domain of meaningful concern, as noted by Francisco Varela, Eleanor Rosch, and Evan Thompson: "[…] cognition is not the representation of a pregiven world by a pregiven mind but is rather the enactment of a world and a mind on the basis of a history of the variety of actions that a being in the world performs" (Varela, Thompson, and Rosch 1991, 9). Hence, there is no "weak enactivism" in which the environment merely influences cognitive processes from the outside.

It is often claimed, therefore, that enactivism is incompatible with all three cognitivist tenets: computationalism, representationalism, and functionalism. This incompatibility stems from enactivism's rejection of core presuppositions shared by these frameworks—namely, that cognition consists of internal operations performed over contentful representations, implemented through functionally individuated mechanisms. In contrast, enactivists posit that cognition arises from the organism's autonomous engagement with its environment, emphasizing sense-making and embodied interaction rather than symbol manipulation. Representationalism is rejected because cognition does not depend on constructing inner models of a mind-independent world, but instead involves "bringing forth" a meaningful world through embodied activity. Computationalism, in its standard form, becomes problematic because it presupposes pre-given inputs and syntactic rules, whereas enactivism maintains that perceptual data are enacted through structural coupling. Finally, functionalism is undermined by the enactivist commitment to embodiment and biological individuation: mental states are not merely defined by causal roles but by their integration within the living

organization of the agent. In this sense, enactivism constitutes not just a spatial redefinition of the cognitive system, but a categorical rethinking of what cognition itself is. Thus, the traditional image of cognition as rule-based manipulation—i.e., computation—of internal representations over pre-given sensory data seems to become untenable, because the data are not pre-given at all—they are enacted through the system's own history of structural coupling with its environment.

Consider what has been called the "outfielder problem" in the literature. From a classical cognitivist perspective, catching a fly ball is explained in terms of internal computation: the brain receives visual input, constructs a model of the ball's parabolic trajectory using parameters like velocity, angle, and gravity, predicts its future position, and issues motor commands accordingly. Enactivism, by contrast, offers a very different explanation. Empirical research by McBeath, Shaffer, and Kaiser (1995), often used as a case study by enactivists, suggests that players do not calculate the ball's trajectory at all. Instead, they employ a strategy known as optical acceleration cancellation, adjusting their movements so that the image of the ball follows a constant velocity in their visual field. By doing so, they remain dynamically coupled with the environment in a way that guides them to the right spot without ever computing the physical path of the ball. The act of catching, on this view, is not the result of detached internal modeling, but of ongoing sensorimotor coordination shaped by real-time perceptual feedback. Each movement—shifting posture, stepping sideways, raising an arm—is modulated on the fly, not according to a stored representation, but through embodied engagement with a structured environment. As Chemero (2009, 114–116) argues, this example illustrates how cognition can be understood as the unfolding of skillful interaction, not as internal processing over symbolic representations.

Rejecting the three main cognitivist presuppositions is not a trivial move, however. Despite the proliferation of 4E discourse in recent decades, much of empirical research in cognitive science continues to operate within frameworks that presuppose some version of representationalism, computationalism, and functionalism. These assumptions are deeply embedded not only in the

methodologies used—from modeling neural computation to designing behavioral experiments—but also in the interpretative frameworks through which data are made sense of. Moreover, cognitive architectures that have proven successful in both explanatory and engineering contexts—such as predictive coding, Bayesian inference models, and artificial neural networks—are typically conceived within this triadic framework. To discard these assumptions wholesale, then, is to sever ties with a robust and mature research program, one that has generated cumulative knowledge and continues to yield testable hypotheses. Any theory that aspires to replace it must meet a high standard of empirical adequacy and conceptual clarity.

This is precisely where enactivism faces its most serious challenge. While its philosophical motivations are compelling, its empirical tractability and methodological foundations remain contested. As Clark (1997, 2008, 2016) and others have pointed out, even if one accepts the centrality of dynamic coupling and sense-making, it is not obvious that enactivist models can offer the same explanatory precision or predictive power as computational ones. Dynamical systems theory, sensorimotor contingencies, and autopoietic models—which are important parts of the enactivist's modelling toolkit and will be discussed in more detail in the next chapters—provide promising alternatives, but they are still in the process of being fleshed out into a comprehensive framework that can rival the computational-representational paradigm across all domains of cognition. For this reason, the claims of enactivism should be subjected to careful scrutiny rather than accepted as a wholesale replacement. In the chapters that follow, I aim to evaluate whether enactivist commitments truly require rejecting these premises or whether some reconciliatory framework might be available that preserves the insights of enactivism without forfeiting the methodological strengths of mainstream cognitive science.

An additional topic of great relevance, particularly for Chapters Two and Three of this study, concern the identification of the main varieties of enactive theorizing. Although enactivism is often treated as a coherent and unified alternative to classical cognitivism, it would be a mistake

to assume that it constitutes a single, internally homogeneous theoretical stance. In reality, the enactive paradigm encompasses relatively distinct research programs, each with its own emphases, conceptual vocabulary, and explanatory ambitions. The literature most commonly distinguishes between three main strands: sensorimotor enactivism, which focuses on the role of sensorimotor contingencies in structuring perceptual experience; autopoietic enactivism, which grounds cognition in the self-producing and self-maintaining dynamics of biological systems; and radical enactivism, which is primarily characterized by its rejection of mental representation as a theoretical posit in cognitive science. At first glance, this tripartite division may suggest a typology comparable to that of other 4E theories, each variant adding further commitments or extending the explanatory reach of the core view. However, a closer examination reveals that the situation is more nuanced. Sensorimotor and autopoietic enactivism share a set of foundational assumptions that justify their grouping under the enactive label. Both are committed to the idea that cognition arises through the ongoing, dynamic coupling of organism and environment; both reject the picture of the mind as a *locus* of input-output transformations governed by internal representations.

Despite their differences—sensorimotor enactivism being more minimalist and restricted in scope, while autopoietic enactivism incorporates a broader biological grounding—these two strands can be seen as complementary attempts to specify how the enactive framework applies at different levels of analysis. Radical enactivism, by contrast, stands on different footing. Rather than offering a positive model of cognition grounded in specific biological or sensorimotor structures, it functions primarily as a critical project aimed at undermining the centrality of representation in cognitive science. Its primary contribution lies in its attempt to dismantle what Daniel Hutto and Erik Myin call the "content-involving" conception of mind, arguing that much of our cognitive activity—especially in its more basic forms—does not involve representations at all. However, radical enactivism does not articulate a substantive alternative to the mechanisms and processes it rejects, nor does it provide a coherent account of cognition as sense-making or autonomous coupling. For this reason, it is more accurate to view it not as a third variety of enactivism proper, but

as a polemical position that operates within the broader 4E discourse, offering a conceptual critique of representationalism rather than a fully worked-out enactive theory.

To navigate this internal diversity without collapsing it into a formless pluralism, we must individuate the main varieties of enactivism according to principled criteria. Two dimensions are particularly helpful in this regard: scope and presuppositions. Sensorimotor enactivism is narrower in scope, primarily concerned with perception, action, and the phenomenology of conscious experience. Autopoietic enactivism, by contrast, adopts a broader biological outlook, aiming to explain not only cognition but also the emergence of autonomous living systems. Their presuppositional commitments also differ. Sensorimotor enactivism presumes that cognitive processes—especially perception and action—are mediated by non-linear feedback loops, often instantiated in the brain and sensorimotor systems, but without thereby committing to a fully biological account of cognition. Autopoietic enactivism, by contrast, maintains that these feedback-mediated organizational patterns are not just relevant to cognition, but are constitutive of life itself. On this view, the domain of cognition is continuous with that of life: what makes a system alive—its autopoietic organization—is also what makes it capable of cognition. This is the so-called life–mind continuity thesis, a central tenet of autopoietic enactivism and one that marks a significant departure from any model that treats cognitive systems as special-purpose information processors instantiated in otherwise indifferent biological matter.

4. Conclusion

In this chapter, I sought to map the relations among the various theories grouped under the label "4E cognition," emphasizing the theoretical and historical motivations for establishing distinctions between them when necessary. This allowed for the following brief description of the 4E *crescendo*: traditional cognitivism ascribes constitutive powers exclusively to the brain—i.e., it is strictly intracranialist. Therefore, weak embodiment theorists committed to this view must sever the link

between interoception and mental content by resorting to sanitized B-formatted representations. Moreover, they can only accept the involvement of any extracranial bodily process in cognition—even beyond interoception—if it plays a merely causal role. This commitment to the cognitivist framework makes weak embodiment fully compatible with representationalism, computationalism, and functionalism. Strong embodiment theorists, by contrast, treat at least some extracranial bodily processes as constitutive of cognition; they thus conceive of the cognitive system as a holistic brain-body coupling. In principle, however, if the concepts of representation, computation, and function are understood in a broader sense, there is no fundamental incompatibility between assuming that the body plays a constitutive role in cognitive activity and endorsing the three traditional cognitivist presuppositions.

Embedded—or weak extension—theorists acknowledge the importance of extracorporeal, environmental factors in cognitive processes but regard them as merely causally relevant. There is no reason to think that this modest thesis cannot be fully accommodated within more hegemonic frameworks, which also incorporate functional-representational computationalism as a core premise. Strong extension theorists, by contrast, take extracorporeal processes to be constitutive of cognition itself and, through the parity principle, interpret cognitive processes in terms that are perhaps even more radically functional than traditional functionalism. This is why the framework is often referred to as "extended functionalism." The case for its compatibility with computationalism and representationalism is further strengthened by Clark's recent reframing of the theory in terms of predictive processing.

Enactivists likewise understand cognition in terms of brain-body-environment couplings, but their theory introduces additional presuppositions that make it significantly more difficult to reconcile with classical cognitivism. First and foremost, cognition is, for them, always action-oriented. Since actions are necessarily extended processes—entailing interaction between the body and its immediate environment—the extrabodily extension must always be constitutive of

cognition, not merely causal. For this reason, enactivism cannot be divided into weaker and stronger versions based on the causation–constitution dichotomy. Even if such a division might be tempting for the sake of parallelism, a "weaker" enactivism would be conceptually unstable. With that caveat in mind, I have attempted to map the theoretical commitments of each of these variants in the following way, inspired by—but modifying—the typification presented by Newen *et al.* (2018, 6):

TABLE 1: Causal and constitutional relations in 4E cognition

Constitution Process	Not constituted by (…); causally dependent on (…)	Partially constituted by (...)
(...) extracranial bodily processes.	WEAKLY EMBODIED	STRONGLY EMBODIED
(...) extracorporeal processes.	EMBEDDED (WEAKLY EXTENDED)	EXTENDED (STRONGLY EXTENDED)
(...) extracranial and extracorporeal processes, and dispositions to act.	~~WEAKLY ENACTIVE~~	ENACTIVE (STRONGLY ENACTIVE)

Finally, enactivism's resistance to analysis in terms of causation and constitution may give the impression that, unlike the other Es, it constitutes a unified theoretical program. This, however, is not quite accurate. The enactive tradition exhibits a certain degree of internal diversity, and scholars commonly distinguish at least three main varieties: sensorimotor, autopoietic, and radical enactivism. In the next two chapters, I will examine this diversity. I begin in Chapter II with sensorimotor enactivism, the most minimalist version in terms of both theoretical presuppositions and explanatory scope. Sensorimotor enactivism is typically presented as a tentative response to problems

in the domains of action, perception, and consciousness. Because issues related to the functional organization of information processing in cognitive systems are central to this framework, I will also take the opportunity to briefly examine its relationship to functionalist theories in traditional cognitive science. In Chapter III, we will turn to autopoietic enactivism, a broader framework that seeks to ground the feedback-mediated processes emphasized by sensorimotor enactivism in biological organization.

Chapters IV and V will address the relationship between the enactive framework and, respectively, representationalism and computationalism. I will examine whether the received view—namely, that the enactive approach is profoundly incompatible with these theses—truly holds up to scrutiny. While this incompatibility is often taken for granted in the literature, it is rarely subjected to detailed conceptual analysis. My aim will be to unpack the underlying assumptions of each position and assess whether a more nuanced reconciliation is possible.

II. SENSORIMOTOR ENACTIVISM

In the previous chapter, I argued that the distinction between causation and constitution—instrumental in differentiating strong and weak variants of embodied and extended cognition—is not applicable to enactivism. This is because, for enactivist theories, the coupling between brain, body, and environment is never merely causally relevant, but always constitutive of cognition itself, and therefore must be treated as a basic, unanalyzable *explanandum*. As a result, there is no weak form of enactivism; no variant that merely treats the environment or the body as an external influence upon cognition. This feature unifies the enactivist camp and sets it apart from the rest of the 4E spectrum.

This formal unity, however, conceals important internal differences, for enactivism is not a single, monolithic research program built upon a universally shared set of premises. Rather, it has developed along multiple lines of inquiry, and is commonly subdivided into three main variants in the literature: sensorimotor enactivism, autopoietic enactivism, and radical enactivism (see Hutto and Myin 2013; Newen, de Bruin, and Gallagher 2018; Shapiro and Spaulding 2021; Ward, Silverman, and Villalobos 2017). Radical enactivism—chiefly associated with the work of Daniel Hutto and Erik Myin (2013, 2017), but occasionally also with that of Anthony Chemero (2009)—is not a theory of cognition in the same sense as the sensorimotor or autopoietic variants. As will become clear in Chapter IV, radical enactivism lacks a concrete explanatory framework and should be understood instead as a metatheoretical critique, with the central aim of eliminating representationalist assumptions from enactivist cognitive science. For this reason, in this chapter and the following, we will focus on the first two theories—sensorimotor and autopoietic enactivism.

Sensorimotor enactivism can be seen as the "basic package" of the enactivist framework, explaining mentality through the dynamic coupling of perception, cognition, action, and

phenomenal consciousness—particularly in its sensory dimensions. Its central concept is that of sensorimotor contingencies, or law-like dependencies between motor activity and sensory change, developed by Alva Noë and J. Kevin O'Regan (Noë and O'Regan 2001; Noë 2004). This model challenges the classical computationalist view of cognition, replacing the idea of discrete inputs and symbolic processing with those of feedback loops and embodied interaction. This theory, its achievements, and its limitations are the topic of the present chapter. In contrast, the next chapter will turn to autopoietic enactivism—the "premium package," one could say—which also accepts the same presuppositions but grounds them in the biology of living systems, specifically through the theory of autopoiesis (Maturana and Varela 1980). In autopoietic enactivism, cognition is seen as continuous with life itself—an idea that is not present in the sensorimotor variety of the theory—and concepts such as biopsychism and sense-making come to the fore. While autopoietic enactivism historically predates the sensorimotor variant, I introduce them in reverse order to reflect a progression from a more modest, targeted account to one with broader and more controversial commitments.

After having presented, in §1 and §2 of this chapter, a general account of the motivations for the development of sensorimotor enactivism and the sensorimotor contingencies theory, I will show, in §3, why they fall short of their goals. Sensorimotor enactivism offers no explicit pathway toward higher-level cognition such as language, memory, or abstract reasoning. Moreover, it also remains largely silent on matters of valence, salience, or the intrinsic normativity of perceptual experience. This happens precisely because it does not engage systematically with the biological grounding of cognition. Autopoietic enactivism, on the other hand, offers a deeper and more comprehensive framework by grounding cognitive activity in the biological organization of living systems. While sensorimotor enactivism emphasizes the dynamic coupling between perception and action, it does not explain how this coupling becomes meaningful or normatively structured for the organism.

This is addressed by autopoietic enactivists by introducing the notion of adaptivity: the capacity of a living system to regulate its internal processes and interactions with the environment in ways that sustain its own viability. This concept is central to understanding cognition as inherently normative and affect-laden, tied to the organism's need to persist as a self-producing system. My conclusion will be, therefore, that although sensorimotor enactivism captures important elements of embodied experience, it lacks the depth required to present a real challenge to mainstream cognitive science. A more adequate alternative is found in its autopoietic counterpart, to which we will turn from the next chapter onwards.

1. THE INPUT-OUTPUT MODEL

Sensorimotor enactivism opposes the "input-output model" or "classical sandwich model," a formulation originally coined and critically examined by Susan Hurley, whose work closely parallels key themes later developed by Noë and O'Regan. Hurley was not the first to challenge this model, and in fact—as she herself recognizes—a significant corpus of contemporary critical literature on the input-output model already existed when she finished writing the essays that make up *Consciousness in action* in the late 1990s (Hurley 1998, 413–414). Still, most of those works focused on offering a holistic account of action and perception applied to specific topics and problems, whereas hers was probably one of the most significant attempts at criticizing it systemically from a broader and strictly philosophical standpoint. Hurley describes the input-output model in the following manner:

> A mainstream view of the mind has two main components. The first is a view of perception and action as separate from each other and as peripheral. The second is a view of thought or cognition as the central core of the mind, at least for creatures with cognitive abilities. Cognition is virtually central, even if the mere implementation of cognitive processes is distributed. The mind decomposes vertically into modules: cognition interfaces between perception and action. Perception and

> action are not just separate from each other; but also separate from the higher processes of cognition. The mind is a kind of sandwich, and cognition is the filling.
>
> (Hurley 1998, 401)

Hurley's main target is the then-dominant view of the relationship between action and perception in analytic philosophy of mind and cognitive science, according to which these are distinct capacities whose interaction is merely *causal*, *linear*, and mostly *transitive*. It is merely causal because the different modules are not properly constitutive of each other, as it is theoretically possible to offer a functional description of the workings of one of them without having to describe the others.[1] It is linear because the flux of information follows a consistent path, originating from perception, traversing cognition, and culminating in action. And, finally, it is mostly transitive because each one of these modules receives all of the relevant information for its proper operation from the one that precedes it, with a few possible exceptions.

Even though there are good reasons to suppose that what Hurley has in mind in this specific passage is Jerry Fodor's modularity theory (1983), the views she presents are neither exclusive to a specific theoretical framework, nor to a particular author, but rather widely spread among a majority of philosophers and cognitive scientists for the most part of the 20th century, computationalist or not. In fact, a consensus around the most important tenets of the input-output model has been gradually built by following different steps since at least the 18th century, as I will attempt to show now. This historical detour will, hopefully, help us understand how this view came to be considered so intuitive by contemporary scholars occupying such different positions within the theoretical spectrum.

[1] This framing, while simplified, roughly captures the spirit of many functionalist accounts, even though causal roles themselves are often treated as constitutive of mental states—an idea that puts pressure on the causation-constitution distinction as we have been using it up until now. I return to this issue later in this section.

Attempts to divide the mind into its principal functional components date back at least to Plato, who posited a tripartite soul composed of reason, spirit, and appetite or desire. The Platonic framework—like those of many of his successors—may strike contemporary readers as peculiar, since the functions attributed to each part do not align neatly with either modern scientific models or folk-psychological understandings of the mind. A more familiar outline only began to emerge during the Early Modern period, particularly in the works of thinkers such as Descartes (1641, AT VII 57), Hobbes (1651/1996, I.6), and Hume (1739–1740, T II.3.iii; SB 413–18). Interestingly, the original motivation for such divisions was not always to provide a metaphysical theory of the mind. Rather, many Early Modern philosophers were concerned with questions we now associate with the philosophy of action: What motivates human behavior? How are desires and motives related? What role does reason play in producing action? That this preoccupation with agency and behavior gave rise to a theory of mental architecture becomes particularly clear in the case of Hume, whose theory of motivation left a profound legacy in philosophy and psychology.

When explaining purposeful human behavior, Hume distinguished between two types of mental states: *beliefs* and *desires*. For him, reason alone cannot motivate us to act. In other words, our beliefs about what is true or false do not by themselves lead us to action; rather, it is our desires that drive us to pursue certain goals and engage in specific, goal-oriented behaviors. Beliefs are our judgments about the way things are, and they are therefore formed by our perception of states of affairs in the external world, while desires are our subjective feelings of approval or disapproval towards these states of affairs, and usually result in us acting in specific ways in order to preserve or change them. A desire for something generates the necessary motivation to act in order to attain that desired object or outcome.

For a very simple example, suppose someone has come to form the pair of beliefs that i) water quenches their thirst and that ii) it is usually available from the kitchen tap. Now, if that person suddenly sees themselves feeling a strong enough desire to drink water—that is, if they

suddenly feel thirsty—one would expect their subsequent behavior to be that of opening the tap in order to get a glass of water. The belief alone that water is usually available from the tap does not suffice to prompt that person to do so; it is also necessary that they feel the desire to drink it.

At first glance, *belief-desire psychology* may seem trivial. Its apparent obviousness, however, should not be seen as a weakness, but rather as a strength: Hume's account seems self-evident precisely because the inferential patterns it describes are part of an innate human faculty, one that enables us to interpret both our own and others' behaviors and motivations. In this sense, belief-desire psychology functions as an idealization of our *commonsense* or *folk psychology*, and often a highly effective and reliable one. From the observation of someone turning on a tap and pouring a glass of water, we are naturally inclined to infer that the person had i) the belief that water quenches thirst and that it is available from the tap, and ii) the desire to quench their thirst. We typically assume that their behavior results from the interaction between those beliefs and that desire. Whether folk psychology can serve as a foundation for a properly scientific theory of mind remains a topic of intense debate—a question I will not address here. Nevertheless, its predictive success and practical utility in everyday life make it at least a candidate worth considering.

The model describes a system governed by a linear and unidirectional flow of information. Information about the world is first received through perception. It is then processed by a distinct rational or cognitive module, segregated from perceptual mechanisms, where it leads to the formation of beliefs. If accompanied by a corresponding desire, this belief results in purposeful behavioral output—that is, action informed by the belief-desire pair. According to this model, the mind comprises at least three distinct faculties—perception, cognition, and action—each dependent on the output of the one preceding it. This cascade-like structure also aligns with a well-known distinction in the philosophy of mind and language: the *direction of fit*. Cognitive attitudes such as beliefs have a mind-to-world direction of fit, since they are shaped by facts perceived in the external world. Desires, by contrast, exhibit a world-to-mind direction of fit, as they aim to bring the world

into alignment with the agent's internal states. Hurley captures this process in the following manner:

> The representations produced by the various streams of input processing converge and are combined by perception. The unified result is sent on to cognition, the central module that interfaces between perception and action. This is where the processes occur on which rational thought and deliberation depend. Rationality is conceived as depending on internal procedures involving the manipulation of internal representations, including those passed on by perception. Based on current and stored input and cognitive processing, a motor plan is formulated, and it is passed on to motor-programming processes for execution. There is a linear sequence of separate processing stages, from perception to cognition to action.
>
> (Hurley 1998, 407)

Interestingly, since its ultimate goal is belief-formation, the end result of perceptual processes must be different in nature from perception itself. After all, beliefs have some properties that percepts do not. Notably, beliefs are i) mediated by inference and background knowledge, which makes them ii) susceptible to rational evaluation, as we may assess them for coherence, consistency, and evidential support. Additionally, and, perhaps, most importantly, beliefs usually have iii) satisfaction or truth conditions, which allows one to assign them either truth or falsehood, something that is entirely absent from the perceptual domain. Something similar could be said about the difference between beliefs and the behavioral outputs they help generate—purpose-oriented actions driven by motivational states. Beliefs interact with each other, therefore, within a different mental domain than that where either perception or action takes place. This is the central, cognitive core of the mind mentioned by Hurley in the quotation at the beginning of this section. The peculiar properties displayed by the contents of the cognitive domain are what led Hurley to add, to her initial characterization of the "mainstream view of the mind," also the following clauses:

> A fully orthodox view has one further feature. Not only is cognition central and distinct from peripheral sensorimotor processes, but the center is classical "at the right level of description." A cluster of related properties of cognition—compositionality, systematicity, productivity, binding, and so on—are to be explained classically, in terms of processes involving symbols and recombinant syntactic structure. The subpersonal processes that explain the conceptual structure of thought mirror that structure syntactically. There is an isomorphism between contents and vehicles, or causal systematicity (see Davies 1991[…]; Fodor and Pylyshyn 1988, etc.). The mental sandwich has a classical filling.
>
> (Hurley 1998, 401–402)

Here, Hurley is also assigning to the mainstream view a second thesis famously put forth by Fodor: the idea that cognition must occur by means of a vehicle with certain syntactic and semantic properties that are typically found in natural languages, such as "[...] compositionality, systematicity, productivity, [and] binding[.]" This is a clear reference to the *language of thought hypothesis*. Within this approach, the need for the "classical sandwich filling" can be understood from the way in which the proper cognitive "central core" of the mind is assigned the task of mediating between perception and action. This is where the actual computation takes place, as conceived in the classical computationalist sense, that is, as symbolic manipulation. The "sandwich filling" is nothing more than the computational core where the conversion of what is perceived or sensed into discrete and syntactically manipulable symbols takes place.

A revised version of the Humean theory of motivation gained substantial influence throughout the 20th century, in part because it fits well with the tendency in analytic philosophy of mind to model mental states in propositional terms. In many frameworks, beliefs and desires are treated as distinct types of propositional attitudes—a term first coined by Bertrand Russell (1918, 227), though the core idea can be traced back to Frege. Propositional attitudes, as the name suggests, are simply mental states directed at specific propositions. For example, consider the proposition "Emmanuel Macron is the president of France." One might fear that this is the case, hope so, intend to bring it about, feel grateful that it is the case, etc. While this diversity of attitudes is

undeniable, it has often been argued that many of them are variations on, or reducible to, the prototypical pair "to believe" and "to desire." Since this reduction has already been extensively discussed in the literature (e.g., Searle 1983, 29–35), I will not attempt to reconstruct it here.

Representing mental content propositionally allows us to account for several allegedly central features of mental states, namely: i) representation, ii) truth-conditionality, and iii) inferentiality. Regarding i), propositions offer a structured and flexible way of encoding the content of beliefs and desires, enabling us to represent states of affairs with precision—highlighting relevant aspects while omitting extraneous details. As for ii), the fact that propositions can be assigned truth values allows us to evaluate the truth or falsity of a belief, as well as to determine the satisfaction conditions of a desire. For instance, if I desire that Macron become the president of France and this eventually comes to pass, we may say that my desire has been fulfilled. Finally, with respect to iii), the ability to assign truth values to propositions enables inferential operations between them: if I believe that Macron is the president of France, and I also believe that France has only one president, I can reasonably infer that Marine Le Pen is not the president of France. This propositional framework thereby supports a depiction of the central cognitive module as possessing what Hurley calls the "classic filling"—that is, the requisite syntactic and semantic features for computational processes. According to Hurley, such properties are essential for any proper computationalist account of mental activity, and they complement the classical modular architecture of the mind.

One of the key features that sets central cognition apart from the peripheral modules responsible for perception and action is its capacity to accommodate normatively constrained content—a capacity typically captured by endorsing a propositional theory of mental content. According to Hurley, however, the classical input-output model rests on a fundamental confusion between two distinct explanatory levels: the personal level, where normatively constrained contents such as beliefs reside, and the subpersonal level, which concerns the causal mechanisms that underlie mental processes. In this framework, perception and action belong to the personal level, whereas input

and output processes are located at the subpersonal level. To address this confusion, Hurley proposes an alternative: the two-level interdependence view (1998, 6–12, 85–88, Chs. 9 and 10). At the subpersonal level, she highlights the presence of not merely one-way causal processes flowing from input to output, but a complex network of feedback loops that operate in the reverse direction as well—from output to input. These feedback mechanisms can be internal, confined to the central nervous system, or extend beyond the organism, involving proprioceptive and visual feedback generated through bodily movement. The mind, on this account, emerges from a rich web of sensorimotor feedback loops—loops that are centered on the organism but stretch outward, incorporating its interaction with the environment.

Although processes at the subpersonal level enable perception and action, Hurley contends that the personal and subpersonal levels cannot be neatly mapped onto one another (2008, Ch. 10). Changes in motor output can influence perceptual content, just as variations in sensory input can affect motivational states such as desires. In this sense, perception and action are deeply interdependent. Crucially, however, the point is not merely that perceptions and motivations causally affect one another—an idea that would still be compatible with the traditional input-output model. Rather, Hurley argues that their interdependence is constitutive, not merely instrumental. The contents of both perception and action are constituted through ongoing interactions between inputs and outputs; they are functions of the structured relationships between them, albeit in different ways. This view fundamentally challenges the notion of the mind as something internally bounded and separate from the world. Instead of imagining cognition as retreating inward, Hurley emphasizes that the mind emerges "out in the open" (1998, 3)—as an integrated aspect of a dynamic feedback system that spans both the organism and its environment.

There is a sense, of course, in which traditional functionalism also treats inputs and outputs as both causally and constitutively relevant for cognition. In most of its standard formulations (e.g., Putnam 1967a, 1967b; Lewis 1972; Fodor 1983; Block 1980), functionalism holds that mental states

are individuated by their causal roles—that is, by the relations they bear to sensory inputs, behavioral outputs, and other internal states. According to David Lewis's (1972) version in particular, a mental state is defined by occupying a specific causal role specified by a psychological theory, often a folk-psychological one, where each state is picked out by its place in a network of interrelated causes and effects. As Ned Block (1980) emphasizes, this makes mental states fully individuated by their systemic role and counterfactual patterns of input-output transitions, regardless of physical realization. Thus, these causal roles are taken to be constitutive of mental states: a belief, for example, just is the state that tends to be caused by certain inputs—e.g., the perception of rain—, interacts with other states—e.g., the desire to stay dry—, and causes certain outputs—e.g., carrying an umbrella.

Importantly, however, even though functionalism allows for realization in diverse substrates, it typically presupposes a clear boundary between the system and its environment—inputs and outputs serve to fix the roles of internal states, but are not themselves constitutive of mentality (Block 1980). Against this background, Hurley's conception of cognition as emerging from continuous sensorimotor feedback constitutes a systemic and metaphysical challenge to the functionalist framework. The argument is not merely that functionalist models ignore feedback loops, but that the very act of individuating internal states by their roles becomes untenable in systems where perception and action are dynamically and constitutively entangled. In such systems, the functional roles themselves are not fixed or isolable: the state of the system at any moment depends on ongoing interaction with the environment, and on reciprocal modulation between sensory and motor processes. The distinction between input and output is no longer well-defined, since outputs recursively influence subsequent inputs in real time.

Moreover, in Hurley's account, the contents of perception and motivation—typically assigned to internal representational states in functionalist theories—are not pre-given or internally encoded, but constituted by the structure of sensorimotor coupling. This blurs the very boundary

between what is "in the system" and what is "outside" it. Functionalism, even in its more ecumenical forms, cannot easily accommodate this picture without abandoning its core commitment to internal-state individuation through abstract causal role. Hurley's view thus reframes the unit of analysis: cognition is not a mapping from input to output mediated by representational states, but a dynamical pattern of activity that unfolds across brain, body, and world in a temporally extended, feedback-sensitive loop.

However, Hurley does not offer a single, unified alternative to replace the input-output model wholesale—nor would such a move be consistent with the thrust of her critique. Given that the sensorimotor feedback loops underlying various cognitive functions exhibit distinct patterns of connectivity, any attempt to provide a general, armchair description of their structure would be misguided. Instead, she draws on a wide range of empirical findings and theoretical arguments from across disciplines, each of which challenges core assumptions of the input-output model in its own way. Her aim is to expose the non-generalizable, function-specific nature of these processes and to show how the vertical, modular conception of subpersonal causal architecture has long been under pressure from developments in multiple branches of cognitive science and related fields.

This view of cognition as deeply intertwined with sensorimotor activity and environmental embeddedness forms the conceptual groundwork for Noë and O'Regan's sensorimotor account. Hurley herself, along with Noë, would later develop this line of thought in more depth by exploring the implications of neural plasticity for consciousness—arguing that plasticity alone is not explanatory unless structured by lawful sensorimotor dependencies (Hurley and Noë 2003). In what follows, I will turn to Noë and O'Regan's theory of sensorimotor contingencies. In presenting their view, I will incorporate supporting empirical and neuroanatomical evidence on an *ad hoc* basis, thereby shedding further light on the feedback-loop model of perception and action that Hurley champions.

2. SENSORIMOTOR CONTINGENCIES

Let us summarize what has been concluded so far. The input-output model treats the relationship between perception and action as merely instrumental. Perception enables the mind to form beliefs about the external world—including, notably, the body. These beliefs are seen as attitudes toward propositions that represent states of affairs precisely and reliably, possessing the semantic properties required to bear truth. Among possible attitudes, desire is especially important for explaining action, as it has the opposite direction of fit: world-to-mind, as opposed to belief's mind-to-world. Thus, perception leads to belief, and desire motivates actions aimed at realizing what is desired. This establishes a linear, unidirectional causal chain from perception to action, taken to explain much of human behavior and cognition. While changing one's position may affect the perspective from which the environment is perceived, this influence is limited and not considered constitutive of perception.

The traditional view holds that perception is fundamentally unidirectional and instrumental, aimed primarily at forming beliefs about the external world through sensory input and cognitive interpretation. While action may influence perceptual content in certain situations, it does not alter perception's basic structure as a belief-forming, sensory-driven process. Sensorimotor enactivists reject this account, seeing action and perception as intimately connected and mutually constitutive within continuous feedback loops—so that understanding one requires understanding the other. From this perspective, sensory input alone is insufficient; motor activity must also be considered. The traditional view, by contrast, treats motor functions as mere outputs. This reflects a broader pattern, echoing the brain-body-environment couplings discussed in the previous chapter, though now at a finer level of analysis. The underlying assumption is that perception depends on the physical integration of sensory organs with the musculoskeletal and peripheral nervous systems that execute the movements needed for perception to function properly.

The most influential account of this coupling is currently that of Noë and O'Regan's (2001) theory of sensorimotor contingencies, which will be the focus of this section. The theory offers an enactive account of perception, action, and phenomenal consciousness. They argue that sensorimotor contingencies explain not only the coupling between perception and action but also two central features of phenomenal experience: *intermodal* and *intramodal* qualitative variation. Their aim is that, by accounting for these variations, the theory may help address broader philosophical problems of consciousness. The first challenge concerns differences between perceptual modalities—for example, we do not confuse visual experiences with auditory ones, except in rare cases like synesthesia. We may call something seen red or round, but applying those terms to sounds is clearly metaphorical. This is known as intermodal variation. The second challenge involves differences within a single modality—such as the contrast between perceiving red versus green—which is called intramodal variation, as it occurs within the same sensory channel.

Noë and O'Regan introduce what they call the "puzzle of visual experience," which concerns the origins of visual phenomenal consciousness. Although they focus on vision, they acknowledge the problem generalizes to other sensory modalities. They argue that most attempts to explain visual phenomenology rely on identifying neural correlates—a common strategy in contemporary consciousness research (Noë and O'Regan 2001, 939–940). For example, the brain contains cortical maps where visual information is organized retinotopically, such that neural activity mirrors the spatial layout of stimuli on the retina. This suggests that much of the visual world is neurally encoded, making these cortical areas likely candidates for visual perception's *locus*. Yet Noë and O'Regan are skeptical: "[…] The presence of these maps and the retinotopic nature of their organization cannot in itself explain the metric quality of visual phenomenology. Nor can it explain why activation of cortical maps should produce visual experience" (Noë and O'Regan 2001, 939).

In short, identifying retinotopically organized regions in the visual cortex does little to clarify the hard problem of visual phenomenal consciousness. Consequently, several supplementary

hypotheses have been proposed to enrich this account. These include a commentary system "situated somewhere in the fronto-limbic complex (taken to include the prefrontal cortex, insula and claustrum […])" (Weiskrantz 1997, 226); correlations between consciousness and "coherent oscillations in the 40–70 Hz range" (Crick and Koch 1990; Llinas and Ribary 1993; Singer 1993; Singer and Gray 1995); Penrose and Hameroff's (1994) quantum microtubule hypothesis; and Edelman's (1989) theory of reentrant signaling between cortical maps. Yet, according to Noë and O'Regan, all these views share the same flaw: they merely shift the explanatory burden—"pushing" the problem of consciousness to a deeper level instead of resolving it:

> For even if one particular mechanism—for example, coherent oscillations in a particular brain area—were proven to correlate perfectly with behavioral measures of consciousness, the problem of consciousness would simply be pushed back into a deeper hiding place: the question would now become, why and how should coherent oscillations ever generate consciousness?
>
> (Noë and O'Regan 2001, 940)

Regarding intermodal and intramodal qualitative diversity—that is, the variation in phenomenal character both across and within modalities—Noë and O'Regan argue that the current debate remains at a stalemate. They claim that standard accounts still rely heavily on Johannes Peter Müller's (1838) notion of *specific nerve energy*, which they paraphrase as follows: "Müller's idea, in its modern form, amounts to the claim that what determines the particularly visual aspect of visual sensations is the fact that these sensations are transmitted by specific nerve pathways (namely, those originating in the retina and not in the cochlea) that project to particular cerebral regions (essentially, cortical area V1)" (Noë and O'Regan 2001, 940). Yet, for several reasons, they argue, this explanation is inadequate:

> But what is it about these pathways that generates the different sensations? Surely the choice of a particular subset of neurons or particular cortical regions cannot, *in itself*, explain why we attribute visual rather than auditory qualities to this influx. We could suppose that the neurons involved are of a different kind, with, say, different neurotransmitters, but then why and how do different neurotransmitters give rise to different experiences? We could say that the type of calculation done in the different cortical areas is different, but then we must ask, how could calculations ever give rise to experience? The hard work is left undone. Much still needs to be explained.
>
> (Noë and O'Regan 2001, 940)

Noë and O'Regan believe that their approach is distinct from the others precisely because they do not treat perception as a phenomenon that occurs exclusively in the brain, and this makes the allegedly fruitless quest for neural correlates unnecessary. I will now attempt to explain what their own positive account consists in. Its starting point is the critique of a certain visual bias generally found in accounts of perception. Noë claims that, due to the relatively high degree of visual development in humans, in comparison to the other perceptual modalities, we are often inclined to consider vision as the paradigmatic case of perception in general. This leads us to generalize the characteristics that we deem most strongly distinctive of our visual experience to our perceptual faculties as a whole. Some of these perceived characteristics are actually misattributed to vision, as we will soon see, since they describe visual perception as following a photographic model, where the visual apparatus plays the passive role of a camera that receives sensory input:

> We tend, when thinking about perception, to make vision […] our paradigm, and we tend to think of vision on a photographic model. You open your eyes and you are given, at once, a sharply focused impression of the present world in all its detail. On this view, the relation between moving and perceiving is only instrumental. It is like the relation between the lugging around of a camera and the resulting picture.
>
> (Noë 2004, 2)

This paradigm is perfectly illustrated by Ernst Mach's (1886, 15) attempt at depicting his own visual field while lying down on a divan at his office. To the left side of the image and downwards, the borders of the visual world gradually blur until the objects depicted—books on a shelf, the philosopher's clothes, and the floor's wooden planks—completely disappear from sight. On the right side and upwards, the perceiver's supraorbital ridge, nose, and mustache frame the visual world. Still, everything that is within the large central area of Mach's field of vision is presented sharply and as fully accessible:

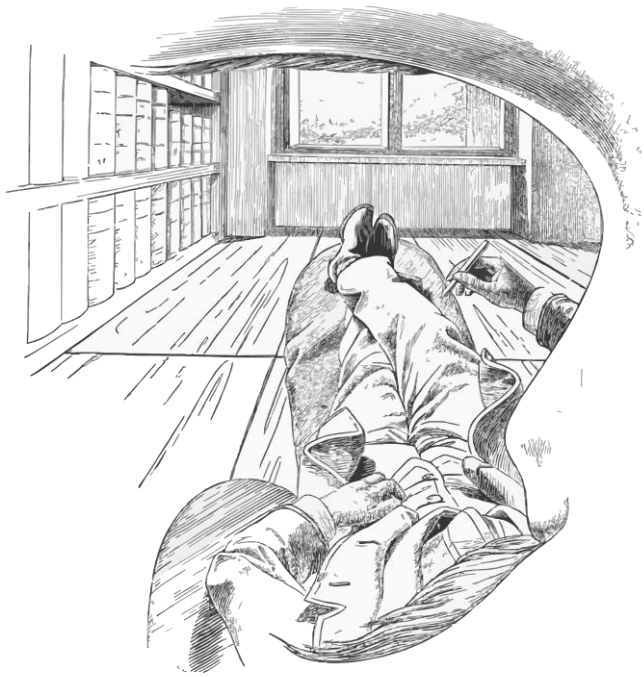

FIGURE 1 – Mach's depiction of the visual field (1886, 15)

According to Noë, the common depiction of visual experience as static and photograph-like is misleading, even if intuitive. His critique aligns with Dennett's attack on the Cartesian theater—the idea that perception is a centralized, passive process in which conscious experiences are presented to an internal observer (Dennett 1991). Dennett's multiple drafts model rejects this view, portraying phenomenal richness as illusory and shaped by attention and memory processes that "edit" experience in real time and retrospectively. Similarly, Noë and O'Regan reject the notion of

vision as passive and fixed, but they go further by emphasizing the active, exploratory role of bodily movements, sensory interactions, and environmental context. They portray vision as an ongoing interaction with the environment, structured by the sensorimotor contingencies of our perceptual systems. In this light, Noë suggests that touch—our most obviously exploratory sense—may serve as a better paradigm for perception:

> […] Perception is not something that happens to us, or in us. It is something we do. Think of a blind person tap-tapping his or her way around a cluttered space, perceiving that space by touch, not all at once, but through time, by skillful probing and movement. This is, or at least ought to be, our paradigm of what perceiving is. The world makes itself available to the perceiver through physical movement and interaction. In this book I argue that all perception is touch-like in this way: Perceptual experience acquires content thanks to our possession of bodily skills. *What we perceive* is determined by what we are *ready* to do. In ways I try to make precise, we *enact* our perceptual experience; we act it out.
>
> (Noë 2004, 1)

The most important tool in this alternative way of understanding perception proposed by Noë and O'Regan is the notion of *sensorimotor contingencies*, occasionally also called *sensorimotor dependencies*. These were first introduced in their influential article "A Sensorimotor Account of Vision and Visual Consciousness," published in 2001. Sensorimotor contingencies refer to the inherent relationship between sensory input and motor actions in an organism's perception and interaction with the environment, wherein the perception of an external event is intricately tied to the specific bodily movements or actions exerted in response to that event, consequently shaping the ongoing sensory experience and modulating subsequent motor responses.

This ultimately constitutes a tightly intertwined cycle of perceptual and motoric processes, which emphasizes the bidirectional nature of perception and action. These contingencies entail the systematic coupling between an organism's sensory receptors, which register and transduce

external stimuli into neural signals, and its effector systems, responsible for generating motor commands that induce appropriate bodily movements, whereby the sensory consequences of these self-generated actions not only guide and fine-tune subsequent motor actions but also contribute to the perceptual interpretation of incoming sensory signals. The integration of bodily engagement, perceptual processing, and active exploration within the fabric of sensorimotor interactions enables organisms to engage with and comprehend their environment through an intricate interplay of perceptual expectations, motor control processes, and real-time sensorimotor feedback mechanisms. According to Noë and O'Regan, the specific phenomenal qualities associated with each perceptual modality emerge out of the lawful coordination between the system's afferent and efferent channels. There is, they claim, nothing intrinsic to the hardware itself that determines its function:

> From the point of view of the brain, there is nothing that in itself differentiates nervous influx coming from retinal, haptic, proprioceptive, olfactory, and other senses, and there is nothing to discriminate motor neurons that are connected to extraocular muscles, skeletal muscles, or any other structures. Even if the size, the shape, the firing patterns, or the places where the neurons are localized in the cortex differ, this does not in itself confer them with any particular visual, olfactory, motor, or other perceptual quality. On the other hand, what does differentiate vision from, say, audition or touch, is the structure of the rules governing the sensory changes produced by various motor actions, that is, what we call the sensorimotor contingencies governing visual exploration. Because the sensorimotor contingencies within different sensory domains (vision, audition, smell, etc.) are subject to different (in)variance properties, the structure of the rules that govern perception in these different modalities will be different in each modality.
>
> (Noë and O'Regan 2001, 941)

For Noë and O'Regan, in the case of vision, there are two types of sensorimotor contingencies: "[...] those that are [i)] specific to the visual apparatus and [ii)] those that are specific to the way objects occupy three-dimensional space and present themselves to the eye" (Noë and O'Regan 2001, 946). Presumably, both types also apply to the other modalities. We could connect this topic

to those discussed in our previous chapter by conceiving of these two types of sensorimotor contingencies as being i) embodied and ii) extended. The first type is embodied because it is dependent upon contingent anatomical features of our perceptual organs; the second, extended, since it depends on contingent features of the type of external stimulus received and processed by the modality in question. Sensorimotor contingencies of the first type—i.e., those that are embodied and, therefore, determined by the characteristics of the visual apparatus itself—are independent of any categorization or interpretation of objects and can therefore be considered a fundamental aspect underlying visual sensations.

One example of this type of sensorimotor contingency would be the way the vestibulo-ocular reflex compensates for someone moving their head to the left by directing their eyes to the right, in order to maintain gaze fixed on a specific object or region of the visual world. Another example is the way the specific distribution and ratio of rod cells and cones on different parts of the surface of our retinas structure our visual world, increasing visual acuity in the foveal area and decreasing it on the periphery of our visual field, allowing us to distinguish colors more precisely at the center, as well as affecting our ability to perceive colors or movement under dim light conditions, among various other parameters.

Accepting the traditional distinction between *sensation* and *perception*, Noë and O'Regan relate the first type of sensorimotor contingency to visual sensation. Sensorimotor contingencies of the second type, on the other hand—those that relate to the attributes of visual objects themselves—are the basis of visual perception. This makes them particularly interesting for Noë and O'Regan's account, since they are what actually distinguishes the phenomenal qualities of different modalities (Noë and O'Regan 2001, 943). These are distinguished by the different ways in which they allow the perceiving subject to explore the perceived environment and the objects it is composed of. Physical objects have physical properties, after all, such as size, shape, texture—and they

occupy space in the physical world that is at a certain distance and angle from the space occupied by the perceiver.

Visual sensorimotor contingencies of the second type are, therefore, the rules that account for the way exploratory engagement with visual objects take place ordinarily. An exhaustive classification and description of them would not only be very lengthy and out of the scope of this study but, also, at the present moment, impossible, since it would depend on empirical studies that were still not conducted. Therefore, I will present here only a brief, open-ended list of sensorimotor contingencies of this second type for the sole purpose of illustrating the argument. With that caveat in mind, some sensorimotor contingencies are related to i) the detection of shapes and contours of visual objects, involving the correction of distortions that may result from the position occupied by the distal source of the stimulus at a specific distance and angle in relation to the perceiver's eyes. Others are related to ii) depth perception and, therefore, also to the tridimensionality of visual objects: they imply a correction of judgments of size that is also based on the distance and position occupied by the object in the visual world. Similarly, there are sensorimotor contingencies that structure iii) color perception and iv) texture or surface perception, and they are profoundly related to lighting and background conditions in the observed environment. These sensorimotor contingencies, according to Noë and O'Regan, are forms of tacit procedural knowledge that are modality-specific and, therefore, cannot be generalized to other forms of perception. In Noë and O'Regan's words:

> What characterizes the visual mode of sampling object properties are such facts as that the retinal image of an object only provides a view of the front of an object, and that when we move around it, parts appear and disappear from view; and that we can only apprehend an object from a definite distance, so that its retinal projection has a certain size that depends on distance. Other characteristics of visual exploration of objects derive from the fact that the color and brightness of the light reflected from an object change in lawful ways as the object or the light source or the observer move around, or as the characteristics of the ambient light change.

> On the other hand, tactile exploration of an object, even though it may be sampling the same objective properties, obeys different sensorimotor contingencies: you do not touch an object from a "point of view"—your hand can often encompass it more or less completely for example, and you don't apprehend it from different distances; its tactile aspect does not change with lighting conditions.
>
> (Noë and O'Regan 2001, 942)

In attempting to ground the theory of sensorimotor contingencies on empirical evidence, Noë and O'Regan often resort to Paul Bach-y-Rita's experiments on sensory substitution (1969), which were originally conducted in the late 1960s. Sensory substitution refers to the process of using one sensory modality to gain information that would typically be acquired through another modality. It involves converting sensory input from one modality into a different modality to compensate for the loss or absence of a specific sensory channel, allowing individuals to perceive and interpret the world using alternative senses. In the groundbreaking experiments conducted by Bach-y-Rita, sensory substitution was employed to explore the remarkable plasticity of the human brain and its ability to adapt to novel sensory inputs.

One notable experiment involved a blind man who was equipped with a device that transformed luminous input into haptic stimuli on his stomach. This apparatus consisted of a camera that captured visual information and converted it into patterns of vibrations delivered through an array of vibrators placed on his skin. After a period of adaptation, the blind man was eventually able to "see" the environment by interpreting the patterns of vibrations. This experiment aligns closely with Noë and O'Regan's notion of sensorimotor contingencies, as the sensory substitution device is claimed to have created a new sensorimotor contingency by establishing a lawful relationship between the visual information captured by the camera and the tactile sensations experienced on the subject's stomach. The vibratory patterns became the sensory consequences of the subject's actions, enabling him to explore and interact with the environment. By actively moving his body

and focusing his attention, he could extract meaningful information from the tactile input, effectively substituting his lack of vision for a new perceptual modality.

This interpretation resonates with the insights developed by Hurley and Noë (2003) in their joint work on neural plasticity and consciousness, where they argue that the functional organization of perceptual systems is not fixed by anatomical structures alone but is dynamically shaped by the agent's sensorimotor interactions with the environment. Drawing from empirical cases of sensory substitution, such as Bach-y-Rita's experiments, they propose a distinction between cortical dominance and cortical deference—concepts meant to capture how neural regions either retain or relinquish their functional role depending on the behavioral context and sensorimotor engagement. Their central thesis is that the neural substrate becomes meaningful only through patterns of active sensorimotor coordination, and that consciousness itself is constituted by these dynamic loops, rather than by any localized brain state. This view reinforces the sensorimotor theory by emphasizing that what matters for perception is not just the neural activation, but the agent's practical mastery of the relevant contingencies. It also suggests that the boundaries between different sensory modalities, and between sensation and perception, are far more plastic and context-sensitive than traditionally assumed, supporting a framework in which perceptual content is not pre-given but enacted.

Let us now stop again to summarize what has been concluded so far. Defended by Noë and O'Regan, the theory of sensorimotor contingencies sets out to challenge the traditional input-output model by advancing an enactive account of perception, action, and phenomenal consciousness. At its core lies the claim that perception is not a passive reception of sensory inputs but an active, embodied process constituted through patterns of sensorimotor engagement. This framework succeeds in capturing the dynamic interdependence between sensory and motor systems, particularly in the domain of visual and tactile experience, and it offers an elegant alternative to representationalist accounts of perceptual consciousness. Yet, for all its ingenuity, the theory

reveals itself to be limited in both scope and ambition. Its applicability to non-visual modalities remains unclear (see Young 2017), and its silence on key aspects of cognition—such as normativity, conceptual content, and higher-order capacities—raises serious doubts about its viability as a comprehensive alternative to mainstream cognitive science.

3. CRITICISM

Although sensorimotor enactivism undeniably represents a significant alternative to traditional theories of perception, it encounters serious difficulties when evaluated as a broader explanatory framework in the domains of psychology and the cognitive sciences. In what follows, I will present a series of criticisms that have been raised in the literature—as well as some that I believe have not yet received sufficient attention—all of which point to important limitations of the sensorimotor model. They can be grouped under the following headings: A) the problematic role of the notion of *knowledge or mastery of sensorimotor contingencies* within the theory; B) the lack of an account of the *development of sensorimotor contingencies*; C) the *lack of individuation criteria for perceptual systems*; D) sensorimotor enactivism's *limited explanatory scope in view of its theoretical ambitions*; and E) the *lack of an account of normativity* within this theory. These criticisms do not render enactivism itself untenable, however. As we will see in the next chapter, there is an alternative formulation of the enactive framework that addresses the majority of these limitations and may offer a more adequate basis for a comprehensive theory of cognition.

One of the most persistent and conceptually troubling features of sensorimotor enactivism is the ambiguity surrounding the notions of A) "mastery" or "knowledge" of sensorimotor contingencies. In Alva Noë and Kevin O'Regan's formulation, perception is made possible by the perceiver's skillful engagement with their environment—specifically, by the implicit grasp of the lawful dependencies between movements and sensory changes. This idea is central to what they describe as the "sensorimotor account" of perception (O'Regan and Noë 2001), and it remains the

theoretical core of Noë's more recent elaborations. Indeed, Noë has increasingly referred to his position as "sensorimotor knowledge enactivism" (Noë 2021), presumably to underscore the view that perceptual experience is grounded not merely in activity, but in an agent's embodied understanding of how their movements bring about changes in sensory input. However, as Hutto and Myin (2013, 23–32) have argued, this invocation of "knowledge" risks reintroducing the very representational structures that enactivism was meant to replace. If perception depends on the possession of knowledge—even in a dispositional, non-conceptual, or embodied form—this invites questions about its ontological status. Is this knowledge something stored in the brain or body? Does it have propositional content? Is it structured in a way that mirrors classical representations, albeit in a distributed or non-symbolic format? Even if one attempts to avoid these implications by insisting that the relevant knowledge is not of the "know-that," but rather of the "know-how" type, the vagueness of the formulation remains problematic.

To be fair, Noë explicitly rejects the idea that this mastery involves any form of stored inner content. For him, the knowledge in question is not a representation of sensorimotor rules but the embodied capacity to bring them forth in action (Noë 2004, 2–5). Nevertheless, even practical knowledge or know-how can be interpreted functionally—that is, as a capacity governed by internal rules or mappings between stimuli and responses. Without a robust alternative to this functionalist reading, sensorimotor enactivism remains vulnerable to the criticism that it merely repackages representations in a more opaque vocabulary. Moreover, without a precise account of what counts as mastery—how it is acquired, individuated, and evaluated—the notion becomes too flexible to do serious explanatory work. It risks becoming a placeholder for perceptual success rather than a genuine theoretical construct. This concern ties into a deeper problem with the theory: its inability to provide a non-circular account of normativity in perception. As I will argue below, perceptual success cannot simply be defined *post hoc* in terms of behavioral adequacy, unless one already assumes a normatively structured background—a notion absent from the original sensorimotor enactivist framework.

A further difficulty with sensorimotor enactivism concerns its silence about the B) development of sensorimotor contingencies—namely, the question of how mastery of sensorimotor contingencies is acquired in the first place. If perceptual experience presupposes the subject's skillful grasp of the lawful dependencies between movement and sensory input, then a troubling bootstrapping problem arises: perception depends on knowledge of sensorimotor contingencies, but the acquisition of that knowledge appears to depend on the agent's prior capacity to perceive and act meaningfully in a structured world. This generates a form of epistemic circularity. The theory, as currently formulated, seems to assume the very capacities it seeks to explain. It begins with an already-formed agent, equipped with a repertoire of perceptual skills, but offers little insight into how such competence could have arisen developmentally. This is particularly problematic in light of empirical work in developmental psychology and neuroscience, which increasingly emphasizes the role of early sensorimotor learning, active exploration, and neuroplasticity in shaping perceptual abilities (Smith and Gasser 2005). Sensorimotor enactivism, by contrast, tends to model perception as if the organism were already a perceptually fluent agent, bracketing the formative processes through which sensorimotor coupling is itself constituted and refined over time.

To their credit, proponents of the sensorimotor view are not unaware of this concern. O'Regan (2011) points to the brain's capacity to continually recalibrate and adjust sensorimotor expectations, particularly in cases of sensory substitution or motor adaptation. Similarly, Noë (2004, 221–226) draws on examples of perceptual learning—such as the acquisition of new sensorimotor schemes through tool use or novel forms of interaction with the environment—to argue that sensorimotor contingencies can be gradually internalized through active engagement. On this view, mastery is not a prerequisite for perception, but an ongoing developmental achievement that emerges through embodied experience. However, these appeals to plasticity and learning do not resolve the core ambiguity. They explain how existing sensorimotor schemes might be updated or refined, but not how the very first perceptual episodes occur—how an infant or novice system initially acquires the minimal coupling necessary for learning to begin at all. In the absence of such

an account, the sensorimotor framework risks treating agency and perceptual competence as unproblematic givens, thereby neglecting the biological and normative processes by which these capacities emerge.

A further point to be made about Noë and O'Regan's appeal to "knowledge" of sensorimotor contingencies concerns its role in the C) individuation of perceptual systems. If such knowledge is interpreted simply as the capacity to coordinate movement and sensory input in accordance with lawful regularities, then any system exhibiting such coordination—no matter how mechanistic or unreflective—could in principle qualify as cognitive. A motion-tracking surveillance camera, for example, registers patterns of movement and adjusts its orientation accordingly. If this counts as an instance of sensorimotor knowledge, then the concept of cognition is in danger of collapsing into mere functional responsiveness. This overextension threatens to trivialize the enactivist project: if every form of regular sensorimotor coupling counts as perception, then perception becomes indistinguishable from basic causal reactivity.

Noë and O'Regan might respond to this worry by emphasizing that sensorimotor contingencies are not merely physical regularities, but meaningful relations enacted by situated agents. Noë (2004, 214–218) often appeals to the idea of presence—that in perceptual experience, objects are not just encountered as stimuli but as present to the agent in a structured field of possibilities for action. This sense of presence, he argues, depends on the organism's ongoing ability to access sensorimotor dependencies, even if they are not currently exercised. On this view, a camera does not perceive because it lacks access to the relevant counterfactuals—it cannot anticipate or navigate changes in its environment in a flexible, situated manner. Similarly, O'Regan (2011, 179–183) argues that perceptual experience arises only when a system is capable of exploring its sensory dependencies and distinguishing between changes caused by itself and changes caused by the world.

While these clarifications do help avoid the most obvious forms of overextension, they are ultimately insufficient for articulating a robust criterion of cognitive individuation. The appeal to

access to sensorimotor dependencies, or to the richness of counterfactual sensitivity, still leaves open the crucial question: why do some systems have such access while others do not? And more importantly, why should access to counterfactuals be taken as the threshold of cognition? Without a deeper account of agency and normativity, such criteria remain descriptive rather than explanatory. They tell us how skilled systems behave but not why such behavior counts as cognitive rather than merely reactive. In this regard, sensorimotor enactivism, as it stands, lacks the theoretical resources to explain why some sensorimotor couplings matter to the system—why they are meaningful engagements rather than inert correlations.

Even if we grant that the concept of cognition can be narrowed in order to avoid overextension, sensorimotor enactivism still faces a more fundamental limitation: its D) limited explanatory scope in view of its broader theoretical ambitions. By design, the theory was developed as a challenge to traditional accounts of perceptual experience—particularly those that treat perception as a passive, internal process mediated by inner representations or sense-data. This reorientation has been notably fruitful for thinking about vision, touch, and other action-oriented modalities, and it has stimulated important empirical and philosophical work on perceptual learning and embodied skill (Noë, 2004; O'Regan, 2011). However, beyond these domains, the theory struggles to provide substantive accounts of more complex cognitive capacities such as memory, imagination, language, or abstract reasoning. At best, it gestures toward these capacities as higher-order sensorimotor patterns, but it offers no principled framework for how these patterns operate offline, become decoupled from immediate action, or integrated into temporally extended and socially embedded practices.

To this concern, Noë might respond by emphasizing that his goal was never to explain all of cognition, but rather to dismantle entrenched assumptions about perception, particularly the idea that perceptual content is something internal to the head (Noë 2021). He could argue that other aspects of cognition—such as memory or imagination—may require additional theoretical

tools, and that sensorimotor enactivism should be judged within the limited domain for which it was originally designed. This position reflects a kind of strategic modesty, and it has the merit of keeping the theory tightly focused and empirically testable. Yet this very modesty becomes a limitation if the framework is presented, as it sometimes is, as the foundation for a broader paradigm shift in cognitive science. A theory that aspires to replace cognitivism at large cannot remain agnostic about the nature of thought, memory, or affect.

Moreover, even the claim that the sensorimotor approach explains phenomenal consciousness remains contested. While the theory has had success explaining the dynamic structure of sensorimotor engagement, it offers little traction on other dimensions of conscious life—such as moods, emotions, reflective awareness, or bodily affect—that do not map neatly onto sensorimotor patterns. These phenomena involve temporally extended, self-modulating processes that are not easily reduced to immediate patterns of action and feedback. For instance, as Giovanna Colombetti (2014) and Evan Thompson (2007) have argued, affectivity is not merely a background modulation of perception, but a constitutive dimension of conscious experience. Without a framework for explaining how perception is shaped by an agent's biological and emotional attunement to its environment, the sensorimotor account risks leaving out precisely those aspects of cognition that make it meaningful for the organism.

This limitation in explanatory scope is intimately connected to a deeper theoretical deficiency: sensorimotor enactivism E) lacks an account of the normative, affective, and sense-making dimensions of cognition. While it characterizes perception as skillful interaction with the environment, it does not explain why certain interactions matter to the organism—why they are experienced as good or bad, relevant or irrelevant, desirable or threatening. In its current form, the theory models the structure of perceptual engagement but remains silent on its significance. Sensorimotor enactivism is structurally descriptive: it explains how patterns of sensorimotor contingencies are enacted but lacks the conceptual tools to make sense of why some trajectories are valued over

others, or how meaning arises from bodily engagement. Without an account of intrinsic normativity—of how a system evaluates its engagements relative to its own needs or goals—sensorimotor mastery remains inert, incapable of grounding genuine agency, motivation, or concern.

Noë and O'Regan might respond by appealing to the idea of pragmatic success: that correct perception is defined by the agent's ability to guide action effectively. For instance, Noë (2004, 210–213) maintains that perceptual accuracy is not about internal correspondence but about the skilled guidance of behavior. Similarly, O'Regan (2011, 201–204) proposes that perceptual content arises from the potential to act on sensorimotor contingencies. From this perspective, normativity is not absent but built into the practical success or failure of interaction: when perception enables appropriate action, it is deemed "correct."

Yet this account of normativity is ultimately external and instrumental. It defines correctness in terms of third-person observable outcomes—successful navigation, appropriate response—without addressing the first-person perspective of the organism. It leaves unexplained how or why a given interaction matters to the system itself. As Thompson (2007, 154–157) emphasizes, what is missing is a notion of sense-making: the idea that a cognitive system interprets and evaluates its environment relative to its own continued viability. Sense-making involves an evaluative stance—a felt orientation toward one's surroundings that includes concern, salience, and affect. It is this internal normativity that distinguishes a perceiving subject from a reactive mechanism. A robot can detect obstacles and avoid them; a living organism can find those obstacles threatening, frustrating, or dangerous.

Without a way to articulate this affective and evaluative dimension, sensorimotor enactivism falls short of offering a complete account of cognition. The patterns it describes are real, but they do not suffice to explain what makes those patterns matter for the system enacting them. In this respect, the theory does not fail because it is wrong, but because it is incomplete. It opens important conceptual territory while leaving key questions unanswered—questions that concern

the very nature of mindedness, agency, and experience. In the following chapter, I explore an alternative strand of enactivist theory—i.e., autopoietic enactivism—that attempts to fill this gap by grounding normativity not in external performance criteria, but in the organism's own self-regulation and affective orientation toward the world.

4. Conclusion

Sensorimotor enactivism has offered a bold and compelling challenge to traditional representationalist models of perception. By reconceiving perception as an active skill enacted through lawful couplings between action and sensory input, it has (rightly, in my view) dismantled the myth of the passive observer and the central Cartesian theater of the mind. Yet, for all its conceptual elegance and empirical resonance, the theory ultimately falters when asked to speak to the full complexity of the mind. It remains entangled in ambiguities about what "mastery" entails, cannot account for the origins or development of perceptual skills, and fails to distinguish truly cognitive systems from mechanistic ones, among other issues. In attempting to sidestep representation, it risks reinstating it in a covert form. Most critically, sensorimotor enactivism is silent on the very features that make cognition more than mechanical: normativity, affectivity, intentionality, and meaning. As such, it succeeds not as a theory of mind, but as a partial insight—valuable, yes, but limited to the periphery of cognition rather than its core.

What is needed, then, is a deeper paradigm—one that does not merely reconfigure the interface between sensation and action, but grounds cognition in the existential predicament of living systems. Autopoietic enactivism answers this call by shifting the explanatory center from sensorimotor coordination to biological self-organization, from perceptual skill to the precarious task of self-maintenance. It recognizes that cognition begins not with data, but with concern—with the organism's need to survive, adapt, and make sense of a world that matters to it. In doing so, it retrieves the affective, normative, and developmental dimensions that sensorimotor enactivism

elides, offering a richer and more encompassing account of what it means to have a mind. If sensorimotor enactivism taught us that perception is not passive, autopoietic enactivism teaches us that cognition is not neutral—it is evaluative, perspectival, and fundamentally rooted in life. It is toward this more foundational vision that we will turn in the next chapter.

III. AUTOPOIETIC ENACTIVISM

As previously mentioned, historically, autopoietic enactivism predates its sensorimotor counterpart. It builds upon the theory of autopoietic systems, a framework in biology and philosophy developed in the 1970s that deployed ideas from the fields of cybernetics and systems theory in an attempt to offer a rigorous definition of the concept of life. Thus, having concerned ourselves mostly with the philosophies of mind and perception in the last chapter, we will now briefly turn our attention towards a different issue—the philosophy of biology—, whose connections to the topics of cognition, action, and perception will hopefully become clear as we move forward.

Just as the works of Hurley, Noë, and O'Regan served as a minimal canon in our discussion about sensorimotor enactivist theory, while investigating the autopoietic variety, we will resort primarily to those of Humberto Maturana, Francisco Varela, and Evan Thompson. In the 1970s, Maturana and Varela worked together in the development of a theory of autopoietic systems that could offer a rigorous definition of life and living systems. Their results quickly invited them to incorporate also an autopoietic account of cognition into the broader theory of autopoiesis, something that was attempted in *Autopoiesis and Cognition* (1979). Later, Varela developed their framework into a full-fledged autopoietic theory of cognition that aimed to present itself as an alternative to the computationalist paradigm that dominated the cognitive sciences.

This culminated in the writing and publishing, together with philosopher Evan Thompson and psychologist Eleanor Rosch, of the book *The Embodied Mind* (1991), which is retrospectively considered an important manifesto of enactivism. Until his premature death in 2001, Varela was the main voice and the most prolific author within the enactivist movement. After that, it was Thompson who gave continuity to the project of fleshing out the theory, as shown in what is arguably the most comprehensive work on the topic until now, his book *Mind in Life* (2007). The

seminal works I mentioned will be our main references when analyzing the most general features of autopoietic enactivism, although I will also resort to more recent publications when delving into some of the more difficult details of the theory.

1. THE AUTOPOIETIC THEORY OF LIFE

1.1. DEFINING "LIFE"

Distinguishing between living and non-living systems is a matter of ongoing conceptual and theoretical interest in biology, as such distinctions help clarify the domain of biological investigation and inform research strategies. While it is certainly debatable whether the science of life requires a precise, universally accepted definition of its subject matter—many sciences, after all, operate productively with vague or open-ended categories—it remains the case that the question "What is life?" continues to motivate important lines of inquiry in theoretical biology. As Thompson notes, definitions of life usually focus on one of its three different aspects: it is possible to characterize life in i) genetic and reproductive terms, focusing on "historical continuity and evolution, on the genetically based linkage of generations and the arising of novel variants within a population as a result of various evolutionary factors;" but it is also possible to characterize it from an ii) ecological perspective, where "individual organisms are seen not only as members of reproductively linked populations, but also as beings that interact constructively with their environments," that is, as "niche-constructing" entities; and, finally, it is possible to define life in view of iii) the individual living system, a here-and-now entity taken as a system and abstracted away from historical and environmental considerations (Thompson 2007, 95–96). What concerns us here, since this is what was sought after by autopoietic theorists, is that last type of definition.

Let us remember that a classical definition is expected to offer necessary and sufficient conditions for something to fall within the scope of its *definiendum*. Although this kind of definition

is standard in mathematics, its applicability to empirical concepts—especially vague and complex ones like "life"—is far more contentious. Still, the attempt to find a general characterization of the concept of life remains relatively common in discussions in biology, both as a heuristic and as an aspirational theoretical goal. In the case of living systems, a relatively long series of criteria has been proposed: homeostasis, organization, metabolism, growth, adaptation, stimulus-responsiveness, reproduction, etc.

However, none of them—and no specific combination of two or more of them, for that matter—seems to be capable of encompassing the totality of entities that we are willing to call "living." Many non-living systems are homeostatic, for example, and whereas reproduction might be present across all species, there are clearly non-living systems that are also capable of self-replication. These are, therefore, insufficient conditions for a successful definition of life. Similarly, many of them seem to be unnecessary: even if the presence of metabolic processes does characterize a great number of living entities, there are seeds and spores that can remain perfectly dormant—from a metabolic perspective—for hundreds or thousands of years before they finally germinate. Thompson presents additional reasons for doubting the strategy of listing necessary and sufficient conditions in the quest for the definition of life:

> The problem with this approach is that it is descriptive, not explanatory. It takes for granted the distinction between living and nonliving, and then it lists some common characteristics of systems accepted as living. But how do we know which characteristics should be included on the list or when the list is complete? Lists describe things but do not explain them. To explain we need a theory.
>
> (Thompson 2007, 97)

This task was undertaken by a number of theoretical biologists and philosophers of biology during the course of the development of that discipline. Autopoietic theory, as developed by Maturana and Varela in the 1970s, is one of them. Theirs, however, was obviously not the first attempt at doing so, and, therefore, before moving on to the details of their own theory, we will

briefly go over other historical contenders in the quest for a definition of life and see why their answers can be deemed unsatisfactory. Most strategies for a definition of living systems as individual entities can be fitted within three broader families: metabolic (e.g., Schrödinger 1944; Morowitz 1968; Rosen 1991, Kauffman 1993), informational (e.g., von Neumann 1966; Monod 1971; Crick 1966; Dawkins 1976; Dennett 1995), and structural-morphological (e.g. Virchow 1860; Maturana and Varela 1980; Margulis 1993).

Now, metabolic and informational theories of life represent two distinct and big families of related perspectives within the field of biology, each attempting to contribute to our understanding of the nature and dynamics of living systems. The metabolic theory of life centers on the notion that the defining characteristic of living entities is their capacity for metabolism, the set of chemical processes that sustain and regulate vital functions. According to this theory, life manifests as a network of interconnected biochemical reactions, wherein organisms continuously exchange matter and energy with their environment to maintain their integrity and perpetuate existence. Central to this theory is the idea that living systems actively engage in the transformation of energy and materials, thus distinguishing themselves from non-living entities.

For the metabolic view, a list of necessary and sufficient features that an entity needs to present in order to be classified as living would include any combination of the following: metabolism, energy transformation, homeostasis, growth and development, organization, etc. Different varieties of the metabolic view will be produced depending on how you combine these criteria. Thus, a focus on physiological processes such as breathing, excreting, and growing as definitional criteria will lead one closer to a *physiological* definition, but it is also perfectly possible to focus on the thermodynamic features of metabolizing systems and define life in related terms, such as lowered entropy in relation to its surroundings due to its organized metabolic activity. In this case, we would have a thermodynamic strategy. Despite this internal variety, I do think it is reasonable to classify such theories together as constituting a single family.

The informational theory of life, on the other hand, accentuates the pivotal role of information processing and storage within living organisms. At the core of this perspective is the recognition that life is characterized by its ability to encode, transmit, and utilize genetic information, usually encapsulated in DNA and other nucleic acids. This hereditary information serves as the blueprint for the development, functioning, and reproduction of living beings. The informational theory thus underscores the significance of molecular codes and regulatory mechanisms in orchestrating the intricacies of life's manifestations. Therefore, it usually includes any combination of the following as necessary and sufficient conditions for membership within the class of living systems: information processing, presence of genetic material, replication, mutation, genetically-mediated evolution, etc.

A different family of approaches is the structural-morphological strategy for a definition of life. As its name suggests, instead of focusing on specific processes that tend to be found in most living organisms such as metabolism or information processing, structural theories tend to focus on structural and morphological features found across all or most living systems. Perhaps the most distinguishable among the myriad of such features is the fact that all life-forms currently known are made up of cells. This is the reason why most structural approaches to the quest for a definition of life tend to focus on the cellular structure of life and, therefore, can also be called cell-based approaches.

1.2. AUTOPOIESIS

Whereas the prefix "*auto-*," meaning "self," needs no introduction, the less well-known Greek word "*poiesis*," in turn, means "creation" or "making." Thus, "autopoiesis," the compound term coined by Varela and Maturana in their 1972 book *Autopoiesis and Cognition*, is simply intended to mean "self-creation" or "self-making." The reasoning behind this neologism will become clear in the next few paragraphs. The autopoietic definition of life comes from what can be seen as this family

of structural or cell-based theories of life. The idea of the cell as "the ultimate irreducible form of every living element" was first developed by Rudolf Virchow (1860) in the 19th century. Despite enormous advancements in molecular biology, the study of life does not seem to have moved significantly from Virchow's basic insight: even if DNA and RNA, for instance, which could theoretically support information-processing definitions of life, do play a tremendously significant role in determining the course of biological processes, they cannot account for the phenomenon of life in isolation from the specific functional organization in which they participate—i.e., the cell as a whole. Moreover, there is a relative consensus that denies the status of living systems to acellular, lifelike entities—again, such as viruses and prions.

However, as noticed by Thompson, there seems to be something wrong with a strictly cell-based structural-morphological definition of life. After all, we identify being a life form with having cells because of the contingent fact that we have never encountered life forms that are not made up of cells on Earth. Hence, the identification is tautological: "[…] life is cellular because there is no life without the cell" (Thompson 2007, 97). Thus, if we endorse a straightforwardly cell-based definition of life, we seem to have been led to the same allegedly failed listing strategy that was mentioned at the beginning of this chapter, and to be victims of the same shortcomings it faced. To go beyond the tautology, we would need a specification of the organization of living entities that is independent of their particular, contingent structure: a minimalist description of the functional organization of life that could help define it without resorting to the notion of "cell" itself. Thompson formulates this problem in a very clear manner in the following passage:

> One strategy for meeting this demand would be to characterize a living system, such as a cell, in terms of its relational form or organization (Maturana and Varela 1980, 1987). A system's organization consists of the relations that define the system as being a member of a specific class. For something to qualify as an automobile, for instance, its parts have to be arranged or related to one another in a certain way. In specifying a system's organization, one abstracts a pattern or set of relations that defines what kind of system the system is. A system's organization is thus not equivalent to its actual structural relations and components because the same organization can be structurally realized in

> different ways, and a system can undergo structural change without necessarily changing its organization. Thus, the organization of an automobile can be realized in different physical materials and mechanisms, and these can change during the automobile's lifetime. Similarly, a single cell, such as a bacterium, undergoes many structural changes during its life cycle without changing its organization as a unicellular organism. Is it possible to specify a definitive organization in the case of living systems? With regards to the cell theory, can one specify the organization of a cell independent of its structure?
>
> (Thompson 2007, 97)

The description of this basic organizational or functional feature of living systems is exactly what autopoietic cell theory attempts to do, and this is what distinguishes it from other contenders within the structural-morphological family of attempts at defining life. In a sense, therefore, one could initially see some surprising and unexpected similarities between the strategy adopted by functionalists in philosophy of mind and autopoietic life theorists. After all, according to autopoietic cell theory, what is really characteristic of living systems is a specific functional organization of its parts, and not the contingent ways in which this organization has been materially implemented throughout evolutionary history. The parallel with what functionalists say about the relationship between mind and brain is quite striking, and it will be further explored in §3. At first sight, therefore, the autopoietic theorist should have no good reason to deny the possibility of alternative material implementations of such an organizational pattern, which leaves open, for instance, the possibility of artificial life—an analogue to the idea of the possibility of consciousness in artificial intelligence.

Moreover, whatever this minimal organization turns out to be, there are good reasons to suppose that it must have already been present in the protocells that are hypothesized to be at the origin of life on Earth—molecular configurations that present some structure separating them from their immediate environment, such as a lipidic membrane, and perhaps some form of rudimentary metabolism, as well as the necessary structures to support it. If this were the case, we would avoid the chicken-and-egg type of problem that plagues the cell theorist, because the origin

of life on Earth could be traced back in time to the moment of the emergence of the first molecular systems displaying this specific minimal organizational pattern. If what is needed is a definition of life in terms of the organization of its most basic structural features, regardless of the contingencies of its implementation, then what is needed is a functionalist theory of life. The concept of autopoiesis is precisely Maturana and Varela's attempt at offering such a basic functional description of the living. Thompson notes that: "The concept of the autopoietic organization arose from an attempt to abstract from the molecular processes of the cell the basic form or pattern that remains invariant through any kind of structural change, as long as the cell holds together as a distinct entity" (Thompson 2007, 97).

The notion of autopoiesis emerged out of the application of methodologies from cybernetics and systems theory to the study of biological phenomena specifically, and hence it is marked by some of the central preoccupations of cyberneticians and systems theorists, such as feedback-mediated exchanges (Wiener 1948) between structurally coupled systems (Ashby 1952; 1956). According to autopoietic theory, a cell is a thermodynamically open system that exchanges matter and energy with the environment continually, and yet preserves a certain functional organization that is necessary in order for this process to continue going on indefinitely—that is, until failure due to wear and tear and, consequently, cell death. This functional organization is characterized by a loop: the cell obtains from the environment the material and energy it needs to build the structure that separates it from that very same environment—i.e., the cell membrane—, thus allowing it to selectively obtain and retain what is needed for the reiteration of this process: "[…] a cell produces its own components, which in turn produce it, in an ongoing circular process" (Thompson 2007, 98).

There are two possible antonyms to the notion of autopoiesis—*allopoiesis* and *heteropoiesis*—each one negating one of two different aspects of autopoietic systems, the active and passive ones. Both "allo-" and "hetero-" are Greek prefixes that convey the idea of otherness, but when it is said that a system is allopoietic, we are focusing on the fact that it *actively* produces something else,

whereas when it is said that it is heteropoietic, the focus is on the fact that it is *passively* produced by something or someone else. For our purpose here, however, these terms can be treated as synonymous, since it is almost always valid to deduce from the fact that a system does not produce its own components that it is not produced by itself. Therefore, from now on, I will use the word "heteropoietic" in a sense that encompasses both these active and passive aspects.

Notice that, just like all man-made machines up to the present, subcellular structures such as ribosomes or mitochondria are also examples of systems that are not autopoietic, since they are not capable of producing the components necessary for their own self-maintenance (Varela, Maturana, and Uribe 1974, 189). This idea of the self-sufficiency of the cell when building itself is a very important concept within autopoietic theory and is called *autonomy*. The fact that the cell—and, hypothetically, any other system that mirrors its functional organization—is capable of self-production by means of its selective interaction with the environment is what makes it autonomous, whereas heteropoietic systems cannot be said to be autonomous precisely because their output is insufficient for promoting self-preservation. While doing so, the autopoietic system gets individuated, since it distinguishes itself from the environment by establishing a clear boundary—i.e., in the case of the cell, the cell membrane—between its internal domain and the external world.

While external molecular interactions are guided solely by physical laws and those that govern the relevant chemical reactions, the interaction between molecules within the internal domain already show some sort of metabolic pattern, since they are additionally governed by biochemical principles and laws of bioenergetics, which are constraints introduced in the system by the selective nature of the cell membrane and by the function of whatever structures are present in its cytoplasm. Thus, Maturana and Varela define autopoiesis in the following manner, exemplifying its concrete realization by means of the basic cell structure:

> The autopoietic organization is defined as a unity by a network of productions of components which

(i) participate recursively in the same network of productions of components which produced these components, and

(ii) realize the network of productions as a unity in the space in which the components exist.

> Consider for example the case of a cell: it is a network of chemical reactions which produce molecules such that
>
> (i) through their interactions generate and participate recursively in the same network of reactions which produced them, and
>
> (ii) realize the cell as a material unity.
>
> Thus, the cell as a physical unity, topographically and operationally separable from the background, remains as such only insofar as this organization is continuously realized under permanent turnover of matter, regardless of its changes in form and specificity of its constitutive chemical reactions.
>
> (Varela, Maturana, and Uribe 1974, 188)

These criteria are satisfied by a large number of entities at the cellular level and, as we have seen, by absolutely no subcellular structure. But what about multicellular organisms? Surely, there is an important sense in which it is meaningful to talk of a tree, a fish, or a man as autonomous and self-producing; hence, as an autopoietic system. In fact, extending the concept of autopoiesis to multicellular organisms will prove essential for defending more advanced tenets of autopoietic enactivism. With that in mind, Maturana and Varela (1987, 87–89) distinguish between first-order and second-order autopoietic systems. Cells are first-order autopoietic systems, whereas systems that include individual cells as their structural components—i.e., multicellular organisms, which they also call metacellular systems—are second-order autopoietic systems. Autopoietic theory, however, was originally developed with the explicit aim of explaining life at the basic level of the cell, and extending it to multicellular organisms, although a desirable goal, can lead us to much more controversial claims.

The idea that living systems continuously produce and regenerate their own components, while maintaining their structural integrity, can possibly be extended to multicellular organisms by considering the interactions between different cells within an organism's body. In multicellular organisms, individual cells contribute to the overall functioning of the organism while also maintaining their own identity as distinct entities. Maturana and Varela argued that the interactions between cells and their specialization in various tasks are organized in a way that contributes to the overall autopoietic nature of the organism: each cell plays a role in maintaining the overall structure and functioning of the organism, and the interactions between cells contribute to the dynamic balance required for the organism's survival as a whole.

For example, in a multicellular organism like a human, cells differentiate into various types—neurons, myocytes, hemocytes, etc.—and work together to maintain the organism's overall functioning. While each cell has its own specific functions, they are interdependent and collectively contribute to the autopoiesis of the entire organism, and can therefore be seen as making up its internal reaction network. The interdependence and cooperation between cells do indeed reflect the core principles of autopoiesis. Maturana and Varela's work emphasizes that the definition of life should not be limited to individual cells but should encompass the organization and interactions of cells within a multicellular context. The central insight remains the same: the continuous self-production and maintenance of the components that make up the living system, whether they are individual cells or the various components within a multicellular organism.

However, there are still some apparent incompatibilities with the initial definition of autopoiesis as applied to multicellular organisms. Most notably, whereas the boundary of autopoietic systems is clear at the cellular level, this is not necessarily the case when one is discussing multicellular organisms. In these, individual cells are specialized and organized into tissues, organs, and systems that work together in order to maintain the overall function of the organism. While individual cells have their own lipidic membranes that provide a clear boundary at the cellular level,

the question arises as to whether the entire multicellular organism can be considered as having a single, coherent semipermeable boundary that separates it from its environment.

In multicellular organisms, there are various interconnected and interdependent cellular components that are in continuous communication with each other. The interactions between cells, the exchange of signals, nutrients, and waste products, and the overall coordination required for the organism's functioning challenge the notion of a single, distinct boundary. Taking a single organ such as the skin as the multicellular equivalent of the cell membrane also presents its own difficulties, since the skin only performs a small subset of the functions that the semipermeable boundary is supposed to perform in autopoietic systems—it does not, for instance, act as the primary filter as to what is allowed inside the system, a function generally performed by the digestive system. The challenge here is that autopoietic theory's strict emphasis on a clear and impermeable boundary as a defining feature of life might not align perfectly with the complexity of multicellular organisms. Take, for instance, Thompson's attempt at formulating a decision tree for determining whether a system is autopoietic or not:

1. *Semipermeable boundary:* Check whether the system is defined by a semipermeable boundary made up of molecular components. Does the boundary enable you to discriminate between the inside and outside of the system in relation to its relevant components? If yes, proceed to 2.
2. *Reaction network:* Check whether the components are being produced by a network of reactions that take place within the boundary. If yes, proceed to 3.
3. *Interdependency:* Check whether 1 and 2 are interdependent: are the boundary components being produced by the internal network of reactions, and is that network regenerated by conditions due to the boundary itself? If yes, the system is autopoietic.

(Thompson 2007, 103)

In view of this algorithm, Thompson seems to think that whether multicellular organisms are autopoietic or not is still an open question, for "[…] much depends on how we interpret

'boundary' and 'internal reaction network' in the criteria of autopoiesis" (Thompson 2007, 106). This vagueness, I think, is a potentially serious issue, for it leads to a significant deflation of autopoietic theory's initial claim to being capable of offering a rigorous definition of "living system." After all, multicellular organisms are indeed living systems, and if they do not conform to the autopoietic requirements for being so, then it is dubious that autopoietic theory can claim success in its attempt at defining the living. In this sense, Varela and Maturana's definition of multicellular organisms as second-order autopoietic systems may seem significantly *ad hoc* if the criterion for admission into this class of systems is precisely that of presenting a significant number of the features that characterize autopoietic systems, but perhaps not others. This would be particularly worrying if among the lacking features were the very central one of possessing a clearly-defined semipermeable boundary—something that, as we have seen, multicellular organisms do not seem to have, at least not in the most literal sense of the expression. For the sake of consistency, Thompson seems to concede that non-unicellular systems without a clear boundary might not qualify as autopoietic, even if they still present all the other properties needed in order to be characterized as autonomous:

> Here Varela's (1979) distinction between autonomy (organizational closure) and autopoiesis is pertinent. An autopoietic system is a specific kind of autonomous system—one having an organizational closure of production processes in the molecular domain—but there can be autonomous systems that are not autopoietic if their constituent processes exhibit organizational closure in their domain of operation. For example, an insect colony or animal group might qualify as autonomous in this sense. On the other hand, taking "boundary" to mean only a unicellular semipermeable membrane or even a multicellular epidermal layer seems too restrictive (plants and insects do not have a skin). Rather, the crucial matter is that the system produce and regulate its own internal topology and functional boundary, not the particular physical structure that realizes this boundary.
>
> (Thompson 2007, 106–7)

The class of the autonomous systems contains in it, therefore, the class of the autopoietic systems. The suggestion seems to be that, even though multicellular organisms and other complex systems might not qualify as autopoietic, they are autonomous and, therefore, a great deal of the consequences of autopoiesis that will be presented later on also apply to them. This insight will be important when, in §3, we return to issues concerning the semipermeable boundary that individuates autopoietic systems. For the time being, however, it is not unreasonable to see this maneuver as a significant slackening of the original rigor that characterized autopoietic theory's conceptual landscape. Additionally, as illustrated by Thompson's quotation, the idea of applying this first iteration of the concept of autopoiesis to multicellular organisms seems to invite further expansions of its scope of application—towards, for instance, explaining intraindividual social systems or even Gaia, the sum of all of Earth's ecosystems, if such a concept is scientifically defensible. These, in turn, present their own sets of additional problems (Thompson 2007, 219ff).

Attempts at treating animal social systems—e.g., ants and bees' colonies—as autopoietic have been made, and even at modelling human social systems in such terms, as epitomized in Niklas Luhmann's work in sociology. Those, however, are still far from uncontroversial, even among autopoietic theorists themselves: Maturana, for instance, disapproved of Luhmann's use of the term "autopoiesis" in the context of human social systems. One significant consequence of adopting such a stance is the further weakening of the claim that the concept of autopoiesis includes everything necessary and sufficient for a definition of life, for societies are obviously not living entities in the same literal sense as the individuals they are composed of. Despite this caveat, I still do not think these apparent inconsistencies within the first iterations of autopoietic theory jeopardize its posterior developments, particularly those which led Varela to formulate his autopoietic enactivist theory of cognition. Let us, therefore, resume our argument in that direction.

1.3. TELEOLOGY AND TELEONOMY

We are now approaching the point where the description of living systems as autopoietic unities connects to their description as cognitive systems. Before that, however, it is essential that we understand how the theory of autopoiesis accounts for the apparent teleological behavior of living systems, as opposed to the behavior that is explainable in strictly mechanistic terms shown by non-living systems, since this is what the autopoietic explanation of cognition will be built upon. Immanuel Kant thought that we could not explain biological phenomena mechanistically, following the paradigm given by other more well-established sciences of his time, but must rather explain it teleologically (Kant 1790/2000). The reasons for that were i) the intrinsically self-organizing nature of living organisms, which are both cause and effect of themselves; ii) the fact that the teleological conception cannot be reduced to the mechanistic paradigm of efficient causation in a world of inanimate matter.

Thompson notices that the Kantian paradigm of a mechanistic explanation was Newtonian physics, and also that his main difficulty was that of fitting the characteristic self-causation of living systems within this specific conception, adding that this connects his insight to the concerns that motivated the development of the autopoietic theory (Thompson 2007, 129–30). Hence, for Kant, explanations of organic nature must make use of the teleological concept of *purpose*, for we have no reason to "[…] hope that perhaps someday another Newton might arise who would explain to us, in terms of natural laws unordered by any intention, how even a mere blade of grass is produced" (Kant 1790/2000). Teleological explanations reverse the cause-effect schema that is traditionally employed in mechanistic explanations, presenting a desired effect—i.e., an end or purpose—as the cause of an agent's initiating the relevant course of action. In teleological accounts, therefore, we have a means-end explanatory schema.

Since Kant's times, however, the landscape of the life sciences has changed substantially. Many would argue that Charles Darwin can perhaps be seen as the much-awaited Isaac Newton of

biology, that is, as the scientist who was first capable of offering a strictly mechanistic explanation of biological phenomena, proving Kant's diagnosis wrong. In this sense, the notion of teleology, which seems to imply purposiveness brought about through divine or human intention—be it intelligent design or the fact that constraints upon our human intellects force us to view organisms in terms of final purposes—, is replaced by that of *teleonomy*, which lacks these connotations. Teleonomy is a concept often used in biology and philosophy to describe the apparent purpose or goal-directed behavior exhibited by living organisms or systems, while highlighting the fact that behind this apparent purposiveness, in fact, lie nothing more than mechanistic cause-effect interactions. Thus, it refers to the idea that certain biological structures, functions, or behaviors appear to be guided by a form of intrinsic purpose or end, even though they may not be the result of conscious intention or design. Contrary to teleology, the notion of teleonomy suggests that purposive behavior or patterns can emerge from natural processes and evolutionary mechanisms without the need for any external guidance or intention.

It is quite clear how teleonomic patterns could arise within a Darwinian framework, according to some interpretations of evolutionary theory. There are two basic constraints on the perpetuation of life forms, and they act at the ontogenetic and phylogenetic levels, respectively: self-preservation and reproductive success. All other teleonomic patterns can, in principle, be understood as derivative from these two basic constraints. Under the pressure of natural selection, organisms that exhibit traits enhancing self-preservation and reproductive success are favored over those that do not. Even if the biological processes by means of which such traits are developed are completely random and intentionless, the iterative cycle of variation, selection, and heredity drives their gradual refinement, ultimately resulting in a complex interplay of behaviors, structures, and functions that align with the imperatives of survival and successful reproduction. Thus, in animals, behavioral patterns such as hunting, foraging, and migrating can be understood in light of the teleonomic imperative of self-preservation, whereas others such as courtship, mating, and parental care can be understood in light of the teleonomic imperative of reproduction. Of course, it is at

least conceivable that the contingencies of the way an organism relates to its environment and to its conspecifics could lead to the development of even more complex and sophisticated behavioral patterns, such as working in an office to buy food or purchasing life insurance. The point is, precisely, that these could also be ultimately traced back to the imperatives of self-preservation and reproduction.

This is how many would phrase a Darwinian response to Kant's teleological problem. Thompson argues, however, that it simply will not work. As I have already mentioned previously (§1.1), when discussing different strategies for defining living systems, explanations of the phenomenology of life are usually multi-levelled: one can focus on its genetic and reproductive aspects, but also on its ecological aspects, or else on the individual organism. Darwin's theory of evolution concerns life at the two first levels, but not at the level of the individual, since, as noticed by Thompson, it "[…] must presuppose biologically organized individuals that reproduce" (2007, 131). Additionally, for this reason, "[…] Darwin's Newtonian framework, in which design arises from natural selection conceived as an external force, does not address the endogenous self-organization of the organism" (Thompson 2007, 131). This would put Kant's teleological problem out of the scope of evolutionary theory, which lacks the theoretical tools for crafting a satisfactory explanation for the organization of life at the level of the biological individual.

For Thompson, the autopoietic theory of life is the only framework capable of accounting for the teleological nature of organisms due to its emphasis on self-organization and intrinsic purposiveness. Unlike mechanistic explanations that focus exclusively on external causes and effects, autopoiesis recognizes that living systems are not passive entities but actively engage in self-production and self-maintenance. This self-organizing nature implies a teleological dimension, as organisms are treated as both cause and effect of themselves, producing and maintaining their own organization. Additionally, as will become clear later on, an expanded version of autopoietic theory is not only capable of offering a satisfactory teleological account of self-production within

organisms—i.e., of explaining how autopoiesis itself operates—, but also their purposive engagement with their immediate environments. That is, since the organism's self-productive functions are always profoundly dependent on the way it interacts with its environment, which is the material and energetic source of the processes by means of which it realizes autopoiesis, autopoiesis should also help us understand teleological functions that are not immediately self-directed. This, as we will see in the next section, is what will connect the autopoietic theory of life to an autopoietic theory of cognition—i.e., autopoietic enactivism.

2. THE AUTOPOIETIC THEORY OF COGNITION

2.1. INTENTIONALITY AND SENSE-MAKING

Varela and Maturana did not initially conceive of the notion of autopoiesis as encompassing any type of intrinsic teleology in the Kantian sense, but rather as a strictly mechanistic explanation of the way cells and multicellular organisms come to organize themselves (Thompson 2007, 141; Di Paolo 2005; Varela and Maturana 1980). In other words, even if teleological patterns could be identified in the behavior of autopoietic unities, they should not be seen as intrinsic to them, but rather as the result of the observer adopting a Dennettian intentional stance towards the object of their observation (Dennett 1987). In this view, as a scientific endeavor that is in the business of offering an objective description of the organization of living systems, it would be out of place for autopoietic theory to ascribe teleology or purposiveness to the organisms it attempts to explain, although teleonomics does not seem, in principle, incompatible with its framework. However, later developments in autopoietic theory led Varela to change his own initial stance on these matters, paving the way for the advent of enactivism as a more ambitious theory of both life and mind, even if Maturana never explicitly committed to these views and in fact restricted his own work to the merely biological aspects of autopoiesis. Varela's change of perspective is clearly illustrated by some of his later writings, where he complements the idea of autopoiesis as a process of an organism's

individuation and identity formation by metabolic means with the notion of sense-making (Varela and Weber 2002).

What prompted the introduction of sense-making into the conceptual repertoire of autopoietic theory was a series of considerations about the apparently teleological orientation of the autopoietic unity—i.e., the organism—towards its immediate environment, something that could not be satisfactorily explained by the immanent, minimalistic teleonomy implicit in the concept of autopoiesis itself. For Varela, there are two apparently teleological dimensions of life that need to be considered in a satisfactory scientific description of the genesis of an organism's purposiveness: self-preservation and externally-directed behavior. Whereas the basic concept of autopoiesis explains self-preservation quite satisfactorily, the notion of sense-making is needed in order for externally directed behavior to be covered. Thompson explains this in the following manner:

> The first mode of purposiveness is identity: autopoiesis entails the production and maintenance of a dynamic identity in the face of material change. The second mode of purposiveness is sense-making: an autopoietic system always has to make sense of the world so as to remain viable. Sense-making changes the physicochemical world into an environment of significance and valence, creating an *Umwelt* for the system. Sense-making, Varela maintains, is none other than intentionality in its minimal and original biological form.
>
> (Thompson 2007, 146-147)

Up until now, in our discussion about the autopoietic realization of organisms, we have focused almost exclusively on the internal, structural, and morphological features of such organisms, abstracted away from the environment from which they obtain the matter and energy needed to carry out their internal metabolic processes. In fact, I have even mentioned in section §1.1 that we would ignore, for the sake of methodology, two important aspects that are usually considered when one seeks a definition of life—the genetic-reproductive and the ecological—, and focus exclusively on the organism as individual. This seems to be aligned to the important role ascribed by

Varela and Maturana to the notion of autonomy during the formative years of autopoietic theory in the 1970s. In earlier versions of the theory, whereas the conditions for the organization of the living are given in terms of autopoiesis, the environment itself, which supplies the system with the components it needs in order to realize autopoiesis, is almost treated as a black box. The truth, however, is that this abstraction conceals some important aspects of the way autopoietic organization is made possible in real-world scenarios.

At least in a weaker, non-constitutive sense, autopoietic systems are undisputably coupled to their immediate environment. After all, they depend on it causally for the constant supply of matter and energy that is used in the maintenance of autopoiesis. Now, the teleonomy of autopoiesis is rather a matter of all-or-nothing: either the autopoietic system has everything it needs in order to work well enough, or else it collapses. This is because the metabolic processes that sustain autopoiesis are indeed of a strictly mechanical nature—as Varela and Maturana initially presumed—, and there are clear, easily quantifiable threshold conditions that need to be satisfied in order for them to be sustained. When interfacing with the external world, however, the autopoietic system needs a more fine-grained type of purposiveness, one that allows it to guide its behavior in a flexible manner, taking into account the characteristic inconsistencies of the external environment (e.g., changes in tides, seasons, prey population numbers, etc.).

In other words, differently from the self-directed form of purposiveness that is characteristic of autopoiesis, environmentally-directed purposiveness also requires the dimension of *normativity*. Thompson notices that: "Sense-making is normative, but the only norm that autopoiesis can provide is the all-or-nothing norm of self-continuance, not the graded norm implied by an organism actively seeking to improve its conditions of self-production (as when a bacterium swims up a sucrose gradient)" (Thompson 2007, 147).

Thompson is referring to the very often mentioned example of a unicellular organism in a liquid medium, which is led to swim in the direction where the sugar concentration is higher. Its

movement, although explainable in strictly biochemical terms, is simultaneously purposive. One could say that the bacterium operates within a very simple, primitive domain of experience where some form of normativity is already present: in this world, there is at least one criterion for distinguishing success—i.e., getting closer to the region where the sugar gradient is higher—from unsuccess—i.e., getting further away from it. But there is also, notice, an entire graded continuum in between these two possibilities that would satisfy, to different degrees, the conditions imposed by the bacterium's autopoietic imperative of self-preservation. At the ecological and reproductive levels, this is what makes possible the transition from the mechanistic description of life, where only causes and effects play roles, to a teleonomic or teleological description, where these can be reinterpreted in terms of means and ends (Thompson 2007, 76).

However, Ezequiel Di Paolo correctly pointed out that Varela and Weber's (2002) introduction of the concept of sense-making within the autopoietic project is not justified if considered only in relation to the theory's initial conceptual repertoire (Di Paolo 2005). Autopoiesis is, after all, a theory about the identity of living systems: that is, the way they get individuated from their surroundings and are capable of preserving their most distinctive structural features, but the introduction of sense-making must be accompanied by the introduction of a discussion about *adaptivity*. As noticed by Thompson, "[…] although autopoiesis is necessary for sense-making, it is not sufficient, but autopoiesis and adaptivity are jointly necessary and sufficient" (Thompson 2007, 148). After all, even if life on Earth first appeared in a specific environment, surrounded by an aqueous medium that reliably provided it with the materials and energy needed in order for autopoiesis to be preserved, it eventually flourished elsewhere and colonized almost the entire surface of the planet. That means, of course, that it had to adapt to an enormous variety of environmental conditions—different acidity levels, temperature ranges, atmospheric compositions, pressure gradients, light availability, nutrient concentrations, etc.—in order to survive and reproduce successfully.

In reality, organisms do not behave as passive autopoietic receptacles of a molecular and energetic influx that is constantly and invariably supplied by the environment, but in fact adapt to whatever environmental conditions they are subjected to, maximizing their chances of self-preservation and reproduction. Of course, by using this language, I am already ascribing teleological purposiveness and even agency to such organisms, and perhaps I should not—or at least not yet—, for everything that has been considered so far is still entirely compatible with a strictly teleonomic view of the organization of biological systems. What allows me to do that, however, and move from the reduced scope of teleonomy back to the openly teleological account of the organization of living systems, is the fact that Varela was interested in incorporating an investigation of sense-making from the first-person perspective into his autopoietic account. We could perhaps say, therefore, that in Varela's later work there is a vindication of teleology itself. This is because, differently from what happens in the case of teleonomic descriptions, purposiveness is not described in a way that is completely opaque to the conscious and experiential aspects of sense-making. Instead, it is simultaneously dealt with from the third-person perspective of the scientific inquiry and from the first-person perspective of phenomenological description. The methodological tool used when bridging this gap will be what Varela called neurophenomenology, the topic of our next section.

2.2. NEUROPHENOMENOLOGY

In our context, phenomenology can be thought of, figuratively, as the translation of the entire narrative about autopoiesis and sense-making from the third to the first-person perspective. In the specific context of human and higher-order mammal cognition, therefore, neurophenomenology would be the method that correlates insights obtained by means of a first-person investigation of consciousness—i.e., a phenomenological investigation—with those coming from the third-person approach that is characteristic of neuroscience. Since their main explanatory target is human cognition, it is obviously of particular interest for autopoietic enactivists doing philosophy of mind

and cognitive science to focus on organisms that have a relatively well-developed nervous system in their attempts at offering this coordinated, twofold approach. Still, the idea that a first-person perspective can and should complement strictly physicalist explanations of cognition already appears at the level of much more basic minds, perhaps even of unicellular organisms. Here, I will try to show how this correlation between investigations of consciousness from within and from without derives from what has been discussed in the previous sections. Let us start by what Varela says, in a 1997 article, about the relationship between the process of identity-formation carried out by autopoiesis and the establishment of a first-person reference point that could ground intentionality and conscious experience:

> i. An organism is fundamentally a self-affirming, identity-producing process based on autopoiesis.
>
> ii. A self-affirming identity establishes logically and operationally the reference point or perspective for sense-making and a domain of interactions.
>
> (Varela 1997)

At this point, we understand quite clearly what autopoietic enactivists mean by the identity-producing capacity of autopoiesis, but what does it mean to say that it is capable of establishing "[…] logically and operationally the reference point or perspective for sense-making and a domain of interactions"? Autopoiesis and sense-making together give rise to the individuation of a self—i.e., the organism—and its otherness—i.e., the environment. As we have seen, this self exists in a state of constant disequilibrium, for even though it is functionally capable of constant self-renovation, it lacks the raw materials and energy that are needed in order to do so. Its alterity, the environment, which was defined by the same autopoietic processes that individuated the self in the first place, is what provides the elements needed in order to prevent its collapse. However, in order to

acquire these elements, the organism needs to interact with the environment in very specific ways. These types of interactions were naturally selected for and reproduced, whereas others, which led to system collapse at the ontogenetic level and species extinction at the phylogenetic level, were not passed on.

The organism-system became, therefore, coupled to its environment in specific ways, since it reliably responds to specific stimuli in specific manners and, in consequence, alters the environment by means of its reaction. Thus, at least until an external cause disrupts this delicate balance, one could say that an ecological niche was formed. But, interestingly, for Varela and other autopoietic enactivists, the difference between this organism-environment coupling and other types of couplings that usually characterize inorganic systems is not simply one of degree. Although an organism-environment structural coupling can also, in principle, be described in strictly mechanistic terms, such a description would not capture all of the intricacies that interest us when studying the relationship between them.

From the organism's perspective, the environment is what biologist Jakob Johann von Uexküll (1934/2010) has called an *Umwelt*—a term that simply means "environment" in German, but acquired in his work the special connotation of being an organism's immediately meaningful domain of experience. Therefore, from the organism's perspective, the environment can be treated as already meaningful—although in a very restricted sense of the word—, because it seems to present an intrinsic normative scale of preferences. After all, certain of its features prove to be more relevant than others, in view of the ultimate autopoietic imperative of self-preservation: food is usually welcome, whereas predators are to be avoided. These act as points of attraction and repellence for the organism.

Additionally, Varela affirms that, when taken from the organism's perspective, the environment presents a "surplus of value." Let us try to elucidate what this means by resorting to our toy example of a bacterium swimming in a water-and-sugar solution. Suppose we assume the

standpoint of a strictly objectivist and reductionist biologist, which is ultimately interested in a merely physicochemical description of the environment, taking teleonomy as nothing more than a mostly irrelevant epiphenomenon of the chemical reactions and physical processes being investigated. In our research, we could surely describe a series of correlations between the chemical reactions and physical processes involving the organism-system, taken only as an aggregate of molecules, and the sucrose molecules. Surely, the tendency of the organism-system to behave in a certain way when confronted with a higher concentration of sucrose in a certain region of the solution could be explained in strictly mechanistic terms.

Now, there is a sense, however, in which such a scientific description of the phenomena in question could be considered unsatisfactory. When doing microbiology, the sucrose molecules suspended in an aqueous medium are not, after all, merely sucrose molecules: they are also *food for bacteria*. In other words, in an organism-environment coupling, the set of the properties of the coupling's individual elements is not coextensive with their set of properties as mere physicochemical systems—in other words, there is a surplus of value. As Thompson puts it:

> There is no significance in sucrose except when bacteria migrate up-gradient and metabolize sucrose molecules, thereby enhancing their autopoiesis. The food significance of sucrose is certainly not unrelated to the physics and chemistry of the situation; it depends on sucrose being able to form a gradient, traverse a cell membrane, and so on. But physics and chemistry alone do not suffice to reveal this significance. […] Whatever the organism encounters it must evaluate from the vantage point established by its self-affirming identity. At its simplest, this evaluation takes the form of the dual valence of attractive or repulsive and the correlative behaviors of approach or escape. In this way, sense-making "lays a new grid over the world: a ubiquitous scale of value."
>
> (Thompson 2007, 154)

The upshot is: we surely can look at the environment from a strictly reductionist perspective, wherein we describe chemical reactions and physical processes that produce specific organic

phenomena in strictly causal or correlational terms—i.e., in non-teleonomic and non-teleological terms. Still, we cannot do biology this way and, hence, neither can we fully understand life as a natural phenomenon. In order to carry out biological research, to understand and explain life, we need to be capable also of adopting the perspective of a living, embodied agent that is coupled with its environment in specific manners, and the scientific language used needs to be capable of expressing that. In illustrating this, Thompson resorts to philosopher Hans Jonas' maxim that "[…] life can only be known by life" (Jonas 1966/1982, 91). Within the physicalist language of a reductionist science, there are no reasons to call the sucrose molecule "food," for it is indeed food only in the context of the coupling between the bacterium, its aqueous solution, and the sugar gradient. And yet, at the level of biological explanations, we do need the very concept of food in order to explain a lot of important phenomena: ecosystem dynamics, predation and herbivory, trophic levels, symbiotic relationships, etc.

Before moving on to the next section, it might be opportune to discuss some noteworthy, but less pressing matters related to the neurophenomenological approach. One often neglected but quite important part of Varela, Rosch, and Thompson's proposal in *The Embodied Mind* consisted in the incorporation of insights from Buddhist philosophy and meditative practices into, respectively, the theory and the methodology of the cognitive sciences. Presumably, these topics are related to Varela's conclusion that it is necessary to introduce a neurophenomenological approach to the methodological repertoire of an autopoietic approach to cognition, if one wanted to study the process of sense-making both from within and without. Whereas it is usually regarded as bad practice to rely on introspection and one's own intuitions to formulate scientific theories, the general idea seems to be that an appropriate training in meditative practices grants one with a certain degree of expertise or familiarity with the phenomenological. Claims to that effect have been met with skepticism or reticence by many in the cognitive sciences and philosophy of mind. At the same time, the general perception seems to be that they are not an essential part of the wider project sketched in enactivism's manifesto. In any case, it seems unlikely that these features would be

inherited by a successfully unified enactivist theory, one that would have overcome the sensorimotor-autopoietic dichotomy.

2.3. MIND-LIFE CONTINUITY

Because of this intrinsic teleological dimension of the autopoietic explanation of living systems' organization, autopoietic enactivists contend that the autopoietic theory of life, if supplemented by the notion of sense-making, also entails an autopoietic theory of cognition and, possibly, even of mentality as a whole. This idea is called the strong continuity thesis. It was originally presented by Peter Godfrey-Smith (1996, 320) and it is succinctly formulated by Andy Clark in the following terms:

> The thesis of strong continuity would be true if, for example, the basic concepts needed to understand the organization of life turned out to be self-organization, collective dynamics, circular causal processes, autopoiesis, etc., and if *those very same concepts and constructs* turned out to be central to a proper scientific understanding of mind.
>
> (Clark 2001, 118)

In other words, the thesis, which has also been termed biopsychism, puts forward the view that life and mind are profoundly interrelated, representing two facets of a single, unified phenomenon. The relationship between them, according to this view, is one of coextension: every living system presents some degree of mentality, even if in a quite incipient or rudimentary form; conversely, and with equal force, all systems that can be properly described as minded are also living systems. Take again the example of the bacterium swimming to the area of higher sugar concentration in a liquid gradient. For autopoietic enactivists, this activity must be classified as cognitive already. This view challenges the conventional association of cognition with brain-bound

processes—an association that makes the attribution of cognitive capacities to organisms lacking a central nervous system appear counterintuitive or even implausible to many.

It is important to note how the biopsychist hypothesis distinguishes itself from panpsychism—i.e., the hypothesis that consciousness is a fundamental property of all matter and, therefore, present at least to some degree in everything that is material. The mind-life continuity thesis, or biopsychism, offers a far more constrained and biologically grounded account, decisively rejecting the notion that mentality is an indiscriminate property of all physical systems. Instead, it argues that mind is intrinsically and exclusively linked to a very specific kind of material organization: autopoiesis, the organization of life itself. Hence, it is not matter *per se* that is minded, but matter that has become organized into a self-producing and self-maintaining autopoietic system, and mentality emerges precisely from the complex dynamics that allow such a system to sustain its organizational pattern in the face of entropy. Thus, contrary to panpsychism, biopsychism anchors mentality and cognition firmly in the concrete, goal-directed activities of living agents, whose very existence depends on their capacity to make sense of their world in order to sustain themselves.

An important implication of this thesis is that there can be no absolute primacy granted to neural tissue in the explanation of cognition, a conclusion that radically decenters the brain from its long-held throne as the seat of the mind. This perspective aligns closely, of course, with the arguments presented by Alva Noë and J. Kevin O'Regan against the primacy of the neural, discussed in the beginning of §2 of the previous chapter (Noë and O'Regan 2001). In this view, the nervous system should be understood as a highly specialized and powerful organ that has evolved to facilitate and modulate the dynamic sensorimotor coupling between an already living, autonomous organism and its environment.

Hence, it surely does play a privileged role in making viable and sustainable the coupling between organism and environment, since the capacity for environmental adaptation is, as we have seen, an essential property of living and cognizing systems in general. This, however, does not

mean that it should be treated as a *necessary condition* for cognition to occur in the first place, as the bacterium-in-a-sugary-medium example tries to illustrate: it is a mediator, a facilitator, and a potentiator of cognitive activity, but not its ultimate source. The nervous system is a tool that allows for more complex, flexible, and far-reaching forms of sense-making, but the fundamental capacity for sense-making itself is a property of the living organization as a whole, an organization that precedes the evolution of neurons by billions of years. The relevant explanatory unit, however, continues to be brain-body-environment couplings.

This, of course, seems to suggest a close relationship between enactivism and the two strong E-theories—i.e., strong embodied cognition and extended cognition—, which also attribute constitutive powers in cognitive activity to extracerebral matter. Thus, considering that the three of them attribute to brain-body or brain-body-environment couplings a central explanatory role in their conception of a new science of cognition, it seems natural to conceive of these frameworks as overlapping in significant points. As we will see in the next section, however, it is still not entirely clear that the relationship between enactivism and extended cognition, in particular, is indeed a sustainable one. In fact, there are *prima facie* good reasons to suspect a significant degree of incompatibility between them. After all, enactivism introduces the notion that a boundary is needed in order to differentiate a living system from its environment and explain its resistance to entropic pull, whereas what many extended cognition theorists seem to defend is precisely the idea that cognitive boundaries are mere heuristic abstractions.

3. AUTOPOIETIC ENACTION AND FUNCTIONALISM

It is now time to elucidate and resolve two important points of tension within the framework of autopoietic enactivism as presented in the previous sections. The first and perhaps most pressing of them arises from an apparent clash between the foundational premises of autopoietic theory and the claims of strong extended cognition, sometimes considered not only to be compatible with,

but essential to the enactivist view. As we have seen, autopoietic theory, in its initial and most rigorous formulation, centers on the biological cell as the paradigmatic living system, explaining life through a process in which a network of reactions produces a boundary—i.e., in a cell, a semipermeable membrane—that, in turn, encloses and sustains that very same network. This conception, which emphasizes a clear and materially demarcated boundary as a necessary condition for life, seems *prima facie* irreconcilable with the strong extended cognition thesis, which forcefully argues that cognitive processes can and often do constitutively include parts of the external environment, thereby dissolving any prospect of a strict boundary between the cognitive agent and the world. If a person's thoughts can literally be constituted by scribbles in a notebook, then the autopoietic insistence on a self-enclosed, metabolically-defined unity appears to be fundamentally challenged.

In their initial efforts to scale the theory beyond unicellular organisms, Maturana and Varela introduced the distinction between first-order and second-order autopoietic systems as their primary means of accounting for more complex forms of life, as we have seen in §1.2 above. First-order autopoiesis pertains to unicellular organisms, which neatly instantiate the theory's core criteria of self-production and physical boundedness through their cellular membranes. Second-order autopoiesis, by contrast, was their attempt to apply this same logic to multicellular organisms, conceiving of them as composite systems, or metacellulars—as the authors called them—, which are composed of vast networks of first-order autopoietic systems interacting in a coordinated, mutually sustaining manner. This theoretical maneuver was not merely an afterthought but an essential step if the theory was to have any relevance for understanding a significant portion of life on Earth, acknowledging that complex organisms are not merely scaled-up cells but integrated communities of cells.

However, this very extension, while necessary, inadvertently creates a significant internal crisis for the theory. While the boundary of a unicellular organism is clearly and unambiguously

defined by its lipidic membrane, the corresponding boundary of a second-order, multicellular organism is far from determinate. Again, one might be tempted to nominate the skin as the obvious candidate, yet, as I mentioned before, the skin fails to perform many of the crucial functions of the autopoietic boundary, such as selectively incorporating matter and energy. Yet these systems are internal, convoluted, and topologically complex, lacking the clear inside/outside demarcation that the theory's initial formulation demanded. This ambiguity threatens the very definitional rigor that autopoiesis was originally intended to provide, making the designation of multicellular organisms as "second-order autopoietic systems" seem uncomfortably *ad hoc* and straining the central concepts to a breaking point. The problem is not merely terminological; it strikes at the conceptual heart of the theory, raising profound doubts about whether the strict criteria of autopoiesis can adequately capture the organizational logic of complex living systems without losing their explanatory force.

Yet, it is precisely this crisis, this failure of a direct and literal extension, that motivates and ultimately necessitates a crucial conceptual shift within the theory's evolution: the move from the strict, structurally-defined notion of autopoiesis to the more general and flexible, functionally-defined concept of autonomy. Autopoiesis, to reiterate, is defined with structural and material precision: it requires a physical, self-produced boundary that topographically encloses the very network of production processes that created it. Autonomy, by contrast, is defined in purely functional and organizational terms: it characterizes any system that exhibits organizational closure, a condition wherein the network of processes constituting the system recursively depends on and regenerates itself, thereby defining the system's identity and distinguishing it from its environment, even in the complete absence of a single, continuous, physically demarcated border.

This distinction allows the theory to accommodate systems whose boundaries are not strictly physical but are instead constituted by the coherence and self-maintenance of their functional and organizational relations. This shift from a structural to a functional definition is the key

that unlocks the puzzle. By applying this more powerful conceptual tool, the apparent conflict with extended cognition dissolves. A coupled human-notebook system, for instance, is manifestly not autopoietic in the strict sense, for it lacks a single, self-produced physical boundary that encloses both the human and the notebook in a shared metabolic web. It can, however, be robustly and coherently defined as an autonomous system. Its boundary is not the skin, but the functional and organizational closure of the coupled processes of reading, writing, thinking, and remembering that sustain a particular cognitive task. This reframing resolves the tension not by rejecting one thesis for the other, but by showing how extended cognitive systems can be understood as a legitimate, if higher-order, manifestation of the core enactive principle of self-constituting identity, thereby accommodating the insights of extended cognition without sacrificing a commitment to organizational closure. It is important to notice, however, that this is a significant step in the direction of framing autopoietic theory itself as a type of functionalist theory—if not in the classical sense of the term, at least in an extended sense.

With this in mind, it is opportune to now consider a second major tension to be addressed, which concerns precisely the nature of the relationship between enactivism and functionalism. This debate is often framed by a pervasive but, I contend, flawed misconception that these two theoretical frameworks are fundamentally and irreconcilably antagonistic. The standard incompatibility argument typically proceeds along the following logical steps: it is stated that enactivism, particularly in its autopoietic form, is committed to the idea that cognitive phenomena require a living and adaptive organism to take place. Then, it is stated that functionalism, in its classical form, is defined by its commitment to substrate neutrality, the idea that mental states are defined by their causal roles and can therefore be realized by any physical substrate that supports the requisite functional organization. From these two premises, the flawed conclusion is drawn: enactivism, therefore, with its emphasis on the specific material organization of life, must be an anti-functionalist view. This argument, however, rests on a profound misunderstanding of both the nature of autopoietic

enactivism and the broader possibilities of functionalism itself, creating a false dichotomy that obscures a deeper compatibility.

The resolution to this apparent conflict lies in recognizing that autopoietic enactivism is not the antagonist of functionalism but is, in fact, best understood as a more specific and sophisticated form of functionalism—a biologically-grounded and organizationally-specific type of functionalism. In this framework, the function being realized is not an abstract, disembodied mental state or a syntactic operation within a computational system, but rather the concrete, organizational, and dynamic pattern of "life-mind" itself. Cognition, from this perspective, is the function of maintaining a living identity through sense-making interactions with an environment. Consequently, the core functionalist principle of multiple realizability is fully preserved, but it applies not to isolated mental states but to the autopoietic functional organization as a whole. This means that any system, regardless of its specific material makeup—be it carbon-based, silicon-based, or something else entirely—that successfully realizes the functional organization of a self-producing, adaptive, and autonomous system would, by definition, be considered both alive and cognitive. This nuanced perspective thereby gracefully leaves the door open for the principled possibility of artificial life and artificial cognition, provided that the requisite organizational and functional patterns of autonomy and sense-making are properly instantiated. Far from rejecting functionalism, autopoietic enactivism enriches it, grounding it in the concrete biological principles of self-organization and providing it with a more robust, less abstract, and scientifically tractable subject matter.

Synthesizing the resolutions to these two central tensions reveals a profound conceptual unity at the heart of autopoietic enactivism's theoretical development. The crucial move used to solve the first tension—that is, the conceptual shift from the strict, physical requirements of autopoiesis to the more general, organizational pattern of autonomy—is itself a quintessentially functionalist move. It is an explicit prioritization of function and organization over any specific material implementation or structure, demonstrating that the theory's own internal logic pushes it towards

a functionalist interpretation in order to solve its own scaling problems. However, it is absolutely critical to emphasize that this functionalist expansion does not, in any way, render the core biological concept of autopoiesis irrelevant or obsolete. On the contrary, it simultaneously affirms and reinforces its foundational and indispensable role. The reason for this is that higher-order autonomous systems, whether they be second-order autopoietic systems like multicellular organisms or even more abstract autonomous systems like a person-tool coupling, are not free-floating, disembodied functional structures. They are, without exception, materially grounded in, and emergent from, the vast, underlying network of first-order autopoietic systems—the living cells—that constitute and sustain them.

The entire explanatory power of the life-mind continuity thesis, which is the central pillar of autopoietic enactivism, flows directly from the basic, material, metabolic, self-producing activity of the cell—or some other hypothetical structure that is functionally equivalent to it. Autonomy is the powerful and necessary conceptual tool that allows us to understand how this fundamental principle of life can scale up to explain the existence of complex organisms and even their technologically-mediated cognitive extensions, but this explanatory reach is possible only because the concept of autonomy never fully detaches from its ultimate biological and material foundation. In this way, autopoietic enactivism maintains its conceptual rigor and its profound explanatory depth, successfully navigating the challenges posed by both extended cognition and functionalism, and in doing so, securing its place as a uniquely coherent and compelling framework within the broader landscape of the philosophy of mind and cognitive science.

4. Conclusion

This chapter was divided in two parts, covering, respectively, the autopoietic theory of life (§1) and the autopoietic theory of cognition (§2). Thus, we started by tracing the conceptual development of autopoietic theory in §1, beginning with its foundational concern: the definition of life. The

autopoietic theory developed by Maturana and Varela defines living systems as organizationally closed, self-producing, and self-distinguishing unities. This definition emphasizes not the material constituents of life but the functional organization that sustains it. Drawing on cybernetics and systems theory, the notion of autopoiesis describes life as a circular, self-maintaining process, initially modeled on the basic unit of the cell but later extended—albeit not without some theoretical difficulties—to multicellular organisms. In §2, we examined how this autopoietic framework was extended into a theory of cognition, giving rise to what has become known as autopoietic enactivism. This theoretical shift was largely spearheaded by Varela, who introduced the concept of sense-making to capture the purposive, norm-sensitive interactions between organism and environment.

In doing so, Varela moved beyond the mechanistic interpretation of autopoiesis toward a richer, more teleological view in which cognition is not confined to brain-bound processes but is instead distributed across the organism's embodied interactions. The notion of sense-making reframes cognition as an intrinsically normative activity grounded in the organism's self-maintaining identity and environmental embeddedness. This leads to the thesis of biopsychism, which posits a deep continuity between life and mind. This idea suggests both that all living systems, insofar as they are autopoietic and adaptive, already exhibit a minimal form of cognition; and also that cognitive phenomena are necessarily tied to living systems. We have also seen, in §3, how this position—initial intuitions notwithstanding—makes autopoietic enactivism even closer in spirit to functionalism, provided that we apply the functionalist framework not only to cognitive phenomena, but also to the notion of life itself. In this view, life may be multiply realizable and, therefore—provided that the life-mind continuity thesis holds indeed—, so may be cognition.

In the previous chapter, I have identified some key limitations in sensorimotor enactivism's approach to the normative and purposive dimensions of cognition. While that theory foregrounds the embodied and interactive nature of perception, it often remains silent on what grounds the meaningfulness of those interactions. I hope to have shown here that the autopoietic variety of

enactivism, by contrast, addresses these shortcomings by grounding cognition in the very organization of life itself. Through the concept of autopoiesis, it provides a robust biological foundation for sense-making, offering an account of how intrinsic normativity emerges from the self-maintaining dynamics of living systems.

Finally, while autopoietic enactivism offers a promising foundation for grounding cognition in the biological autonomy of living systems, several important questions remain open in light of the broader aim of this study: assessing the compatibility of enactivism with the central presuppositions of traditional cognitive science—namely representationalism, computationalism, and functionalism. As this chapter has shown, autopoietic enactivism can be interpreted as a distinctive form of functionalism, one grounded not in input-output mappings or syntactic operations, but in the organism's dynamic organization and self-maintaining structure. However, whether this organizational functionalism is ultimately compatible with traditional accounts of representation and computation—or whether it points to a fundamentally different explanatory framework—remains to be seen. The challenge, then, is to determine whether enactivism revises the conceptual foundations of cognitive science or simply rearticulates them in novel biological and systemic terms.

To address these open questions, the next two chapters of this dissertation turn to the relationship between enactivism and the other two pillars of mainstream cognitive science: representationalism and computationalism. Chapter IV focuses on the role of mental representation, investigating whether enactive approaches can do justice to the explanatory roles attributed to content without invoking internal representations. This discussion will also introduce radical enactivism, a more deflationary strand of the enactivist tradition that rejects representational posits altogether and seeks to reframe cognition in purely relational and interactional terms. Chapter V will then examine the compatibility of enactivist theory with contemporary accounts of computation, including mechanistic and non-representational models. Together, these chapters aim to clarify whether enactivism constitutes a reformist or a revolutionary stance within cognitive science—and

to what extent it can coexist with, or must ultimately reject, the core assumptions of its predecessor paradigms.

PART II: ENACTIVISM, REPRESENTATION, AND COMPUTATION

IV. ENACTION AND REPRESENTATION

In the first chapter, we have concluded that enactivism differs from other 4E approaches to the extent that it necessarily takes the coupling between brain, body, and environment as constitutive of cognitive activity, and not merely as causally relevant to it. An additional feature that is very often assigned wholesale and indistinctively to the entire family of enactivist theories is antirepresentationalism. In philosophy of mind and the cognitive sciences, antirepresentationalism is the view that proper explanations of perception and cognition should not resort to the notion of mental representation, as it does not adequately account for what actually takes place during the course of cognitive and perceptual processes. There are two different levels at which one can be an antirepresentationalist, and I shall call their corresponding views "moderate" and "radical" antirepresentationalism.

A moderate antirepresentationalist might be willing to accept mental representation as a useful theoretical posit in cognitive models, even if it does not actually track anything that participates constitutively in our cognitive and perceptual processes. On the other hand, a radical antirepresentationalist will usually deny mental representations both the status of entities, as well as their explanatory usefulness as theoretical posits within these models. Here, we shall be dealing mostly with the self-declared radical variety of antirepresentationalism, such as the one defended by Daniel Hutto and Erik Myin's radical enactive cognition, or by Anthony Chemero's radical embodied cognitive science (Hutto and Myin 2013; Chemero 2009). For the sake of brevity, I will use Hutto and Myin's term REC—an acronym that stands for "radical enactive (or embodied) cognition"—to refer to the whole family of projects that are critical of mental representation. Whenever this shared use cloaks important differences in detail among the authors' critiques, I will disambiguate the variant in question by referencing the author nominally.

In §1, we will briefly consider some of the main reasons why antirepresentationalism has become such an important feature of enactive and embodied theories of perception, cognition, and action in general. Moreover, we shall also see that, even if there is a sense in which enactivist research programs in general usually imply a deflationary view of the explanatory power of mental representations, radical antirepresentationalism is not strictly a requirement in order for a theory to be properly categorized as enactivist, as illustrated by the role played by the notions of "knowledge of sensorimotor contingencies," in sensorimotor enactivism (see Ch. 2). While discussing the relationship between enaction and representation, I will also correct a frequently made, but false assumption: namely, that REC is a type of enactivist theory, a theoretical rival of sorts to the sensorimotor and autopoietic varieties of enactivism. This, we will see, is not even an aspiration of REC's defenders themselves, who tend to see their own endeavor mostly as a critical, metatheoretical comment on the role assigned to the notion of representation within those two branches of enactivism.

After that, in §2, I will reconstruct one of the most central arguments presented by defenders of antirepresentationalism in enactivism, which was called by Hutto and Myin the "hard problem of content." Briefly stated, Hutto and Myin argue that, up until now, all attempts at naturalizing mental representation—or, a synonym for them, mental content—were grounded on the notion of "information-as-covariance," with the explicit aim of arriving, by means of this simpler, more naturalistically acceptable notion, at the idea of "information-as-content," a widely used and yet ambiguous theoretical posit in the cognitive sciences and related fields. However, information-as-covariance, according to them, lacks the most important semantic property of information-as-content: satisfaction conditions. Hence, they arrive at the principle that defines the hard problem of content: COVARIANCE DOES NOT CONSTITUTE CONTENT. Hutto and Myin's proposed radical solution to this paradox is simply the denial of any role whatsoever to representation or content at the level of what they call "basic minds," a domain which seems to involve mostly prelinguistic action and perception, and perhaps cognition that is not yet language-dependent.

In §3, I will distinguish a few important concepts whose differences are sometimes neglected in the literature, and whose conflation breeds a lot of confusion in the current debate about the ontological status of mental representations. They are symbolic and neural representations, personal and subpersonal representations, and, finally, representational vehicle and representational content. We will see how these distinctions help us better understand what is at stake in the different types of debates around the topic of mental representation, as well as the possible theoretical stances in each one of them. In §4, I will defend the view that certain instrumentalist approaches to representation, as exemplified by Frances Egan's deflationary account of mental representation, can avoid the hard problem of content. Thus, they can be made entirely compatible with both of the two main varieties of enactivist theory. I will explain how Egan's instrumentalist perspective avoids metaphysical commitments by treating representational content as a heuristic device rather than as something with intrinsic semantic value, illustrating how enactivists can adopt her perspective without undermining their core commitments. Finally, in §5, the conclusion, I will argue that this perspective dissolves Hutto and Myin's hard problem of content by reframing representational talk as a practical tool for scientific modeling, rather than as a substantive claim about the nature of cognition itself.

1. ANTIREPRESENTATIONALISM AND 4E COGNITION THEORIES

Although most often associated with works published during the last two decades by authors such as Hutto, Myin, and Chemero, the core tenets of antirepresentationalist, radical embodied or enactive cognition were already put forward in Andy Clark's seminal 1996 book *Being There*, where he briefly entertained an already quite radical view of embodied cognition:

> Thesis of Radical Embodied Cognition: Structured, symbolic, representational, and computational views of cognition are mistaken. Embodied cognition is best studied by means of noncomputational

and non-representational ideas and explanatory schemes involving, e.g., the tools of Dynamical Systems Theory.

(Clark 1996, 148)

Clark himself was not committed to this type of antirepresentationalism, which was being entertained as a mere theoretical possibility in his book, but many philosophers and cognitive scientists with sympathies towards enactivism seem to be attracted by this picture. A reasonable first question would be, therefore: Why does not only enactivism, but 4E theories in general, seem to gravitate towards antirepresentationalism? (Hutto and Myin 2013, Ch. 4) The causes for this widespread sympathy are diverse, but perhaps there are a few of them that can be more easily identified: i) a late reassessment of the merits of James J. Gibson's ecological psychology and the corresponding theory of direct perception from the early 1990s onwards (Gibson 1966, 1979); ii) developments in robotics that showed many solutions to simple tasks to be much more complicated and demanding if implemented in terms of computational processing than by means of a readjustment of the coupling between the robot and its environment (Brooks 1991); iii) the application of dynamical systems theory to cognitive processes (Thelen and Smith 1993, 1994), which allowed cognitive scientists to model many complex phenomena, particularly in the fields of action and perception, while making no reference whatsoever to mental representations.

Still, as we have seen in the previous chapter, among the various theories within the 4E-cognition family, there are some that are overtly committed to the notion of representational content, so that antirepresentationalism cannot be taken as a distinctive feature of embodied and extended approaches to cognition in general. Our first example was Goldman and de Vignemont's recent attempts at making sense of mental content that is interoceptive in origin, but has been coopted for a secondary, usually higher-level cognitive function, in terms of B-formatted representations (Goldman and de Vignemont 2009; Gallese and Sinigaglia 2011; Alsmith and de Vignemont 2012; Goldman 2014). An additional example is Clark's project of establishing connections

between his extended mind hypothesis and the burgeoning field of predictive processing, which relies significantly on the idea of mental representation. After all, in perception, a prediction must be accounted for in terms of a representation of what is expected to be perceived, which is then compared to the actual perceptual input and corrected by means of backwards propagation (Clark 2016; Clark 2023).

It could be argued, however, that although embodied and extended cognition by themselves do not imply any significant commitment to antirepresentationalism, a truly coherent form of enactivism does. Although more plausible, given the relatively widespread hostility towards representational talk among enactivists, this contention is still questionable. Some of the currently leading figures in sensorimotor enactivism, such as Alva Noë, have expressed serious doubts concerning the purpose and the necessity of REC's eliminativist purge (Noë 2021). Similarly, as recognized by radical enactivists themselves, some of the most important tenets of autopoietic enactivism—such as the notion of sensemaking—may imply at least a partial acceptance of the core principles of representationalism (Hutto and Myin 2013, 32–36).

This hesitation in fully adopting the antirepresentationalist stance is what REC aims to correct within enactivism at large. Although it is often treated as a third member of the enactivist theoretical family—or even as a rival theory to its sensorimotor and autopoietic counterparts—this is simply not the case. It should, in fact, be seen as a critical metatheoretical endeavor, targeting both traditional computationalist accounts and what is perceived by its defenders as the residue that decades of representationalist hegemony have left on the two main enactivist theoretical frameworks. Hence, Hutto, for instance, recognizes that: "REC is not an alternative version of enactivism with distinct explanatory tools in its own right. […] Its analyses and arguments are designed to cleanse, purify, strengthen, and unify a whole set of existing anti-representational offerings. REC's aim is *to radicalize* existing versions of enactivism and related explanatory accounts through a process of philosophical clarification" (Hutto 2017).

As we will see in this chapter, therefore, instead of proposing its own alternative to representational theories within the various subdomains of the cognitive sciences, REC limits itself to comment on the uses and misuses of the representationalist jargon by cognitive scientists and philosophers. The task would indeed be Sisyphean if it had more traditional accounts as its main target—for the use of the concepts of "representation" and "content" is, for better or worse, so deeply rooted in cognitivist discourse. Luckily, however, with the increasing acceptance of enactivist theories into the mainstream of the cognitive sciences, a more amenable target can be chosen.

2. THE HARD PROBLEM OF CONTENT

In the previous section, I presented an admittedly limited summary of the main reasons for the widespread adoption of strong antirepresentationalism among defenders of embodied, extended, and enactive cognition. Those were mostly related to diverse and scattered insights afforded by developments in the cognitive sciences and robotics, particularly in the 1990s. A true motivation for endorsing full-fledged antirepresentationalism, however, must be complemented with a proper argument, such as the one given by Hutto and Myin in their 2013 book *Radicalizing Enactivism*. Such an argument would constitute a perfect example of the type of clarificatory work REC theorists see as the core of their intellectual endeavor.

With that in mind, they present the so-called "hard problem of content" as a more definitive blow against representationalist accounts. It is introduced along the following lines: It is a widely shared view among theorists of content-involving cognition that an explanation of mental content must be given in naturalistic terms, which means that there must be no explanatory gap between a strictly external, non-interpretative informational relation and the internal, representational symbol. Any such naturalistic explanation of content-involving cognition must therefore depart from the concept of information as an objective, mind-independent property or relation between objects in the world. For Fred Dretske, for instance, one of the first and most notorious

defenders of these views, what serves as the raw material for the subsequent emergence of meaning as a symbolic relation is precisely this more primitive, naturalized type of information, which "[…] do[es] not require or in any way presuppose[s] interpretative processes" (1981, 1). Hutto and Myin complement this insight with the following passage by Pierre Jacob:

> […] The relevant notion of information at stake in informational semantics is the notion involved in many areas of scientific investigation as when it is said that a footprint or a fingerprint carries information about the individual whose footprint or fingerprint it is. In this sense, it may also be said that a fossil carries information about a past organism. The number of tree rings in a tree trunk carries information about the age of the tree.
>
> (Jacob 1997, 45)

For Hutto and Myin, these two views supposedly express the mainstream scientific—and perhaps also philosophical—stance on the nature of information and the role it plays as the raw material of cognition. The notion of information-as-covariance, which is implicit in them, can be summarized in the following manner: "[…] s's being F 'carries information about' t's being H *iff* the occurrence of these states of affairs covary lawfully, or reliably enough" (Hutto and Myin 2013, 66). For Hutto and Myin, however, the main problem is that all naturalistic explanations of representation currently available depend on a very peculiar concept of content, namely, informational content that has satisfaction conditions. By "satisfaction conditions" they mean, as does a significant part of the relevant literature, informational content that can be assessed as either true or false, in the case of beliefs and propositional thought; successful or unsuccessful, in the case of actions and intentions; accurate or inaccurate, in the case of perceptions; and similarly for other putative mental states.

Information-as-covariance, however, lacks these properties. After all, the number of rings in a tree trunk may covary with its age, but this does not mean that it says or conveys anything true

about it. These can only be found in information-as-covariance's "richer cousin," semantic or intentional information. Hutto and Myin call the variation of the concept of information that has these properties "information-as-content" and, according to them, we have no grounds to derive it from information-as-covariance. They introduce, therefore, the principle COVARIANCE DOES NOT CONSTITUTE CONTENT, and use it in their formulation of the hard problem of content: if we want a naturalistic explanation of basic cognition and we are not allowed to derive information-as-content from the only available naturalistic concept of information—i.e., information-as-covariance—, what is it that allows us to introduce information-as-content in our explanation in the first place? (Hutto and Myin 2013, 67)

Clark was not only responsible for presenting the possibility of a radically antirepresentationalist stance in the philosophy of cognitive science, but also probably the first philosopher to discuss in detail the challenges to be faced by defenders of such a position. These challenges are mostly related to what he called "representation-hungry domains" of cognition: anticipation, absence, non-existence, and counterfactuality. As noted by Jan Degenaar and Erik Myin (2014), the most common types of criticisms from representation-hungry domains can be divided into two main groups: criticisms from absence, when "[…] cognitive activity in domains involving the absent necessitates mental representations as *explanantia*," and criticisms from abstraction, when these are necessitated by "[…] cognitive activity in domains involving the abstract." The extent to which mental activity related to these domains is language-dependent is debatable, but it seems intuitive to think that, at the very least, a significant part of our reasonings regarding absent or abstract entities could not happen in non-linguistic minds (Degenaar and Myin 2014, 12). As it stands, therefore, Hutto and Myin's REC attempts to offer a non-representational explanation only of what they call "basic minds," which is a category involving mostly action and perception:

> Catching a swirling leaf, finding one's way through unfamiliar terrain, attending and keeping track of another's gaze, watching the sun rising at the horizon—the vast sea of what humans do and

> experience is best understood by appealing to dynamically unfolding, situated embodied interactions and engagements with worldly offerings. […] Where we find such familiar activity, we find basic minds.
>
> (Hutto and Myin 2013, x, emphases added)

In fact, as we have seen in Chapter II, scaling up from existing dynamic models of action and perception to explain all of cognition is still very far from reality, and therefore Hutto and Myin are forced to recognize their project's scope limitations:

> Let us be clear. In pressing for REC, we do not say that CIC [Content-Involving Cognition] is never true. We do not say that cognition is never informed by or never involves content. We have no truck with that claim. We are not advancing Really Radical Enactive or Embodied Cognition as a thesis about the nature of all minds. Some cognitive activity—plausibly, that associated with and dependent upon the mastery of language—surely involves content.
>
> (Hutto and Myin 2013, xviii, emphases added)

A number of scholars, however, when assessing Hutto and Myin's work, have expressed disapproval precisely about what has been called the problem of scaling up—or else, the problem of continuity. Some believe these issues disavow the whole of REC's project (e.g., Miłkowski 2015, Matthen 2014); others suggest that it shows Hutto and Myin are not nearly as radical as they claim to be (e.g., Moyal-Sharrock 2019, van den Herik 2020). Marcin Miłkowski (2015), for instance, criticizes them for what he sees as a double standard in their attitude towards "basic minds," on the one hand, and "linguistic minds," on the other. According to him, extending Hutto and Myin's approach to linguistic minds would result in "semantic nihilism, which states that nothing ever is true or false, or veridical" (Miłkowski 2015, 73). In other words, the view that language has no satisfaction conditions. He proceeds to argue that "It is a minimal requirement for rational

argumentation in philosophy; one has to assume that one's statements can be truth-bearers. If they cannot have any truth-value, then it's no longer philosophy" (Miłkowski 2015, 74).

Miłkowski's ultimatum to Hutto and Myin is then given in the following terms: "Language should not be given any special treatment, so for consistency, Hutto and Myin should insist on representationalism, unless they want to embrace semantic nihilism" (2015, 85). Similar criticisms have been raised by other authors, even if with completely different agendas, who believe Hutto and Myin to be contradicting themselves when they claim that "[…] content does not exist in the natural world, and cannot be meaningfully explained in terms of how an organism makes use of a putatively representational entity, but language-dependent cognition is, nonetheless, contentful" (Harvey 2015, 113). Along the same lines, others add that their willingness to reintroduce content at higher levels of cognitive activity without first offering a solution to their own hard problem undermines their criticisms of content-involving accounts of cognition from more traditional sources. After all, if radical enactivism can set aside the hard problem of content whenever this suits its purposes, then any competing theory should be allowed to do the same.

It is important to notice that, *pace* Miłkowski, it is debatable whether eliminativism about linguistic content is indeed universally rejected, since the position has been entertained even by an authoritative figure such as Noam Chomsky, as well as by some influential names within analytic philosophy of mind (Chomsky 2003, 274; Collins 2007; Churchland 1981). In view of this, the question of why Hutto and Myin commit to representationalism at the level of linguistic minds might indeed seem puzzling at first, since it seems to introduce either a flagrant contradiction or at least a significant explanatory gap in their account.

The contradiction, of course, is this: If covariance does not yield content for basic minds, how come it does for linguistic minds? The explanatory gap, on the other hand, presupposes the existence of a satisfactory answer to that question, which is not yet available in radical enactivist literature. At least in the case of Hutto, this is partially illuminated by considering other aspects of

his philosophy, specifically his account of reasons in terms of folk psychological narratives (Hutto 2008). Apparently, then, the tentative answer would involve the status of language as a public representational device and its scaffolding by our evolutionarily constrained sociocultural practices. Still, considering the boldness of REC's claims, I do believe a more in-depth exploration of this topics is warranted. I will not discuss these ideas in detail here, however, for this would lead us too far from the main scope of this chapter. Instead, I will now show how Hutto and Myin's hard problem of content can be relatively easily overcome if we resort to a more nuanced view of the nature of mental representational content, and the explanatory role it plays in cognitive science.

3. PERSONAL AND SUBPERSONAL REPRESENTATION

Perhaps a substantial part of the disagreement concerning the nature of mental representation and content hinges upon our lack of clarity about the concepts themselves. As currently used in the study of the mind, after all, "mental representation" appears to be polysemous: often, it seems that what is referred to by the expression are abstract symbols with determinate content that are syntactically manipulated by an amodal, central cognitive processor (see Ramsey 2007, 8). This type of mental representation would be, therefore, an abstract mental entity that fits perfectly well within the role of an abstract symbol in Mentalese. Let us call this type of representational entity "*symbolic representations*."

Although usually conceived of in terms of encapsulated cognition, isolated from action and perception, it is not difficult to imagine that many of the main characteristics of symbolic representations do in fact also apply to their counterpart in perceptual processing—that is, that perception could also be treated, in principle, as something analogous to symbol manipulation. What distinguishes these two types of representation from others, which will be discussed subsequently, is the fact that they often appear in psychological explanations restricted to the personal level. In philosophy of mind and the cognitive sciences, the personal level of explanation is the one that

deals with the individual's conscious experiences, beliefs, desires, and actions, and it is often opposed to the subpersonal level, which is concerned with the underlying mechanisms—such as neural processes—that enable cognition and perception to take place, with no reference to intentionality or consciousness. Thus, I will use here the more encompassing expression "*personal-level representations*" to refer both to instances of paradigmatically symbolic mental representations and other types of structurally-similar mental representations that appear in personal-level explanations of mental phenomena, such as perceptions and intentions.

At other times, however, it seems that what is referred to by the expression "mental representation" are the neural correlates of our mental states—that is, neural networks or specific patterns of activation that are correlated with what is being represented in a variety of different manners. What is being represented can be, for instance, either a distal source of sensory stimuli in the environment, in the case of perceptual representations; an individual's past experiences, in the case of episodic memory; as well as an intended course of action, in the case of motor representations; among other possibilities. Let us call this concept "*neural representations*." Now, unlike symbolic representations, which tend to feature in personal-level explanations, neural representations function at the subpersonal level, where cognitive processes are explained without invoking conscious experience or intentional states.

By framing neural representations as part of subpersonal mechanisms, we can focus more precisely on the underlying neural and physiological processes that support cognition, perception, and action. This shift allows us to address the mechanistic aspects of cognition, which, while not directly accessible to conscious awareness, are crucial for understanding how personal-level phenomena emerge from these subpersonal systems. Therefore, to align with the distinction between

personal and subpersonal levels, I will henceforth refer to these as "*subpersonal representations*" in order to better capture their explanatory role in mechanistic accounts of mental functioning.[2]

Another relevant distinction is that between *representational vehicles* and *representational contents*. This is a particularly important point, since the conflation of these two concepts, illustrated by the often-interchangeable use of the words "representation" and "content" in radical enactive literature, contributes significantly to the confusion and lack of consensus that characterizes the debate. A representational vehicle is the medium that carries or encodes information about what is being represented; one can think of them as the "how" of a representation. A representational content, on the other hand, is its meaning; correspondingly, one can think of them as the "what" of a representation. Before moving on to the consequences of this distinction to the thornier case of mental representations, let me try to clarify it by using an example from natural language. While the English and French written or spoken words "dog" and "*chien*" are representational vehicles, the corresponding concept DOG, which is the meaning of them both, is their representational content. We now have two important pairs of distinctions: first, representations can be *personal* or *subpersonal*; additionally, we can distinguish between representational *vehicle* and representational *content*. As we will see in the next few paragraphs, these distinctions will prove essential for understanding the possible theoretical positions one can assume in the debate about the ontological status of mental representation.

In what concerns mental content and representation, there are three main stances one can assume: realism, eliminativism, and different types of instrumentalism. However, when we discuss

[2] In *Representation Reconsidered*, William Ramsey distinguishes instead between structural and receptor representations. Structural representations, linked to classical computational theories, differ from receptor representations, often favored by connectionist and dynamic approaches, which involve patterns of activation responsive to environmental stimuli. Ramsey avoids the common distinction between symbolic and neural representations because structural and receptor representations cover a broader range, and perhaps correspond more precisely to the distinction in terms of personal and subpersonal levels. In his account, symbolic representations, tied to language-like properties, are a subset of structural representations, which also include modality-specific forms like visual or auditory representations. Similarly, receptor representations encompass not just neural patterns but also external systems, especially under dynamicist views that extend beyond the brain. Because the symbolic-neural distinction is more common, however, I have used it throughout this chapter whenever I am not directly engaging with Ramsey's contributions.

the ontological status of mental representation, it is essential to take into consideration that there are actually two distinct debates taking place: the first concerns the reality of personal representations; the second, of subpersonal representations. Unfortunately, those two debates are often conflated and, since there is no immediate mapping between the theses, arguments, and conclusion contained in each one of them, an already confusing theoretical landscape often becomes even more intractable.

The debate about the ontological status of personal representations is part of a wider discussion concerning the ontological status of the posits of folk psychology—for instance, propositional attitudes such as beliefs and desires, as well as other putative representational entities that populate such theories. In this specific context, realism implies a commitment, of course, to the reality of these posits. Realists about personal representations tend to extend their realism to both personal representational vehicles and personal representational contents. Within this very same debate, eliminativism is the position according to which personal representations have no reality and, hence, no causal powers of their own. The theory is often, but not exclusively, defended in the context of reductionist physicalism and is exemplified by the works of Paul and Patricia Churchland, among others. Among defenders of this position, there is usually the expectation that further developments in the neurosciences will make attempts at formally modelling folk psychology obsolete and unnecessary. An eliminativist will deny reality to both personal representational vehicles and to personal representational content. Still within the debate about the ontological status of personal representations specifically, a wide variety of positions may be more or less precisely called pragmatic, since they recognize the theoretical usefulness of modelling cognition and behavior in terms of propositional attitudes, but at the same time do not commit to their reality—some examples are fictionalism and instrumentalism.

As I said, a different albeit closely related debate concerns the ontological status of subpersonal, neural representations. Since very few people deny the reality of a neural basis for cognition,

at the center of this debate is instead the question of whether neural states can be meaningfully conceived of as representations, or whether they should instead be seen as an intermediate link in a causal chain that goes from the distal source of a stimulus to its broader effects on the cognizing agent's organism. In fact, what seems to be always contentious, in debates about the reality of subpersonal representations specifically, is the status of subpersonal representational content.

Whereas some, coming from the most diverse camps, seem to take for granted that the covariance relation between neural states and their distal causes is enough to characterize a relation of representation, others might view this approach as problematic. In this specific context, realists will usually attribute reality both to subpersonal representational vehicles and to subpersonal representational content, which amounts to the recognition of the representational status of neural pathways, networks, and patterns of activation (i.e., subpersonal representational vehicles) and, consequently, that the relation between this subpersonal representation and what it represents is an intrinsic one (i.e., subpersonal representational content). Eliminativists about subpersonal representation, on the other hand, simply deny that neural states are representations at all, and thereby reject the reality of both subpersonal representational vehicles and subpersonal representational content.

Notice that, interestingly, even though the realist position is not that much affected by the transposition from one debate to another—as reality is ascribed by realists to both personal and subpersonal representational vehicles and contents alike, in both debates—, the eliminativist view is substantially changed. The eliminativist about personal representations was typically a materialist that hoped for a neuroscientific reduction of folk psychology, with a description of cognitive processes in terms of subpersonal representations as its centerpiece, whereas the eliminativist about subpersonal representations simply denies the possibility of subpersonal representations altogether. This is the radical position, as I understand it, defended by many dynamicists and supporters of radical embodied or enactive cognition, such as Hutto, Myin, and Chemero.

This is also the position defended by William Ramsey in *Reconsidering Representation* (2007), who presents what I see as the most compelling rationale for endorsing such a radical position. For Ramsey, while cognitive science is effectively moving away from representationalism by favoring connectionist and dynamicist models of the mind, it is still very much committed to the representationalist jargon of classic cognitivism. He believes the history of science to be rife with such mismatches between the content of a theory and the language used to express it, mentioning the maintenance of the Ptolemaic notion of "celestial spheres" in Copernican cosmology, despite the lack of a clear functional role for them within the new theory. Thus, for defenders of this type of eliminativism, even if subpersonal representation of the sort used in connectionist models do respond selectively to specific stimuli, there is no sense in talking about its content in the same way as one talks about the content of personal representations.

In this view, subpersonal representations do not *mean*, *refer to*, or, in fact, even *represent* their distal source, but rather are simply part of the same causal chain of events that starts at that distal source and ends in a specific pattern of neural activation, and which in fact produces the covariation which is taken by realists to ground representational relations of a semantic nature. This, I believe, is what is at the core of Hutto and Myin's hard problem of content. For the sake of illustration, in the case of visual perception, this causal chain would look roughly like this: photons reflect off the surface of an object and enter the eye, where they are transduced into electrical signals by the retina, which are then transmitted to the primary visual cortex via the optic nerve, where they are processed by various neural mechanisms that extract features such as shape, color, motion, and depth. Hence, according to eliminativists, the relation between a so-called subpersonal representation and its distal source is not really representational, but rather a merely causal one.

The main problem with both types of eliminativism, as voiced by some of its critics, is that they are largely based on promissory notes. Eliminativism about personal representations tends to rely on the expectation that future advancements in neuroscience would potentially render our

current folk-psychological understanding of the mind obsolete. In other words, the hope is that these advancements will reveal a complete picture of mental processes purely in terms of neural activity or other non-mentalistic terms, thereby making notions such as beliefs and desires redundant in scientific explanations of cognition. Similarly, eliminativism about subpersonal representations is based on the expectation that developments in dynamical systems modelling will make the representational framework in which explanations of neural processing is couched redundant.

However, even if certain areas of the cognitive sciences do seem to be undergoing an intensive non-representational makeover, the hard core of what is traditionally conceived as properly "cognitive" activity still depends extensively on representational devices. Moreover, there is a widespread worry among scholars engaged in this debate that, even if the neurosciences do advance significantly and prove capable to elucidate the neural mechanisms that underly most of the phenomena traditionally addressed in representational and symbolic terms by computationalist approaches, there might still persist a gap between the language and concepts used to understand the mind at the psychological level, on the one hand, and the neural level, on the other. In my view, a more prudent strategy would be to develop both approaches simultaneously, as I mentioned in chapter two when discussing the use of formal models in the study of embodied and enactive cognition.

With that in mind, I would like to focus now on a third possible stance in the debate about the ontological status of subpersonal representations specifically. When discussing the pragmatic stance in what concerns the reality of folk psychological theoretical posits, I said a variety of different theories could be thought of as constituting a broader theoretical family. This is an additional point in which the two debates differ: in what concerns subpersonal representations, there is not a wide variety of more or less pragmatic, philosophically interesting positions available. Specifically relevant to us, therefore, is the type of instrumentalism about subpersonal representational content

proposed by Frances Egan, who has dedicated a substantial part of her philosophy to a defense of what she calls a "deflationary account of mental representation."

Subpersonal representational content appears in this view as a mere heuristic, part of an informal explanation of the theory which is shaped by pragmatic considerations, but still, importantly, one without which we would not be capable of making sense of the theory's scope and explanatory targets (Egan 2010, 2014; Coelho-Mollo 2020). Notice that the notion of subpersonal representational content is seen as worthy of preservation as a theoretical entity, given its pragmatic value in our scientific theorizing and, for the time being, the lack of a more explanatorily powerful tool or expedient to replace it. The status of subpersonal representational vehicles themselves is less clear within this theory, and part of what I hope to be able to clarify in the next paragraphs. Let us now try to unpack these ideas by resorting to some of the examples given by Egan.

4. A DEFLATIONARY ACCOUNT OF MENTAL REPRESENTATION

According to Egan, most explanations of content are committed to two basic presuppositions, which are nonetheless very difficult to substantiate: a) that a mental representation has its content essentially, that is, that "[…] if a particular internal structure had a different content it would be a different (type of) representation;" and b) that content is determined by a "privileged naturalistic relation holding between a state/structure and the object or property it is about" (Egan 2018, 251).

She resorts to an analysis of David Marr's influential theory of early vision—specifically, his explanation of how edge detection takes place in the visual system—to show how her instrumentalist stance on content can account for the explanatory power of the theory while prescinding from these dubious presuppositions. Egan explains the mathematical content of Marr's theory by highlighting the function-theoretic characterization employed in computational cognitive science. Marr's theory of the early stages of visual processing aims to explain edge detection by positing the

computation of the Laplacean of a Gaussian of the retinal array (Marr 1982). This computation involves taking intensity values at points in the image and calculating the rate of intensity change over the image, effectively computing a particular smoothing function.

In general, according to Egan, theories in computational cognitive science employ a strategy known as function-theoretic explanation, where the mathematical characterization central to the theory provides a mapping from sets (the arguments of the function) to other sets (its values). This approach explains a cognitive capacity by appealing to an independently well-understood mathematical function under which the physical system is subsumed. For instance, other computational models in perceptual systems compute smoothing functions to eliminate noise, or compute vector subtraction to explain motor control. Therefore, Egan's explanation of the mathematical content in Marr's theory and computational models in general emphasizes the role of function-theoretic characterization in providing a mathematical description of the computational processes involved in cognitive tasks. This mathematical content forms the basis for understanding the computational mechanisms underlying cognitive capacities, and it is a key component of Egan's analysis of content in computational models. It is not sufficient, however, to characterize the cognitive dimension of the theory: it offers only a strictly functional description of the computational mechanism. In her own words, Egan classifies the elements composing the non-cognitive core of the theory in the following way:

> What I will call the *computational theory proper* comprises a specification of (i) the mathematical function(s) computed by the device (the [functional-theoretic] characterization), (ii) the specific algorithms involved in the computation of the function(s), (iii) the representational structures that the algorithms maintain, and (iv) the computational processes defined over these structures. I shall call elements (i)-(iv) the *computational component* of the theory proper. These core elements provide an environment-independent characterization of the device. They have considerable counterfactual power: they provide the basis for predicting and explaining the behavior of the device in any environment, including environments where the device would fail to exercise any cognitive capacity at all.
>
> (Egan 2014, 252)

In order to account for cognitive content, however, one also needs (v), which Egan calls the "ecological component of the computational theory," and which is nothing more than the covariance between features of the computational system and the inputs provided by distal sources in the environment. Notice that, even with the introduction of this environmental embedding of the system and its consequent covariance relations, there is still no semantic content involved:

> Of course, the theorist must explain how computing the value of the mathematical function, in the subject's normal environment, contributes to the exercise of the cognitive capacity that is the explanatory target of the theory. Only in *some* environments would computing the Laplacean of a Gaussian help an organism to see. In our environment this computation produces a smoothed output that facilitates the detection of sharp intensity gradients across the retina, which, when they occur at different scales, typically correspond to physically significant boundaries—changes in depth, surface orientation, illumination, or reflectance—in the scene. Thus the 'theory proper' will also include (v) such environment-specific facts as that a co-incidence of sharp intensity gradients at different scales is likely to be physically significant, corresponding to object boundaries in the world. I shall call element (v) the *ecological component* of the computational theory proper. Together these five elements of the theory proper suffice to explain the subject's manifest cognitive capacity.
>
> (Egan 2014, 252–253)

This is the point where Egan's account connects most closely to the discussion prompted by the hard problem of content. After all, this covariance is nothing more than what Hutto and Myin claimed to be insufficient for characterizing a relationship of semantic representation. And, indeed, in Egan's content-pragmatic account, they are not yet to be considered representational: they are strictly causal, and what they are capable of offering is a causal explanation of the functioning of the visual mechanism when coupled to a specific environment, where elements such as light and the reflective surface of materials behave in a specific manner.

Now, construing a representational relation out of such covariance already requires moving on to a later stage of the theory's formulation, which comprises what Egan calls the *intentional gloss* on the "theory proper." But notice that, in fact, when we have features (i)-(v), we already have a complete formal description of the computational device. Of course, for the sake of brevity and when this way of talking does not produce significant misunderstandings, we may licitly employ representational metaphors to describe aspects of the theory proper: "To say that the structure [in (v)] *represents* edges is 'shorthand' for the facts that constitute the ecological component of the theory, typically facts about robust covariations between tokenings of the structure and distal property instantiations under normal environmental conditions" (Egan 2018, 254). But if cognitive representational content is not part of the theory proper, what role is left for it in the theoretical endeavor?

Egan explains that the primary function of cognitive content is to illustrate, in a perspicuous and concise way, how the computational theory—i.e., the functional-theoretical description of the computations performed by the device—addresses the intentionally characterized phenomena with which the theory is concerned. Cognitive content is ascribed to internal structures on the basis of the cognitive capacity to be explained, what is happening in the subject's normal environment when the structures are tokened, and various pragmatic considerations. For example, in the context of Marr's theory of vision, Egan highlights the intentional gloss associated with the computation of the Laplacean of a Gaussian, which results in the identification of structures such as EDGE.

She argues that calling the structure EDGE highlights its role in the complex process whereby the subject ultimately comes to recover the three-dimensional layout of the scene. This intentional gloss, including the assignment of cognitive content, links the sub-personal mathematical capacities posited in the theory to the manifest personal-level capacity that is the theory's explanatory target. Therefore, Egan's account of cognitive content goes beyond the function-theoretic characterization of mathematical content by emphasizing the heuristic and explanatory role

of cognitive content as an intentional gloss that connects the computational theory to the cognitive capacities it aims to explain.

To elucidate how content ascription participates in the constitution of a theory, Egan presents us with a thought experiment: the Visua and Twin-Visua case (Egan 2014, 125–126). This is nothing more than a hypothetical scenario designed to illustrate the concept of cognitive interpretation and its relation to environmental differences. In this scenario, Visua is a mechanism that computes the depth of objects and surfaces in the immediate vicinity based on retinal image information. Much like our own visual apparatus, it achieves this in part by representing edges, as its states covary with changes in depth or edges in the immediate environment. Now, imagine an exact physical duplicate of Visua, called Twin-Visua, but situated in a different environment, Twin Earth. Despite being computationally identical to Visua, Twin-Visua operates in an environment where light behaves differently, causing its states to covary with shadows rather than edges.

The thought experiment highlights the challenge of providing a single cognitive interpretation that specifies what this mechanism represents in both worlds: since it does not track shadows on Earth or edges on Twin Earth, neither EDGE nor SHADOW are suitable candidates for its intrinsic content. Tentatively, Egan's hypothetical scientist in the story proposes a more general and ambiguous entity called EDGEDOW, one that subsumes both edges and shadows, allowing for a unified cognitive interpretation that is applicable to both Visua and Twin-Visua. Egan then prompts us to reflect on how much these changes would impact our own understanding of what Visua is supposed to process, as well as its role in visual processing. As she notices, a textbook describing the workings of Visua on Earth would be much more readily understandable by Earthlings than an environment-neutral counterpart that attempted to explain its functioning both on Earth and Twin Earth: "Besides introducing a new vocabulary containing such unfamiliar predicates as 'edgedow,' the new edition [would require] cumbersome appendices appropriate to each world, explaining

how to recover answers to questions about the organism's capacities in its local environment, questions that motivated the search for an explanatory theory in the first place" (Egan 2014, 126).

Although her examples are usually focused on perception and motor activity, I believe what Egan attempts to show is that conflating the roles of subpersonal representational vehicles and representational contents leads to significant confusion in the cognitive sciences in general, and not only in its subfields that are concerned with perception. Having presented a general outline of Egan's deflationary account of mental content, I will now conclude with a few remarks on its relationship with Hutto and Myin's hard problem of content.

5. Conclusion

For the sake of clarity, let us now briefly restate the steps of the arguments presented in the previous paragraphs. We started, in §1 and §2, by investigating the various historical and theoretical motivations for the antirepresentationalist stance in enactive philosophy of mind and cognitive science, as well as the more specific reason offered by radical enactivists when defending it—i.e., Hutto and Myin's hard problem of content. In §2, the hard problem of content was presented in terms of the impossibility of naturalizing the richer notion of semantic representational content—i.e., content with satisfaction conditions—by grounding it on the strictly informational notion of representational content as covariance. We have also briefly discussed some criticisms to Hutto and Myin's resistance to generalize their antirepresentationalism to higher levels of cognitive activity, such as language and reasoning. Instead, they present their account as valid only for the domain of "basic minds"—i.e., roughly, prelinguistic minds engaged in perceptual and motoric activity.

In §3, we have made some important distinctions between representations at the personal and subpersonal levels of explanation, as well as between representational vehicles and representational content. In doing so, we also had the opportunity of distinguishing between various types of realist, eliminativist, and deflationary theories of mental representation, according to their

respective explanatory targets. From that point onwards, our primary focus was subpersonal representations and processes, since this is the relevant level of explanation when discussing Hutto and Myin's basic mentality. With that in mind, in §4, I have shown how one particular type of deflationary theory of subpersonal mental representation, Egan's deflationism about subpersonal representational content, is capable of offering a solution to the hard problem of content. Unlike realist approaches that treat representational content as something metaphysically robust, or essential to the function of a mental state, Egan proposes that we view it as a mere theoretical tool—a shorthand that helps cognitive scientists map computational processes onto cognitive tasks without making strong metaphysical commitments. By reframing representation as a non-essential feature of our explanatory frameworks, Egan dissolves the tension inherent in Hutto and Myin's hard problem of content: the question of how content could arise from mere covariance is revealed as misguided because representational content never needed to arise from it in the first place.

This line of reasoning leads us to the following conclusion: the opposition frequently assumed between enactivism and representationalism is often overstated, particularly if we adopt a more pragmatic or instrumentalist perspective on the nature of mental representation. Although it is tempting for defenders of radical enactivism to claim that cognitive science could and should dispense entirely with this notion, such claims rest on an overly narrow understanding of representation's role within cognitive models. As we have seen throughout this chapter, talk of mental representation need not be understood as a metaphysical commitment, and much of the resistance to it hinges on a failure to recognize its pragmatic or instrumental function within scientific explanations of cognition. Thus, when seeing representation as a concept to be discarded, many enactivists overlook its utility as a theoretical tool, both at the personal and subpersonal level, and especially within the current epistemic framework of cognitive science.

Nonetheless, Egan's deflationary account of mental content proves to be compatible with the enactive view by treating representational content as a pragmatic, non-essential feature of

cognitive models: representational content functions, in fact, as a useful heuristic that aids in explaining computational theories, even if it is not intrinsic to cognitive activity. It preserves the explanatory power of representational language while avoiding, at the same time, the metaphysical commitments that enactivists critique. In this light, the hard problem of content—formulated by Hutto and Myin as the impossibility of naturalizing mental content through covariance relations—becomes less of a genuine problem and more of a theoretical illusion, one that is simply dispelled when we abandon the notion that representations need to have intrinsic semantic content.

Furthermore, an instrumentalist approach to representation helps to reconcile what might initially seem like a deep theoretical divide between enactivism and representationalism. The instrumentalist view, by allowing for the use of representational constructs without ascribing to them any intrinsic metaphysical status, shows that representation need not be discarded wholesale in order to maintain the integrity of enactivist theories. The rejection of representationalism, particularly in its more moderate, deflationary forms, risks throwing out valuable explanatory tools simply because they do not conform to the more radical antirepresentationalist agenda. This agenda, however, seems overly ambitious, especially given that cognitive science still very often relies on representational models to describe many aspects of cognitive systems.

Finally, most of what has been said in the second half of this chapter applies more directly to the case of subpersonal representational content, as this seems to be the main focus of Egan's approach. In view of radical enactivists' self-declared concerns, however, this needs not be a significant problem, for they do in fact limit the scope of their own antirepresentationalism to basic mentality, whose mechanics is expected to be explained subpersonally anyway. Still, I presume a possible objection to the solution presented here could be that it is not capable of explaining the presence of representational posits at the personal level, and that it hence falls short of its stated goal. Some radical enactivists, after all, do not stop at the level of basic minds, but instead extend their iconoclasm to higher-level cognitive phenomena, such as language and reasoning. I believe,

however, that similar strategies can be adopted in attempting to elucidate the ontological status of representational content at the personal level too, and I would like to end this chapter by briefly pointing out the directions this would take.

One possible way to extend the deflationary strategy to personal representations is by treating them also as pragmatic constructs—useful for predicting and explaining behavior but not reflecting intrinsic features of cognitive systems. In other words, just as subpersonal representations serve as heuristics in computational models, personal representations such as beliefs, desires, and intentions can be viewed as tools that help make sense of personal-level cognition without committing us to their ontological reality. This would mean that the explanatory role of personal representations lies in their utility for understanding how agents navigate their environments, rather than in their metaphysical status as genuine mental entities.

This position is sometimes referred to as *mental* or *folk-psychological fictionalism*, as it has gained some traction in recent debates about the nature of the posits of folk psychology. Fictionalists, such as Adam Toon, argue that personal-level representations—e.g., beliefs and desires—should simply be understood as useful fictions that help us organize and predict behavior, even though these posits do not correspond to actual, metaphysically real entities (Toon 2023; Demeter, Parent, and Toon 2022). In my view, fictionalism provides a compelling parallel, at the personal level, to Egan's deflationary approach by offering a similar strategy for explaining personal representations—viewing them as part of an explanatory framework rather than as reflections of cognitive reality.

V. ENACTION AND COMPUTATION

One prominent feature of contemporary theorizing about the mind is the computer metaphor, which laid the foundations for the computational theory of mind, also called computationalism. This chapter examines the compatibility of this picture of cognition with the enactive approach. As we will see in §1, the mind-as-a-computer metaphor broadly equates the mind—and, very often, the brain—to information-processing systems such as digital computers, viewing cognitive functions as computational operations over symbolic representations. Thus, in the context of this analogy, the brain's neural circuitry functions as a piece of hardware that executes software-like mental programs, receiving perceptual inputs, performing internal computations, and generating behavioral outputs. As with other scientific metaphors, framing cognitive processes in terms of algorithms and data structures allows us to develop formal models that help us understand complex mental phenomena, and the relative success of the cognitive sciences throughout the last half century or so is indeed much indebted to this way of seeing and thinking about the mind.

Many scholars, however, contend that describing mental processes as computational is not metaphorical at all (Miłkowski 2018; Piccinini 2015). For them, computation is a natural kind instantiated in both artificial and biological systems, rather than an artificial, observer-dependent concept whose extension is negotiated between cognitive and computer scientists. On this view, the function of performing computations cannot be arbitrarily attributed to any system according to our own research interests, but its defining characteristics must instead be identified in paradigmatic systems such as abacuses, calculators, digital computers, brains, or minds. However, this strategy raises significant concerns of circularity: to prove that the mind or the brain are computational systems, we demonstrate that they meet criteria defined based on what we already considered computational in the first place, the mind and brain themselves included. This circularity and related issues have led many philosophers to be skeptical of literalist approaches to the computational

mind, preferring to treat computation as an artificial kind, whose use in the characterization of systems is context and purpose-dependent (Searle 1980; Churchland 1986; Haugeland 1985; Putnam 1987; and even Fodor 2000). I will call this position "computational perspectivism."

At first glance, the claim that the brain or the mind are computers appears incompatible with enactivism. This incompatibility stems from the fact that computation is often understood in representational terms—with the symbols manipulated by computational systems *representing* something. Indeed, this dependency—which is perhaps best illustrated by Jerry Fodor's slogan that there is "[…] no computation without representation"—still holds strong hegemony in the cognitive sciences (Fodor 1975). For a variety of reasons, however, a number of scholars have recently attempted to offer alternative accounts of computation that do not necessarily involve representation (Miłkowski 2011; Fresco 2010; Piccinini 2015). In §2, I will discuss both the motivations and the general arguments in favor of Gualtiero Piccinini's mechanistic, non-semantic account of computation, which is perhaps the most detailed and well-developed among such proposals.

Even though Piccinini himself is neither an enactivist nor overtly committed to the idea that mental computation specifically—as opposed to computation in general—can be entirely explained without the concept of representation, some philosophers have argued that his account could make computationalism compatible with the broader enactive framework (Dewhurst 2014b, 2016b; Dewhurst and Villalobos 2017; Piccinini 2015). This is because the tension between enactivism and computationalism allegedly arises only under semantic views of computation. In §3, we will discuss some of these proposals for harmonizing enactivism and computationalism, with an emphasis on the unifying role envisioned by some for the notion of *wide computationalism* in the context of 4E cognition. However, in §4, we will see that one aspect of Piccinini's proposal introduces significant difficulties not only to this syncretic project, but also to the coherence of his own theory.

The main difficulty with Piccinini's proposal concerns his reliance on the notion of "proper function"—including the proper function of computing—which he argues is an entirely objective property. In §4, I will explain how this commitment introduces difficulties, not only for reconciling computationalism with enactivism but also for the coherence of his own account. To address these challenges, I will explore an alternative approach to computational function that, while rejecting strong objectivism, avoids pancomputationalism by grounding computation in the structural characteristics of physical systems. This perspective will be supported by examples of systems where computational interpretation is constrained by intrinsic causal organization. Ultimately, I will argue that while traditional objectivist computationalism may be at odds with enactivism, a deflationary and perspectival approach can preserve explanatory power while aligning with enactivism's emphasis on observer-relativity and dynamic, embodied cognition.

1. COMPUTATION AND COMPUTATIONALISM

Let us begin by distinguishing the general notion of computation, applicable across various domains, from computationalism, a much narrower philosophical doctrine which posits cognition as fundamentally computational. Although the concept of computation is employed differently across diverse fields, Alan Turing's formulation remains central both to theoretical computer science, as well as to philosophical and cognitive-scientific discussions about the nature of the mind. In his influential paper "On computable numbers, with an application to the *Entscheidungsproblem*," Turing introduced the concept of the *Turing machine*—a simple yet powerful abstract device capable of executing any algorithmically expressible function (Turing 1936). By formalizing the concept of an *algorithm*—which is nothing but a precise, step-by-step procedure for calculating the output of a logico-mathematical function from given inputs—Turing provided the first rigorous definition of computation. A key aspect of his model is its universality, as what has been called a *universal Turing machine* can simulate any other computational system through suitable programming.

The broader implications of Turing's work are encapsulated in the Church-Turing thesis—thus called because Alonzo Church arrived at similar results simultaneously—, which states that any function computable by an algorithm is, in principle, computable by a Turing machine. This thesis laid foundational groundwork not only for the fields of computer science and artificial intelligence, but also for cognitive science, where the mind itself has often been modeled as a computational system. Computationalism, emerging from this intersection, is the view that cognition involves manipulating symbolic mental representations—such as beliefs and desires—through formal, algorithmic processes analogous to those performed by a Turing machine. Articulated explicitly in theories like Newell and Simon's *physical symbol system hypothesis* (1972) and Fodor's *language of thought hypothesis* (1975), computationalism became central to classical cognitive science.

Nevertheless, classical computationalism has faced significant criticism, particularly regarding the problem of representation: Turing machines manipulate symbols based solely on their formal, syntactic properties, making no reference to semantic content, and thus leaving open the question of how meaning arises from purely formal computational processes. This issue has motivated alternative accounts, notably Piccinini's mechanistic approach, which reconceives computation in non-semantic terms by emphasizing physical processes and causal-mechanistic organization, thus avoiding reliance on problematic representational commitments. In what follows, I will explore Piccinini's proposal, its theoretical rationale, and its broader implications for cognitive science.

2. NON-SEMANTIC, MECHANISTIC COMPUTATIONALISM

Before presenting Piccinini's core ideas, I must introduce a caveat. Notice that we here face a two-tiered problem: establishing that computing systems in general are non-representational does not automatically imply a non-representational account of the computational mind. Even if representation is neither necessary nor sufficient for defining computation broadly, it may nonetheless turn out to be indispensable—under the guise of mental representation—in explaining mental

computation specifically. However, from the standpoint of compatibility with enactivism, this uncertainty is still much more preferable. If representation were inherently essential to computation in general, it would inevitably be required for mental computation as well, thereby eliminating the possibility of the enactive reinterpretation to be explored in subsequent sections. Having a non-representational notion of computation, on the other hand, leaves open the possibility of developing a computational account of mentality along enactive lines, emphasizing dynamic systems, embodied interactions, and emergent processes rather than symbolic representation. Keeping this in mind, we now turn to Piccinini's framework.

2.1. SIX DESIDERATA FOR A THEORY OF COMPUTATION

In his book *Physical Computation: a Mechanistic Account*, Piccinini presents an alternative framework for understanding computation that is applicable across all domains where the notion plays a central role, such as computer science, philosophy of mind, and cognitive science. For him, a solid theory of computation must respond to six *desiderata*. The first one is 1) *objectivity*—the idea that determining whether a system is performing a specific computation should be a matter of fact, free from subjective interpretation (Piccinini 2015, 11). Second, the theory must provide a framework for 2) *explanation*, enabling the capacities of a system to be understood through the computations it performs (Piccinini 2015, 12). A third crucial aspect is ensuring that 3) *the right things compute*: the theory should accurately identify which systems are genuine computing mechanisms, including paradigmatic examples such as digital computers, calculators, both universal and non-universal Turing machines, and, probably, brains (Piccinini 2015, 12).

Conversely, the theory should also ensure that 4) *the wrong things do not compute*: systems that we do not typically consider computational—such as planetary systems, hurricanes, and the human digestive system—should not be classified as performing computations under a valid theory (Piccinini 2015, 12). Moreover, the theory must be able to explain 5) *miscomputation*, providing insight

into how and why a system might fail to compute correctly (Piccinini 2015, 13–14). Finally, 6) *taxonomy* is important, as different computing mechanisms have varying capacities. The theory should account for these differences—whether the limited operations performed by logic gates, the finite yet significant capabilities of non-programmable calculators, or the broad computational potential of ordinary digital computers, constrained only by memory and time. A theory that can shed light on these distinctions is more valuable than one that does not recognize them (Piccinini 2015, 14).

Most of these *desiderata* will play a role in the arguments I will develop in this chapter, but for reasons that will become clear in a moment, the first of them—i.e., 1) *objectivity*—is particularly relevant for us. The most distinctive feature of Piccinini's proposed theory for meeting the six criteria is that it is "mechanistic" and "non-semantic." The requirement that it be mechanistic is clearly a consequence of the hegemony of naturalistic and physicalistic methodologies in the natural sciences. The requirement that it be non-semantic, however, deserves further exploration. Even though his theory has occasionally been mobilized in favor of enactive approaches, Piccinini himself is not an enactivist, and his project is not motivated by the same concerns that often lead enactivists to reject representational computationalism. Instead, he seems to have two main motivations for pursuing a non-semantic account of computation: i) the perceived redundancy of the notion of representation in the individuation of the concept of "computation," and ii) the difficulty in preserving objectivity within semantically individuated accounts of computation.

The first of these two motivations stems from the following finding, which is presented in the third chapter of his book: according to Piccinini, semantic ways of individuating computational processes cannot stand on their own, for they all presuppose a non-semantic way of individuating the very same processes (Piccinini 2015, Ch. 3). The converse, however, does not hold, as it is possible to successfully individuate computation on strictly mechanistic grounds, with no resort whatsoever to semantic notions such as "content" or "representation," as he attempts to show in

the positive part of his work. Hence, a non-semantic individuation of computation is not only necessary for establishing the theory of computation on a solid foundation, but also suffices for it.

As for the second motivation, Piccinini contends that semantic accounts often make computational properties dependent on how observers interpret them, which can lead to subjective variability. This, he claims, violates his first *desideratum* for objectivity in the characterization of computational processes. In the type of interpretational semantics that characterizes a large family of current approaches to the issue of mental representation, including Egan's content pragmatism, which I defended in the last chapter, a system's semantic properties are determined by an observer's choice to ascribe meaning to its states, which undermines the objectivity of computational descriptions. In contrast, his non-semantic account allows for a more objective description of computational systems, which aligns better with the practices in computer science. For Piccinini, this should be made evident by the fact that we can perfectly well talk about computation without any inherent semantic content, whereby systems manipulate physical states and symbols based on syntactic rules, without needing those states or symbols to have semantic meaning.

Let us clarify the problem at hand. If we accept Piccinini's premise, an objective account of computation is necessary. However, representational accounts of computation—whether teleological (e.g., *à la* Dretske and Millikan) or interpretational (e.g., *à la* Egan and Dennett)—do not satisfy this requirement. Teleological accounts struggle with content individuation, while deflationary accounts render representations observer-relative, both of which contradict objectivity. Piccinini's solution is to retain objectivity by rejecting representations in computational individuation, which necessitates a non-semantic approach. Thus, before proceeding to a more detailed investigation of the role of objectivity in computation, we must first understand Piccinini's positive proposal for individuating computation non-semantically. This discussion will lay the groundwork for our subsequent analysis of why mechanistic accounts may be instrumental in reconciling computationalism with enactivism in §2.3.

2.2. CRITIQUE OF MAPPING AND SEMANTIC APPROACHES

When arguing for the redundancy of representational content in computational individuation, Piccinini's first targets are not semantic, but mapping accounts. These are not mutually exclusive, however, as the mapping from states in a computational description to states of a physical system are usually incorporated into semantic approaches. Still, the distinction is pertinent, since a mapping account needs not necessarily incorporate any semantics. Mapping accounts posit that any physical system can be considered a computing system if there is a mapping between its microphysical states and the states of a computational description. By "microphysical states," Piccinini and others mean the detailed physical states of a system at a very fine-grained level, encompassing the fundamental physical properties and configurations of said system's components, typically described in physical terms: e.g., positions, velocities, and interactions of particles or the states of various subatomic and atomic entities. In mapping accounts, they are correlated with states in a computational description of the system, which are abstract and concerned with the functional or logical operations being performed: e.g., binary digits being processed in a computer, calculations being performed in an abacus, among others.

One important issue is their inability for accounting for miscomputation: because these approaches are based solely on the existence of mappings, they cannot explain why a system might produce incorrect outputs or fail to compute properly. But his main criticism of mapping accounts *simpliciter* hinges upon their over-inclusiveness. Since most if not all physical phenomena can be described using computational models, mapping accounts seem to entail that all physical systems compute, an idea that violates his third and fourth *desiderata*—i.e., that only 3) *the right things compute* and that 4) *the wrong things do not compute*—, thereby diluting the concept of computation to the point of triviality.

Indeed, one often mobilized argument against the computational theory of mind hinges upon this alleged triviality, claiming that, under certain descriptions, *anything* computes (Searle

1990). In other words, in mapping accounts, any system with a mapping from physical states to computational states can be said to perform a computation, an idea that has been called "pancomputationalism" by some. In general, pancomputationalism is deemed undesirable for a series of reasons. For instance, it dilutes the concept of computation to the point where it loses any specificity and explanatory power and also puts at risk highly valued findings in scientific fields that depend on a much more robust notion of computation. Hence, Piccinini also argues that mapping accounts lack explanatory power, thereby violating his second *desideratum*—2) explanation—, as they do not distinguish between computational explanations, which attribute a system's behavior to its computational processes, and ordinary causal or dispositional explanations.

Pancomputationalism is a fringe position in philosophy, although more casually endorsed by some scholars in other fields. Usually, the very idea of pancomputationalism only appears in *reductiones ad absurdum* of positions such as mapping and semantic accounts of computation, and is seldom positively presented as a thesis in itself. In fact, one of the main motivations of Piccinini's mechanistic account is to show that non-semantic approaches to computation do not necessarily collapse into pancomputationalism. It is important to notice the difference between pancomputationalism, which consists in the belief that all systems are computational, and a position that we could call "computational modeling universalism," for the lack of a better name, consisting in the view that all systems can be the target of computational models. Atoms, chemical reactions, planetary systems, and food digestion can all be modeled computationally, but this is not the same as saying that they are all computing systems. The belief in the universality of computational modeling is much less controversial than computationalism, and in fact can be seen as part of a general optimistic stance concerning our epistemic capacities that characterize the scientific endeavor as a whole.

Semantic approaches, on the other hand, attempt to overcome the difficulties faced by mapping accounts *simpliciter* by introducing additional constraints onto the notion of computation.

They address shortcomings such as collapse into pancomputationalism by emphasizing the meaningful manipulation of symbols and the functional role of states in computation. Thus, pancomputationalism could allegedly be avoided by restricting valid mappings and by focusing on meaningful computation, with "valid mappings" being defined precisely as those that preserve semantic content. This focus on meaningful computation, in turn, distinguishes between genuine computational processes and arbitrary state transitions. For Piccinini, however, semantic accounts fail for two main reasons. The first one is the inherent vagueness of semantic properties. Semantic properties are usually taken to be more obscure and less well-understood in comparison with syntactic or purely formal ones, which makes them a weak foundation for a robust theory of computation. I will not investigate these shortcomings extensively here, since they were already discussed in Chapter IV when we analyzed the hard problem of content and the difficulties it poses for philosophical projects aiming at a naturalization of semantic relations.

The second reason why semantically individuated computation is generally unsuccessful hinges on its ultimate dependency on non-semantic properties. Piccinini argues that even when computations are individuated semantically, this presupposes a prior non-semantic—syntactic or formal—individuation. Thus, semantic individuation remains secondary to a more fundamental mechanistic or syntactic one. Computation can exist without semantic content, but semantic content necessarily depends upon underlying computational processes. Consider a digital computer manipulating binary digits (0s and 1s). Internally, the computer processes these digits solely according to predefined syntactic rules encoded in its machine table, entirely independently of what these digits might represent externally. A binary sequence such as "01000001" can mean various things externally—a decimal number ("65"), an ASCII character ("A"), or something else—depending on how it is interpreted. However, these external meanings do not influence the computational process itself. The system treats "01000001" purely as formal symbols, manipulating them based exclusively on syntactic properties. Thus, semantic interpretations are layered on top of formal computational processes by external observers. Piccinini's critique emphasizes precisely this point:

semantic individuation of computation necessarily presupposes syntactic individuation. Since the computational process itself is entirely formal, semantics, while relevant for interpreting computational outcomes, is secondary and dependent upon the underlying syntactic structures that define computation.

2.3. PICCININI'S MECHANISTIC COMPUTATION

Since he considers semantic content insufficient to define computation and to avoid pancomputationalism, Piccinini appeals to alternative individuation criteria in his mechanistic account. For him, the mark of the computational is teleological function. Piccinini argues that physical computing systems are a subclass of functional mechanisms—systems of components with causal powers organized to perform a function, in this case, computation. Computing mechanisms differ from non-computing ones in their function: they manipulate vehicles (i.e., data or information) based solely on differences between their portions, following rules and possibly considering internal states:

> A Physical Computing System is a system with the following characteristics:
>
> - It is a functional mechanism—that is, a mechanism that has teleological functions.
> - One of its teleological functions is to perform computations. Generic Computation: the processing of vehicles by a functional mechanism according to rules that are sensitive solely to the differences between different portions (i.e., spatiotemporal parts) of the vehicles.
>
> (Piccinini 2015, 121)

In this passage, Piccinini defines a physical computing system as a functional mechanism—that is, a system that has specific purposes or functions it is designed or evolved to fulfill. Crucially, one of these teleological functions must be to perform computations, which implies that computation is not just an incidental property, but a core purpose of the system. The concept of

"teleological function" in Piccinini's definition refers to the purpose or goal that a system is designed or evolved to fulfill. This notion is crucial for distinguishing genuine computing systems from other physical processes that may be described computationally but are not actually computing. The teleological functions of man-made computing mechanisms are intentionally designed and implemented by human engineers to fulfill specific computational purposes, which makes them typically more narrowly defined and optimized for particular computational tasks. Thus, unlike natural systems, the functions of artificial computing systems are explicitly programmed and architected, rather than emerging from complex interactions of simpler components.

When it comes to natural computing mechanisms like brains, their teleological functions arise, of course, through evolutionary processes and natural selection. The brain evolved to perform cognitive functions that aided survival and reproduction, and they emerged gradually over long periods of time as adaptations to environmental challenges. As such, the teleological functions of natural computing systems tend to be broader and more flexible, capable of dealing with a wide range of situations. Despite these differences, however, Piccinini's mechanistic account aims to provide a unified framework that can encompass both natural and artificial computing systems.

Piccinini then provides a definition of "generic computation"—i.e., the teleological function that such mechanisms must perform—that forms the basis of his account. He describes it as the processing of vehicles by a functional mechanism according to rules. These rules are sensitive only to differences between portions or parts of the vehicles being processed, which means the computational process depends only on the structure or states of the vehicles, and hence are *medium-independent*. This means that they depend neither on specific material properties of the vehicles, nor on any semantic content or meaning they might carry. This has two important consequences: the first one is that Piccinini's idea of medium-independence is broader than the more widespread notion of multiple realizability, it requires that the teleological function of a mechanism be completely independent of the physical medium, going beyond just being realizable by different

mechanisms under different conditions. Additionally, it shows that the notion of medium-independence is one of the key elements in distinguishing his own approach from semantic accounts of computation. By defining computation in terms of a mechanism processing vehicles according to difference-sensitive rules, Piccinini provides an objective, non-semantic basis for understanding physical computation.

Let us now see how Piccinini's definition fares in regard to the six *desiderata* he laid out in the beginning of his book. The first, i) *objectivity*, is guaranteed by tying computation to the causal structure and functional organization of physical systems, stipulating that the proper function of computing consists in the manipulation of medium-independent vehicles according to rule-based transformations. The second desideratum, ii) *explanation*, is satisfied because an analysis of the causal structure of computing mechanisms is capable of fully explaining their behavior.

The definition of computation in terms of medium-independent vehicle manipulation guarantees that iii) *the right things compute*, since this framework is capable of accommodating digital, analog, and neural computation, and allows for distinctions between hardware and software. As for iv) *the wrong things do not compute*, non-computing systems—which might nonetheless be amenable to computational modeling—are excluded in this framework, since they clearly do not perform the teleological function of computing. For instance, although digestion can be computationally modeled, the digestive system's function is biochemical rather than computational; crucially, it does not manipulate discrete, medium-independent vehicles according to rules sensitive only to differences among discrete spatio-temporal parts. Instead, digestion involves continuous biochemical processes defined by specific chemical properties of substrates, enzymes, and molecular interactions, which clearly fall short of Piccinini's definition of computation. The fifth desideratum, v) *miscomputation*, is met by grounding computation in physical mechanisms with teleological functions, with miscomputation occurring when a mechanism fails to execute its proper function.

Finally, the sixth *desideratum*, vi) *taxonomy of computing systems*, requires that a computational theory align with hierarchical distinctions recognized in computer science and cognitive science. Piccinini's taxonomy also satisfies this requirement, as it categorizes computing systems by their components, operations, and organizational structure. Primitive computing components, such as logic gates, perform elementary operations but cease to be computational if they are further decomposed. They can, however, be combined into higher-level structures like arithmetic-logic units, memory modules, digital calculators, and general-purpose computers. The taxonomy also is capable of accommodating both analog computers, which operate on continuous variables, and neural networks, both artificial and biological. Thus, Piccinini's mechanistic account meets the six *desiderata* by grounding computation in causal structure and functional organization. Defining computation as medium-independent vehicle manipulation, his approach should in principle offer an objective, explanatory, and taxonomically structured framework that captures genuine computational systems while avoiding pancomputationalism. We will now see why this functional, non-semantic approach to computation has been seen by some as offering a basis for reconciling enactive theories of cognition with the computationalist framework.

3. COMPUTATIONALISM AND ENACTIVISM

As established in the previous chapter, most enactivists fundamentally reject the notion of intrinsic mental representations at both subpersonal and personal levels, insisting that cognition is not a matter of manipulating symbols but instead an emergent, dynamic process rooted in the ongoing interplay between organism and environment. Thus, if such representations are indeed essential for defining computational systems, it might initially appear that computationalism cannot be reconciled with enactivism. At this juncture, however, we confront a crucial question: beyond rejecting intrinsic representations, do enactivists have any further reason to resist computationalism? In this section, I introduce the perspective of a group of philosophers who answer this question in the

negative (Dewhurst 2014, 2016; Dewhurst, Deane and Kersten 2017; Dewhurst and Villalobos 2016, 2017a, 2017b). Moreover, these authors maintain that harmonizing the enactive and computationalist frameworks could prove mutually advantageous.

For computationalists, it would facilitate the incorporation of insights from the enactive camp, which are frequently portrayed as threatening the very foundations of their paradigm. For enactivists and other proponents of 4E cognition, it would help dispel many of the apparent incompatibilities between enactivism and the related embodied, embedded, and extended approaches, thereby unifying them under the broad scope of wide computationalism. In order to implement this proposal, they resort to Piccinini's mechanistic conception of computation. Nevertheless, as we shall see, Piccinini's emphasis on objectivity in the individuation of computational functions—encapsulated in his requirement that such functions be "proper"—introduces significant constraints that complicate this proposed integration.

3.1. WIDE COMPUTATIONALISM

The principal motivation behind Joe Dewhurst and others' endorsement of a computationalist version of enactivism seems to be the following. As we have seen in the first chapter of this dissertation, some tensions arise among the multiple "Es" of 4E cognition: for instance, embodied cognition's insistence on the significance of anatomical and physiological contingencies appears at odds with extended cognition's commitment to radical functionalism, just as extended cognition's insistence on a seamless organism-environment continuum seems incompatible with autopoietic enactivism's emphasis on an organism-environment boundary. Because enactivism is arguably the most encompassing theory within the 4E spectrum—encompassing both body and environment as constitutive elements of cognitive activity, and thus, in a certain sense, also a form of embodied and extended cognition—it is incumbent upon enactivist theorists to provide a compelling resolution to these apparent incompatibilities.

Thus, Dewhurst, Deane and Kersten undertake the task of reconciling these divergent strands of 4E cognition by demonstrating how *wide computationalism* (Kersten 2017), a term originally proposed by Robert Wilson (1994), can accommodate key enactivist insights while preserving the explanatory power of computational theories. At the heart of their proposal is the notion that computational systems need not be confined within the boundaries of skull or skin but may incorporate bodily and environmental elements. This "wide" approach to computation thereby facilitates a reconfiguration of embodied and enactive perspectives in computational terms, all without relinquishing the crucial emphasis on the integral role of the body and environment in cognitive processes.

> Wide computationalism is the view that some of the units of computational cognitive systems reside outside the individual[.] [...] Wide computationalism gains a theoretical foothold via the location neutrality of computational individuation. Since formal systems are indifferent to physical medium and computation is a formal system, it is possible that at least some states and processes relevant to a computational system may reside outside the individual. Nothing in the method of computational individuation precludes the possibility of wide computational systems.
>
> (Dewhurst, Deane, and Kersten 2017)

Naturally, a key step in their argument is the adoption of a functional and mechanistic account of computation, drawing heavily on Piccinini's work, which distinguishes their position from other variants of wide computationalism, such as those advocated by Wilson, Clark, or Chalmers (Wilson 1994; Clark 1997, 2008; Chalmers 1998). As we have seen, this account conceives of computation in terms of functional mechanisms that operate on medium-independent vehicles, rather than in terms of symbolic representation or information processing (Piccinini 2015). By invoking this mechanistic framework, they effectively avoid traditional enactivist objections to computationalism—objections that typically hinge on a rejection of representationalism—and

thereby argue that their mechanistic, wide computationalism is capable of resolving two major tensions in 4E cognition.

The first tension emerges between the "body-centric" orientation of embodied cognition—which stresses the central and constitutive role of the body in shaping cognitive processes—and the "distributed" outlook of extended cognition, which treats the body as merely one component within a broader brain-body-world continuum. Meanwhile, the second tension arises from the enactivist insistence on organism-environment boundaries, which appears incompatible with the extended claim that cognitive processes may span brain, body, and world (Dewhurst, Deane, and Kersten 2017). In response to the first tension, Dewhurst, Deane, and Kersten propose reframing body-centric claims in terms of "wide functional mechanisms" anchored within the brain-body system. Thus, instead of viewing bodily contributions as uniquely privileged, they become simply one instance of a broader class of cognitive processes instantiated by mechanisms extending beyond the confines of the brain. At the same time, this computational framework preserves the distributed and medium-neutral perspective typical of extended approaches.

> The wide account of computation extends the mechanistic reasoning to brain-body-world systems. It maintains that whether or not functional mechanisms, ones that process medium-independent vehicles, are constituted by spatiotemporal components squarely localized within the individual or crisscrossing into the world is an a posteriori question. Since the mechanistic conditions on concrete computations are medium and location neutral, the question of wide computational systems is an open one – some physical computing cognitive systems may be ensconced within the body, while others may be spread out over brain, body and world.
>
> (Dewhurst, Deane, and Kersten 2017)

Accordingly, Dewhurst, Deane and Kersten view wide computationalism as a framework that enables a universal translatability among the various 4E approaches, offering a common language of mechanism and computation capable of incorporating the core insights of embodied,

extended, and enactive perspectives. In their account, body-centric claims are thus reformulated as statements about wide mechanisms, while autopoietic notions of autonomy are reconceived in computational terms of self-maintenance (Dewhurst, Deane, and Kersten 2017). Through this maneuver, they carve out a middle ground between internalist and externalist theories of cognition, retaining the emphasis on bodily and environmental structures so central to 4E approaches, yet preserving the explanatory apparatus of computational cognitive science.

A central component of their argument is that wide computationalism offers precisely the right level of granularity to unify different strands of 4E cognition: it is sufficiently abstract to encompass extended and distributed processes, but remains anchored in physical mechanisms in a manner that does justice to the enactivist and embodied insistence on the concrete, material dimensions of cognitive activity. In this way, the mechanistic framework paves the way for multilevel explanations that accommodate both the fine-grained details of bodily engagement and more general, functional analyses of cognitive processes.

The second tension, of particular relevance here, concerns the apparent conflict between enactivism's emphasis on a boundary between organism and environment and the "location-neutral" nature of extended cognition theories. To address this problem, Dewhurst, Deane, and Kersten also propose that wide computationalism, predicated on a mechanistic, medium-neutral view of computation, offers a promising path to reconciliation. In their framework, cognitive systems are defined by the functional organization of their components rather than by their physical limits. Consequently, the organism-environment boundary—central to enactivist thought—need not serve as a strict demarcation for delineating cognitive processes. While autopoietic systems, foundational to enactivist approaches, often seem confined to the organism because of their focus on autonomy and self-regulation, the authors contend that such confinement is more an empirical observation than a theoretical requirement. Indeed, computational systems can likewise maintain autonomy through distributed mechanisms that extend across organism and environment. Their

account gains plausibility if one considers the significant difficulties, mentioned in the third chapter of this dissertation (Ch. 3, §1.2), surrounding the task of providing a strict definition for "boundary" in the case of non-unicellular autopoietic systems—e.g., multicellular organisms and putatively autopoietic social systems.

In summary, by interpreting the autonomy of autopoietic systems in computational terms—through mechanisms that integrate both internal and external components—wide computationalism ostensibly demonstrates its compatibility with autopoietic enactivism, in keeping with its broader contention that computational systems are inherently location-neutral, able to incorporate resources from within and beyond the individual without sacrificing functional coherence. This reconciliation proceeds further by disputing the assumption that computationalism must invariably be wedded to representational or information-processing frameworks; in fact, the mechanistic strain of wide computationalism is deliberately agnostic regarding representations, rendering it more palatable to enactivists who challenge representation-centric views of cognition. By maintaining this neutral stance, wide computationalism ultimately serves as a conceptual bridge, integrating enactivist perspectives into the distributed, systemic paradigm cherished by extended cognition theorists.

3.2. The problem of proper functions

In "Computing mechanisms without proper functions," Dewhurst identifies at least one significant obstacle in Piccinini's account that could, *prima facie*, undermine its suitability for a mechanistic-computationalist version of enactivism. This obstacle stems from Piccinini's reliance on proper functions—i.e., objective teleological ends—which leads to a potential circularity: one must understand a mechanism's function in order to discern its structure, but to pinpoint that function already requires knowing how the structure is organized (Dewhurst 2018, 574). The difficulty, as Dewhurst frames it, arises when Piccinini insists that, to determine a system's mechanistic structure,

we need to establish its proper function, namely what it is "supposed" to do in an objective sense—yet identifying this proper function in turn presupposes a prior grasp of the relevant parts and their causal organization. In other words, Piccinini's theory demands an appeal to an objective teleological purpose—whether survival-oriented in nature or design-driven in artifacts—to select the components fulfilling that purpose, even though recognizing a system's purpose already seems to depend on understanding its structured composition.

That interdependence risks derailing Piccinini's goal of combining a purely mechanistic account of computation with the idea of an objective, non-historical—i.e., not merely etiological—proper function. If we cannot break out of the loop—defining function to pin down structure, defining structure to specify function—then it is unclear how we can ever decide which physical processes truly count as computations versus incidental happenings. Dewhurst's solution is to adopt Carl Craver's perspectival approach to mechanistic functions (Dewhurst 2018, 575; Craver 2013), which attributes functions according to an explanatory viewpoint rather than infusing them with fixed teleological ends. By anchoring functional ascriptions in the system's physical and causal organization, this approach avoids strict teleological commitments and circumvents circularity, all while preserving the explanatory power of computational descriptions. Even though I am in principle sympathetic to this strategy, in the next sections I will explore a related solution that I think has the advantage of being more directly connected to central ideas in the enactivist framework.

4. Deflationism about functions

As I briefly mentioned in the introduction to this chapter, Piccinini's requirement that the function of computing be objective can be understood as the requirement that "computation" be a natural kind. I will here defend the view that recognizing the artificialness of this category does not compromise the legitimacy of computational descriptions as robust explanatory tools. Moreover, it is a precondition for integration with the enactive framework. My account will avoid

pancomputationalism by showing that computational descriptions should not be arbitrarily imposed on physical systems, since they are constrained by the specific causal organization and structural regularities that those systems display, which means that valid computational attributions must reflect the stable and systematic input-output mappings supported by the physical configuration of the relevant mechanism. Moreover, I will show that this approach is corroborated by distinctions made in more canonical pieces of enactivist literature, particularly Maturana and Varela's distinction between "relations" and "interactions" in their joint work *Autopoiesis and cognition.*

4.1. PHYSICAL STRUCTURE CONSTRAINS COMPUTATIONAL ASCRIPTION

The crucial factor that prevents computational attributions from lapsing into triviality, thus avoiding the specter of pancomputationalism, is the presence of organized structural constraints in a physical system. In any putative computing mechanism, whether an engineered silicon chip or a naturally evolved neural network, certain internal causal regularities and stable interaction patterns must be demonstrably in place. These regularities reveal why only some ascriptions of computation remain viable for a given arrangement of components, even if multiple computational interpretations might be entertained in principle. In this sense, although computational descriptions are to some degree observer-relative, they are nonetheless subjected to a substantive reality check imposed by the mechanism's intrinsic causal organization.

Equally significant is the point that such organizational constraints need not be interpreted through the lens of teleological "proper functions." By abstaining from claims that a system possesses an objective, predetermined goal—such as having been designed or evolved specifically to compute—our approach ties function contextually to the manner in which the mechanism's parts interact. This stance lets us put aside worries about circularity: rather than fixing a proper function first and then delineating structure, we let the physically grounded structure, with its reliable cause-and-effect relationships, guide us in deciding which computational attributions are scientifically

illuminating. Thus, while teleological language may be useful for describing artifacts like calculators or smartphones, it is not mandatory for pinning down whether a system performs a specific mapping from input to output.

Given our broader interests here—i.e., assessing the compatibility between the enactive and the computationalist frameworks—it is important to notice that this structural perspective fits well with insights drawn from Maturana and Varela's earlier works on autopoietic theory. One of the main preoccupations of Maturana in "Biology of cognition" (1980), for instance, is precisely delimiting to what extent the autopoietic organization of living systems can be understood as dependent on the observer's role in individuating their components and functions. Maturana emphasizes that system boundaries, functional attributions, and even the very distinction between system and environment are not intrinsic features of the world but arise from the explanatory stance of an observer.

With that in mind, he distinguishes between *interactions* and *relations* between different parts of a system. Interactions are actual physical engagements that occur between entities, events that take place independently of whether or not they are observed (Maturana and Varela 1980, 8–9). In the enactive framework, these can be conceived in terms of structural coupling: the mutual determination between an organism or part thereof and its environment, with changes in one system triggering corresponding changes in the other, without necessarily involving any cognitive or intentional mediation. Relations, on the other hand, are not intrinsic to the world but rather conceptual distinctions made by the observer when attempting to describe said interactions. Again, in the enactive framework, they are confined to the domain of sensemaking.

As an illustration, consider the process of digestion in a human body: when food is consumed, it is broken down by enzymes and acids in the stomach and intestines—this is an interactional process, consisting of physical and chemical reactions between enzymes, acids, and the food molecules. These interactions, of course, occur regardless of whether anyone is observing them.

However, describing part of that process as "extracting nutrients" or "providing energy" is a relation introduced by the observer. The molecules themselves don't "intend" to be nutrients or "provide energy"—they are simply broken down through chemical interactions. The observer, however, interprets these interactions as having a functional role in sustaining the organism, based on patterns recognized in biology and nutrition, but without the observer's conceptual framework, the same interactions could just be described as a series of chemical reactions with no inherent meaning.

These distinctions, I believe, are also useful when thinking about the individuation of computational phenomena: computational attributions arise from the interplay between the observer-defined relations and the system's intrinsic, stable interactions. Hence, while a mechanism's interactions occur in the "real world," the observer's task is to impose relations that highlight consistent input–output mappings, thereby constraining the range of viable computational interpretations. This dual perspective—acknowledging both the undeniable reality of physical interactions and the observer's role in delineating relations—ensures that computational descriptions are not arbitrarily imposed. Although different researchers might emphasize alternative aspects of a mechanism's causal dynamics, only those descriptions that faithfully reflect the system's organized structure can claim explanatory validity. In effect, the physical structure provides a "skeleton" of interactions within which any observer's conceptual framework must operate, limiting the descriptive leeway and preventing the collapse into pancomputationalism.

Finally, this structural grounding prepares the way for concrete examples. By examining systems such as silicon chips with clearly delineated logic gates or neural circuits exhibiting stable patterns of activation, we can detect the specific ways in which a "computational" arrangement emerges. The interplay between the observer's relational definitions, which are necessarily limited by the need to describe stable, reproducible interactions, and the system's intrinsic causal structure ensures that computational ascriptions remain meaningful and non-arbitrary. In the next section,

we shall see how these principles play out in practice—demonstrating that while the observer's role is inescapable, it is firmly disciplined by the physical architecture of the system.

4.2. EXAMPLE: PHYSICALLY CONSTRAINED COMPUTATIONAL ASCRIPTION

In order to illustrate how a system's physical structure constrains the range of possible computational interpretations, Dewhurst presents the very elucidating example of a voltage-sensitive mechanism that processes two input voltages and generates an output voltage according to specific transition rules—i.e., a typical silicon chip embodying a logical gate. For instance, inputs of 0V and 5V might produce an output of 0V, while two inputs of 5V might result in an output of 5V, as illustrated by the first table below.

Dewhurst argues that this system can be interpreted as performing different computations, and yet its physical structure limits the range of plausible interpretations. Even though an observer might choose to represent the voltages digitally, they could also adopt an analog interpretation, mapping the voltage levels to continuous values instead of discrete ones. For the sake of simplicity, however, let us restrict ourselves to digital systems, a case in which there is already relevant interpretative plurality. In this case, assigning the value "0" to 0V and "1" to 5V would make the physical system the embodiment of an AND-gate, as shown by the second table below, which corresponds to the truth table for conjunction. Doing the opposite, however, and assigning "1" to 0V and "0" to 5V would produce an OR-gate, as shown by the third table, which corresponds to the truth table for disjunction. Despite this variability in possible interpretations, however, the observer's freedom in the assignment of computational functions to the system is not limitless. It is constrained by the system's responses to different voltage levels, and this responsiveness dictates which mappings are acceptable, preventing the arbitrary assignment of any computational function whatsoever to the system and, thus, preventing pancomputationalism.

TABLE [2]: […] [T]ransformations […]

Input	Output
0 V, 0 V	0 V
5 V, 0 V	0 V
0 V, 5 V	0 V
5 V, 5 V	5 V

TABLE [3] […]: A simple processor

String 1	String 2
0, 0	0
1, 0	0
0, 1	0
1, 1	1

TABLE [4]: Another simple processor

String 1	String 2
1, 1	1
0, 1	1
1, 0	1
0, 0	0

(Dewhurst 2018, 579–580, adapted)

Of course, when we shift our focus from engineered to natural systems like the human brain, the task of ascribing computational functions becomes much more complicated. Unlike engineered artifacts, the brain emerged through a long and complex evolutionary process, shaped by

the pressures of survival and reproduction, rather than a predefined computational blueprint. Consequently, identifying and individuating computational components within the brain is a much more challenging and interpretive endeavor. To illustrate how individuation might work in the case of the brain, consider neuroscientists investigating the visual cortex, a region responsible for processing visual information. They might observe that specific groups of neurons respond selectively to certain features of the visual input, such as edges, lines, or colors. These neurons, with their specialized response properties, could be considered as potential candidates for computational components. However, unlike the clearly defined components in the voltage-sensitive mechanism above, the boundaries and functions of these neuronal groups are not always clear-cut. Thus, individuating computational components in the brain requires not only identifying neurons with specific response properties but also understanding their functional roles within the broader network dynamics, considering their interactions with other brain areas, and accounting for the dynamic and plastic nature of neural processing.

This task involves interpretation and inference to a much larger degree, guided by neuroscientists' theoretical frameworks, experimental paradigms, and the available technologies for observing and manipulating brain activity. Thus, while the challenge of individuating computational functions in biological systems is significant, it does not undermine the broader point: computational ascriptions are constrained by physical structure and causal organization, even in complex, evolved systems like the brain. This perspective allows us to avoid the pitfalls of pancomputationalism without requiring a rigid teleological framework. Rather than treating computational function as an objective natural kind, we can understand it as a stable and systematically constrained explanatory strategy, deeply tied to the mechanistic interactions within a system.

5. Conclusion

Over the course of this chapter, I have argued that while computation remains a central explanatory framework in cognitive science, its compatibility with enactivism necessitates a fundamental re-thinking of its ontological commitments. The traditional computational theory of mind, emerging from the foundational insights of Turing and later developed through symbolic and mechanistic computationalism, posits that cognition consists of formal symbol manipulation. However, as we have seen, this view faces significant challenges, particularly in its reliance on representations, which enactivism fundamentally rejects. More recent mechanistic and non-semantic theories, such as those proposed by Piccinini, attempt to salvage computationalism by anchoring it in physical mechanisms that manipulate medium-independent vehicles. Yet, this move, while dispensing with explicit representational commitments, still implicitly relies on the notion that computation is an objective feature of the world, which, as I have argued, remains incompatible with the epistemological foundations of enactivism.

By engaging critically with Piccinini's six desiderata for a theory of computation, I have shown that while his account offers valuable insights into the mechanistic nature of computing systems, it ultimately fails to accommodate enactivism's emphasis on cognition as an emergent, relational, and observer-dependent process. The fundamental tension resides in Piccinini's insistence on computation as an objective phenomenon, a stance that enactivism, grounded in constructivist epistemology, cannot endorse. However, rather than rejecting computational descriptions outright, I have defended a perspectival and deflationary approach to computation, which, while acknowledging the observer-relativity of computational attributions, preserves their heuristic and explanatory utility within cognitive science.

In this concluding section, I shall elucidate how the theoretical framework presented here successfully meets all but one of Piccinini's reasonable desiderata—namely, the criterion of objectivity, which, as has been demonstrated in the preceding discussion, is more accurately

characterized as a falsely self-imposed demand rather than a fundamental requirement for a robust theory of computation. By expanding on each of these desiderata in a manner that preserves the explanatory depth of my argument, I aim to illustrate how a perspectival approach to computation not only remains viable but, in fact, offers a more coherent synthesis of enactivist commitments with the explanatory tools afforded by computationalism.

Let us start with the criterion of 1) *objectivity*, as formulated by Piccinini, which posits that computational individuation should be an observer-independent feature of the world, allowing for a principled and non-arbitrary classification of physical computing systems. As I have argued, this assumption rests on a misunderstanding of the nature of computation itself—one that a deflationary approach may help us correct. Rather than treating computational attributions as objective features of physical mechanisms, I have demonstrated that they are, in fact, relationally determined, emerging through the explanatory practices of observers who delineate functions in accordance with theoretical and practical considerations. This does not imply an unbridled subjectivism, for computational attributions are constrained by the structural organization and causal capacities of the systems to which they are applied. Thus, while Piccinini's criterion of objectivity cannot be fully preserved within this account, it is replaced by a more nuanced understanding of computational individuation as an observer-relative yet systematically constrained practice.

While objectivity is relinquished, 2) *explanation* is not. Computational descriptions remain indispensable heuristic instruments for understanding cognitive and neural systems, facilitating the identification of regularities and causal dependencies that would otherwise be difficult to discern. By embracing a perspectival approach, the account presented here aligns with contemporary scientific methodologies, which routinely employ models that are not necessarily ontologically committing but nonetheless remain indispensable for guiding empirical research. This approach parallels Egan's deflationary stance on mental representation, as discussed in the last chapter, in which representational ascriptions serve an explanatory function without committing to a metaphysical

claim regarding the existence of intrinsic mental content. Likewise, computational descriptions, though not reflective of inherent properties of physical mechanisms, retain their value by structuring our scientific understanding in a coherent and pragmatic manner.

Piccinini's third *desideratum* is that 3) *the right things compute*. A common concern with observer-relativity in computational individuation is that it might open the door to arbitrary attributions, leading to a trivialization of computational explanations. However, this is not necessarily the case. Physical constraints ensure that computational descriptions, though perspectival, remain meaningfully constrained. Consider a voltage-sensitive mechanism that responds to specific input-output mappings: while it may admit multiple legitimate computational interpretations—e.g., an AND-gate or an OR-gate depending on voltage assignments—, it does not permit entirely arbitrary attributions. Similarly, neuroscientific investigations into cognition rely on identifying stable and systematic neuronal activation patterns that align with specific computational frameworks. The fact that different researchers may emphasize different computational descriptions does not undermine the legitimacy of these classifications; rather, it highlights that computation is a relationally attributed structure constrained by the underlying causal organization of the system under study.

As for 4) *the wrong things do not compute*, one of the central objections to overly permissive definitions of computation is that they risk collapsing into pancomputationalism, whereby any physical system can be described as performing a computation. However, by highlighting the role of physical constraints and causal organization, I believe the approach presented here avoids such an outcome. While it is true that any system can be computationally modeled, this does not mean that all systems should be considered as physically computing in the mechanistic sense described by Piccinini. A planet's orbit, for instance, can be simulated using computational models, but this does not entail that the planet itself is performing a computation. Thus, computational ascriptions remain viable explanatory tools precisely because they are constrained by the physical architecture of the systems they describe, even if they do not correspond to intrinsic computational properties.

Another crucial aspect of Piccinini's framework is that it must account for 5) *miscomputation*—that is, cases where a system fails to function according to the computational rules it is supposed to follow. Within an observer-relative perspective, miscomputation can still be meaningfully explained: an artifact such as a calculator malfunctions when its physical components fail to produce the expected results according to its designed function. In biological systems, miscomputation can be similarly understood as the failure of stable functional mappings in neural processing. Thus, even if computational ascriptions are observer-relative, the notion of miscomputation remains well-grounded in deviations from established functional regularities. The key point is that computational functions are not arbitrarily imposed but are constrained by the interactions between the system's components, which means that miscomputations correspond to systematic failures in those interactions rather than to violations of preordained, observer-independent computational properties.

Finally, Piccinini's framework demands a principled 6) *taxonomy* of computing systems, capable of distinguishing different forms of computation in a systematic manner. This account satisfies that requirement by organizing computational ascriptions according to functional and structural constraints. It acknowledges that computational taxonomies emerge from interpretative practices but insists that these practices reflect genuine structural differences across computing systems. Digital logic gates, analog computing mechanisms, artificial neural networks, and biological neural systems each exhibit unique causal structures that shape their computational interpretations. Rather than assuming a rigid, objective taxonomy, this approach recognizes that taxonomic distinctions emerge from empirical research and remain sensitive to theoretical advancements, thus ensuring that computational classifications retain their scientific rigor without reifying computation as an intrinsic natural kind.

In sum, while Piccinini's requirement of objectivity has been critically reevaluated and found to be an unnecessary constraint, this revised framework nonetheless meets all other

desiderata in a manner that preserves explanatory rigor while remaining compatible with enactivist commitments. By acknowledging the relational and perspectival nature of computational ascriptions, this approach prevents the pitfalls of both strict objectivism and unconstrained pancomputationalism. Computational descriptions, far from being metaphysical truths about the world, function as indispensable interpretative tools that allow cognitive science to uncover and articulate the complex causal architectures underlying cognition and information processing. Thus, this deflationary yet systematically constrained view of computation provides a coherent and scientifically useful alternative to both traditional computational realism and radical anti-computationalism.

CONCLUSION

This dissertation set out to critically assess a prominent issue in contemporary philosophy of mind and cognitive science: the alleged incompatibility between enactivism and the core triad of classical cognitivism—namely, representationalism, computationalism, and functionalism. Our guiding questions were: Is enactivism truly best understood as a radical break with the cognitivist tradition? Or can it, through careful conceptual clarification and historical reflection, be brought into a more productive dialogue with the core explanatory tenets that have shaped cognitive science for over half a century? And what would this harmonization—or lack thereof—indicate about the evolving shape of the sciences of mind more broadly?

I have approached these questions along three complementary axes. First, I situated enactivism with respect to other 4E frameworks, clarifying both the shared motivations and significant points of divergence within this family of theories. Second, I traced the internal development of enactivism itself, distinguishing between its main strands—sensorimotor, autopoietic, and radical enactivism—and extracting from each their core philosophical commitments. Third, I examined the supposed tensions between enactivism and each member of the cognitivist triad, presenting not only the nature and sources of apparent incompatibility, but also the grounds and strategies for possible reconciliation—where such reconciliation is plausible. In this concluding chapter, I will synthesize my main findings, restate the arguments in relation to the original questions, and reflect on their broader significance for philosophy of mind, cognitive science, and the ongoing renewal of interdisciplinary research into cognition. I will also acknowledge the study's limitations, suggest fruitful directions for further inquiry, and close by situating these results in the wider, evolving landscape of contemporary thought on mind, body, and world.

Enactivists often define themselves through opposition, by rejecting mental computation, representation, and abstract functional roles as dominating explanatory paradigms. From this

perspective, the advent of this theoretical endeavor should mark an epochal "paradigm shift," rendering much of the previous tradition obsolete or, at best, in need of radical revision. This dissertation critically interrogated such narratives, showing that the history and conceptual genealogy of enactivism is far more entangled with the cognitivist tradition than is commonly acknowledged. The detailed analysis of 4E cognition theories provided in Part I led to a critical clarification: while the embodied, embedded, and extended approaches each propose important, often complementary modifications to the classical cognitivist picture, they remain compatible with the core tenets of mainstream cognitive science. The distinction between weak and strong variants, operationalized through the causation–constitution framework, showed that much of the apparent radicalism of 4E approaches is a matter of emphasis or conceptual scope, rather than fundamental metaphysical challenge.

Enactivism, however, genuinely disrupts this pattern. By insisting that cognition is always constituted by the ongoing engagement of brain, body, and world, it rejects any tidy separation between "internal" and "external," "causal" and "constitutive," in the very fabric of cognitive activity. Crucially, this move not only resists the weak or strong typology but also demands a rethinking of where, how, and in what sense cognition is realized and individuated. Still, despite these core commitments, the close internal analysis carried out in Chapters II and III showed that enactivism itself is far from monolithic. Its sensorimotor and autopoietic forms differ not only in theoretical scope—with the former focusing on perception, motor action, and phenomenal consciousness, whereas the latter also adds biological constitution to the list—but also in their presuppositions: autopoietic enactivism depends, in order to be properly articulated, on the theory of autopoiesis or biological autonomy, which are inessential to the sensorimotor variant. This recognition supports a call for more nuanced, internally differentiated appraisals of enactivist research programs.

After this investigation of the internal diversity of the enactive research program, the central argumentative work of the dissertation was the systematic investigation of each alleged point of

incompatibility between enactivism and the classical cognitivist triad. The main findings, addressed to the three tenets in sequence, can be summarized as follows. Concerning functionalism, we have seen that autopoietic enactivism, in particular, can be reasonably seen as a biologically-grounded, organizationally-specific form of functionalism. The function that is realized, on this view, is precisely autopoiesis—or, in more recent versions of the theory, biological autonomy. This allows for a principled version of multiple realizability—making room, at least in principle, for artificial or non-carbon-based forms of mind, provided that the appropriate organizational structures are instantiated. In this respect, enactivism does not so much overthrow functionalism as it deepens, specializes, and biologically re-anchors it. The resolution of scaling problems within autopoietic theory, such as defining boundaries in multicellular or "extended" systems, further underscores the utility of functionalist strategies in accounting for the emergence and individuation of cognitive systems. In doing that, however, functionalism itself is also transformed: it gains a rich biological grounding and a more subtle concept of organismic autonomy.

As for representationalism, drawing on Frances Egan's proposal, Chapter IV showed that the positing of representational content within cognitive science need not involve metaphysically robust entities. Instead, representational vocabulary can be retained as a pragmatic, theory-relative tool—a gloss imposed by scientific interpreters to link computational or dynamic models with observed behavioral capacities. This perspective not only resolves the "hard problem of content," as articulated by radical enactivists, but also points to a broader theoretical pluralism: researchers are free to use representational talk where it serves explanatory or heuristic purposes, without mistaking such talk for ontological commitment. This view allows enactivism to accommodate many of the working practices of cognitive science while remaining faithful to its own skepticism about intrinsic, internal content.

While classical computationalism, rooted in symbolic manipulation and the "language of thought," sits awkwardly with the enactivist's emphasis on dynamic, historical, and context-

sensitive coupling, more recent developments in the philosophy of computation—such as Gualtiero Piccinini's mechanistic, non-semantic computation—offer a path for reconciliation. Even so, this path is not trivial: enactivism resists any account that reifies computation as an observer-independent natural kind. Instead, a perspectival, functionally-constrained, and observer-relative conception of computation, anchored in structural regularities but not ontologically committing, remains open for enactivists to employ as a flexible, scientifically-rigorous modeling tool. In this light, computational models are indispensable for mapping causal and informational architecture but should not be mistaken for literal descriptions of "what the system is, in itself." Computation is a powerful, epistemically indispensable lens, not a metaphysical foundation.

The findings of this study contribute to the field in at least three significant ways. By showing that the boundaries between enactivist and mainstream cognitive science are more permeable than hegemonic readings suggest, this dissertation encourages methodological and theoretical pluralism. It urges researchers not to treat paradigmatic divides as insurmountable, but as opportunities for mutual learning, conceptual hybridization, and critical cross-fertilization. Moreover, the study advances a pragmatic attitude towards scientific entities and posits: theories, models, and vocabularies are to be judged not only by their metaphysical commitments but by their scientific, explanatory, and integrative power. Representation, computation, and function are not all-or-nothing commitments, but theoretical constructs to be used, revised, and reinterpreted in light of new discoveries, empirical results, and philosophical reflection. More broadly, this view supports the pursuit of an integrative cognitive science, where enactivist and functional-representational computationalism are not forced into zero-sum antagonism, but engage in a dynamic, critical, and reflective exchange, each probing the limits of the other.

It would be remiss, however, to suggest that the proposals presented here are final. The scope of this dissertation, however ambitious, has been necessarily circumscribed by limitations of space and disciplinary focus. Many important issues remain at the periphery of the present analysis:

the nature of consciousness and affect, the role of sociality in cognition, and the complicated relationship between cognitive science and phenomenology—all of which are deeply intertwined with the questions considered over the last five chapters. These topics undoubtedly warrant further, specialized investigation. Looking ahead, several promising directions for future research emerge.

First, there is much to be gained from a more nuanced inquiry into the metaphysics of function, organization, and normativity—one that moves beyond the individual organism to address developmental, social, and cultural scales. Second, the ongoing dialogue between cognitive science and philosophy of science merits closer attention, particularly regarding the role of models, representations, and forms of explanation in interdisciplinary research on cognition. Finally, it would be fruitful to further extend the deflationary account of mental representation, advancing from its current application at the subpersonal or subsymbolic level to a more robust treatment of representations at the symbolic or personal level. Such lines of inquiry promise not only to refine the conceptual distinctions offered in this dissertation, but also to deepen our overall understanding of mind, cognition, and the methods by which we seek to explain them.

To conclude, my central argument was that enactivism should not be viewed as a wholesale replacement for classical cognitive science, but rather as an opportunity to revisit and refine some of the foundational questions regarding mind, world, and organism. Rather than seeking premature theoretical closure, this perspective encourages a more nuanced and pluralistic approach, one that recognizes the value of positioning the embodied, situated subject as central to our accounts of cognition. In parallel, computational and representational theories achieve their greatest explanatory power when they are responsive to insights from embodiment, biologically-grounded normativity, historical context, and the broader environment in which cognition unfolds. Thus, it is my view that, in the philosophy of mind and cognitive science, progress requires a combination of conceptual clarity, methodological openness, and sustained critical analysis. It is in this spirit that the present work has sought to interrogate and, where possible, reconcile the enactivist and

cognitivist traditions, in the hope of contributing to a more integrated and productive understanding of cognition.

REFERENCES

Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. Wiley-Blackwell.

Alsmith, A., & de Vignemont, F. (2012). Embodying the mind and representing the body. *Review of Philosophy and Psychology*, 3(1), 1–13.

Anderson, M. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4), 245–266.

Ashby, W. R. (1952). *Design for a Brain: The Origin of Adaptive Behavior*. Chapman and Hall.

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.

Aydede, M. (2020). What is a pain in a body part? *Canadian Journal of Philosophy*, 50(2), 143–158.

Bach-y-Rita, P., Collins, C., Saunders, F., White, B., & Scadden, L. (1969). Vision substitution by tactile image projection. *Nature*, 221, 963–964.

Bain, D. (2003). Intentionalism and pain. *Philosophical Quarterly*, 53(213), 502–522.

Bain, D. (2007). The location of pains. *Philosophical Papers*, 36(2), 171–205.

Barrett, L. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 1, 12(1), 1–23.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–660.

Block, N. (1980). What is functionalism? In N. Block (ed.), *Readings in Philosophy of Psychology, Volume 1* (pp. 171–184). Harvard University Press.

Broadbent, D. (1958). *Perception and Communication*. Pergamon Press.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1–3), 139–159.

Bruner, J. (1966). *Toward a Theory of Instruction*. Harvard University Press.

Ceruti, M, & Damiano, L. (2018). Plural embodiment of mind. Genealogy and guidelines for a radically embodied approach to mind and consciousness. *Frontiers in Psychology*, 9.

Chalmers, D. (1998). On implementing a computation. *Minds and Machines*, 4(4), 391–402.

Chemero, A. (2009). *Radical Embodied Cognitive Science.* MIT Press.

Chomsky, N. (1959). A review of B. F. Skinner's 'Verbal Behavior.' *Language*, 35, No. 1, 26–58.

Chomsky, N. (1968). *Language and Mind.* Harcourt, Brace & World.

Chomsky, N. (2003). Reply to Egan. In L. Antony and N. Hornstein (eds.), *Chomsky and his Critics* (pp. 89–104). Blackwell.

Churchland, Patricia. (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. MIT Press.

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

Churchland, Paul. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science.* MIT Press, 1989.

Clark, A. (1997). *Being There: Putting Brain, Body, and World together Again.* MIT Press.

Clark, A. (2001). Visual experience and motor action: Are the bonds too tight? *Philosophical Review*, 110(4), 495–519.

Clark, A. (2008a). Pressing the flesh: A tension in the study of the embodied, embedded mind. *Philosophy and Phenomenological Research*, 76(1), 37–59.

Clark, A. (2008b). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.* Oxford University Press.

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

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Clark, A. (2023). *The Experience Machine: How Our Minds Predict and Shape Reality*. Pantheon.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58, 447–461.

Coelho-Mollo, D. (2020). Content pragmatism defended. *Topoi*, 39(1), 103–113.

Collins, J. (2007). Meta-scientific eliminativism: A reconsideration of Chomsky's review of Skinner's Verbal Behavior. *British Journal for the Philosophy of Science*, 58, 625–658.

Colombetti, G. (2014). *The Feeling Body: Affective Science Meets the Enactive Mind*. MIT Press.

Craver, C. (2013). Function and mechanisms: A perspectivalist account. In P. Hunemann (ed.), *Functions*. Springer.

Crick, F. (1966). *Of Molecules and Men*. University of Washington Press.

Crick, F.; & Koch, C. (1990) Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2, 263–75.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17), 6889–6892.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Degenaar, J., & Myin, E. (2014). Representation-hunger reconsidered. *Synthese*, 191(15), 3639–3648.

Demeter, T., Parent, T., & Toon, A. (eds.). (2022). *Mental Fictionalism: Philosophical Explorations*. Routledge.

Dennett, D. (1982). Styles of mental representation. *Proceedings of the Aristotelian Society*, 83, 213–26.

Dennett, D. (1987). *The Intentional Stance*. MIT Press.

Dennett, D. (1991). *Consciousness Explained*. Penguin Books.

Dennett, D. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon and Schuster.

Descartes, R. (1641/1996). *Meditations on First Philosophy* (J. Cottingham, trans.). Cambridge University Press.

Dehaene, S. (2007). *Reading in the Brain: The New Science of How We Read*. Viking/Penguin.

Dewhurst, J. (2014). Mechanistic miscomputation: A reply to Fresco and Primiero. *Philosophy and Technology*, 27, 495–498.

Dewhurst, J. (2016). Physical computation: a mechanistic account. *Philosophical Psychology*, 29(5), 795–797.

Dewhurst, J. (2018a). Computing mechanisms without proper functions. *Minds and Machines*, 28, 569–588.

Dewhurst, J. (2018b). Individuation without representation. *British Journal for the Philosophy of Science*, 69(1), 103–116.

Dewhurst, J., Deane, G., & Kersten, L. (2017). Resolving two tensions in 4E cognition using wide computationalism. In Glenn Gunzelmann, Andrew Howes, Thora Tenbrink and Eddy Davelaar (eds.), *Proceedings of the 39th Annual Conference of Cognitive Science Society*, pp. 2395–2400.

Dewhurst, J., & Villalobos, M. (2016). Computationalism, enactivism, and cognition: Turing machines as functionally closed systems. In: *International Workshop on Artificial Intelligence and Cognition*, pp. 138–147.

Dewhurst, J., & Villalobos, M. (2017a) Enactive autonomy in computational systems. *Synthese*, 195(5), 1891–1908.

Dewhurst, J.; Villalobos, M. (2017b) The enactive automaton as a computing mechanism. *Thought: A Journal of Philosophy*, 6(3), 185–192.

di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4(4), 429–452.

di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford University Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press.

Dupuy, J-P. (2000). *The Mechanization of the Mind: On the Origins of Cognitive Science*. Princeton University Press.

Edelman, G. (1989). *The Remembered Present: A Biological Theory of Consciousness*. Basic Books.

Egan, F. (2010). Computational models: a modest role for content. *Studies in History and Philosophy of Science*, Part A, 41(3), 253–259.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170(1), 115–135.

Egan, F. (2018). The nature and function of content in computational models. In: *The Routledge Handbook of the Computational Mind*. Routledge.

Egan, F. (2020). A deflationary account of mental representation. In: *Mental Representations*; J. Smortchkova, K. Dolega, and T. Schlicht (eds.). Oxford University Press.

Eisenberger, N., & Lieberman, M. (2005). Why it hurts to be left out: The neurocognitive overlap between physical and social pain. In K. D. Williams, J. P. Forgas, and W. von Hippel (eds.), *The Social Outcast: Ostracism, Social Exclusion, Rejection, and Bullying* (pp. 109–127). Psychology Press.

Fodor, J. (1968). *The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar*. McGraw-Hill.

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press.

Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2), 3–71.

Fodor, J. (1990). *A Theory of Content and Other Essays*. MIT Press.

Fodor, J. (2000). *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*. MIT Press.

Fresco, N. (2010). Explaining computation without semantics: Keeping it simple. *Minds and machines*, 20, 165–181.

Fresco, N. (2015). Objective computation versus subjective computation. *Erkenntnis*, 80, 1031–1053.

Gallagher, S. (2017). *Enactivist Interventions: Rethinking the Mind*. Oxford University Press.

Gallagher, S. (2023). *Embodied and Enactive Approaches to Cognition*. Cambridge University Press.

Gallese, V., & Sinigaglia, C. (2011). What is so special with embodied simulation. *Trends in Cognitive Sciences*, 15(11), 512-519.

Garfinkel, S., Minati, L., Gray, M. *et al.* (2014). Fear from the heart: Sensitivity to fear stimuli depends on individual heartbeats. *The Journal of Neuroscience*, 34(19), 6573–6582.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Greenwood Press.

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

Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9(3), 558–565.

Godfrey-Smith, P. (ed.) (1996). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Goldman, A. (2014). The bodily formats approach to embodied cognition. In: U. Kriegel (ed.), *Current Controversies in Philosophy of Mind*, pp. 91–108. Routledge.

Goldman, A., & de Vignemont, F. (2009). Is social cognition embodied? *Trends in Cognitive Sciences*, 13(4), 154–159.

González, J., Bach-y-Rita, P., & Haase, S. (2005). Perceptual recalibration in sensory substitution and perceptual modification. *Pragmatics and Cognition*, 13(3), 481–500.

Gould, S., & Vrba, E. (1982). Exaptation—a missing term in the science of form. *Paleobiology*, 8(1), 4–15.

Greeno, J. (1998). The situativity of knowing, learning, and research. *American Psychologist*, 53(1), 5–26.

Haugeland, J. (1981). *Semantic engines: An introduction to mind design*. In J. Haugeland, *Mind Design*. MIT Press.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. MIT Press.

Hameroff, S. (1994). Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies*, 1(1), 91–118.

Harvey, M. (2015). Content in languaging: Why radical enactivism is incompatible with representational theories of language. Review of the book *Radicalizing Enactivism: Basic Minds Without Content*, by D. D. Hutto & E. Myin. *Language Sciences*, 48, 90–129.

Hobbes, T. (1651/1996). *Leviathan* (R. Tuck, ed.). Cambridge University Press.

Hume, D. (1739/2000). *A Treatise of Human Nature* (D. F. Norton and M. J. Norton, eds.). Oxford University Press.

Hurley, S. (1998). *Consciousness in Action*. Harvard University Press.

Hurley, S. (2007). The shared circuits model: How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral and Brain Sciences*, 30(1), 1–22.

Hurley, S.; & Noë, A. (2003). Neural plasticity and consciousness. *Biology and Philosophy*, 18(1), 131–168.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Hutto, D. (2005). Knowing what? Radical versus conservative enactivism. *Phenomenology and the Cognitive Sciences*, 4, 4, 389–405.

Hutto, D. (2008). *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. MIT Press/Bradford Books.

Hutto, D. (2017). REC: Revolution effected by clarification. *Topoi*, 36, 377–391.

Hutto, D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. MIT Press.

Hutto, D.; & Myin, E. (2017). *Evolving Enactivism: Basic Minds Meet Content*. MIT Press.

Jacob, P. (1997). What minds can do: Intentionality in a non-intentional world. *Cambridge studies in philosophy*. Cambridge University Press.

Jonas, H. (1966/1982). *The Phenomenon of Life: Toward a Philosophical Biology*. University of Chicago Press.

Kant, I. (1790/2000). *Critique of the Power of Judgment*. (Translated by Paul Guyer and Eric Matthews). Cambridge University Press.

Kauffman, S. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press.

Kersten, L. (2017). A mechanistic account of wide computationalism. *Review of Philosophy and Psychology*, 8, 501–517.

Kiverstein, J. (2018) Extended cognition. In: Newen, A., De Bruin, L. and Gallagher, S. (eds.), 19–40. *The Oxford Handbook of 4E Cognition*. Oxford University Press.

Klein, C. (2015). *What the body commands: The imperative theory of pain*. MIT Press.

Kross, E., Berman, M., Mischel, W., Smith, E., & Wager, T. (2011). Social rejection shares somatosensory representations with physical pain. *Proceedings of the National Academy of Sciences*, 108(15), 6270–6275.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books.

Lakoff, G., & Johnson, M. (2002). Why cognitive linguistics require embodied realism. *Cognitive Linguistics*, 13(3), 245–263.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50(3), 249–258.

Llinas, R. (1987). 'Mindness' as a functional state of the brain. In *Mindwaves*, Colin Blakemore and Susan Greenfield (eds.), 339-58. Blackwell.

Llinas, R., & Ribary, U. (1993). Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences USA*, 90(5), 2078–81.

Margulis, L. (1993). *Symbiosis in Cell Evolution: Life and Its Environment on the Early Earth*. W. H. Freeman.

Marr, D. (1982). *Vision*. MIT Press.

Martínez, M. (2011). Imperative content and the painfulness of pain. *Phenomenology and the Cognitive Sciences*, 10(1), 67–90.

Mach, E. (1886/1997). *The Analysis of Sensations and the Relation of the Physical to the Psychical* (Translated by C. M. Williams and S. Waterlow). Open Court Publishing Company.

Matthen, M. (2014). Debunking enactivism. *Canadian Journal of Philosophy*, 44(1), 118–28.

Maturana, H., & Varela, F. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company.

Maturana, H., Varela, F. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Shambhala.

McBeath, M., Shaffer, D., & Kaiser, K. (1995). How baseball outfielders determine where to run to catch fly balls. *Science*, 268(5210), 569–573.

Menary, R. (ed.). (2010). *4E cognition: Embodied, embedded, enacted, extended*. Special issue of *Phenomenology and the Cognitive Sciences*, 9(4).

Miller, G. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, 7(3), 141–144.

Miller, George A., Eugene Galanter, and Karl H. Pribram. (1960). *Plans and the Structure of Behavior*. Holt, Rinehart and Winston.

Miłkowski, M. (2011). Beyond formal structure: A mechanistic perspective on computation and implementation. *Journal of Cognitive Science*, 12, 359–379.

Miłkowski, M. (2015). The hard problem of content: solved (long ago). *Studies in Logic, Grammar and Rhetoric*, 41(1), 73–88.

Miłkowski, M. (2018). From computer metaphor to computational modeling: The evolution of computationalism. *Minds and Machines*, 28, 515–54.

Minsky, M. (1967). *Computation: Finite and Infinite Machines*. Prentice-Hall.

Monod, J. (1971). *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. (Translated by Austryn Wainhouse). Alfred A. Knopf.

Morowitz, H. (1968). *Energy Flow in Biology: Biological Organization as a Problem in Thermal Physics*. Academic Press.

Moyal-Sharrock, D. (2019). From deed to word: gapless and kink-free enactivism. *Synthese*, 198(1), 405–425.

Müller, J. (1838). *Handbuch der Physiologie des Menschen für Vorlesungen*. Hölsche.

Neisser, U. (1967). *Cognitive psychology*. Appleton-Century-Crofts.

Newell, A., & Simon, H. (1972). *Human Problem Solving*. Prentice-Hall.

Newell, Allen, Shaw, J., & Simon, H. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65(3), 151–166.

Newen, A., de Bruin, L., & Gallagher, S. (eds.) (2018). *The Oxford Handbook of 4E cognition*. Oxford University Press.

Noë, A. (2004). *Action in Perception*. MIT Press.

Noë, A. (2009). *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness*. Farrar, Straus and Giroux.

Noë, A. (2012). *Varieties of Presence*. Harvard University Press.

Noë, A. (2021). The enactive approach: A briefer statement, with some remarks on 'radical enactivism'. *Phenomenology and the Cognitive Sciences*, 20, 957–970.

Noë, A, & O'Regan, J. K. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5), 939–973.

O'Regan, J. K. (2011). *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*. Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Pfeifer, R., & Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press.

Pfeifer, R., & Iida, F. (2006). Morphological computation: Connecting body, brain and environment. In *Lecture Notes in Computer Science*, Vol. 3853, 3–4. Springer.

Pfeifer, R., Iida, F., & Gómez, G. (2006). Morphological computation for adaptive behavior and cognition. *International Congress Series*, 1291, 22–29.

Piccinini, G. (2006). Computation without representation. *Philosophical Studies*, 137, 205–241.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford University Press.

Putnam, H. (1967a). Psychological predicates. In *Art, Mind, and Religion*, W. H. Capitan and D. D. Merrill (eds.), 37–48. University of Pittsburgh Press.

Putnam, H. (1967b). The nature of mental states. In *Mind, Language and Reality*. Cambridge University Press.

Putnam, H. (1987). *Representation and Reality*. MIT Press.

Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge University Press.

Rock, I., & Harris, C. S. (1967). Vision and touch. *Scientific American*, 216(5), 96–104.

Roll, J-P., & Roll, R. (1988). From eye to foot: A proprioceptive chain involved in postural control. In G. Amblard, A. Berthoz, & F. Clarac (eds.), *Posture and Gait: Development, Adaptation, and Modulation*, 155–164. Excerpta Medica.

Rosen, R. (1991). *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. Columbia University Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.

Rowlands, M. (2006). *Body Language: Representation in Action*. MIT Press.

Rowlands, M. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. MIT Press.

Rumelhart, D, & McClelland, J. (1986a). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. MIT Press.

Rumelhart, D., & McClelland, J.. (1986b). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. MIT Press.

Rupert, R. (2010). Extended cognition and the priority of cognitive systems. *Cognitive Systems Research*, 11(4), 343–356.

Russell, B. (1918/2009). *The Philosophy of Logical Atomism*. Routledge.

Searle, John R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Searle, J. (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64(3), 21–37.

Searle, J. (1999). The Chinese room. In R.A. Wilson and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–57.

Shapiro, L. (2005). *The Mind Incarnate*. MIT Press.

Shapiro, L. (2011). *Embodied Cognition: New Problems of Philosophy*. Routledge.

Schrödinger, E. (1944). *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press.

Shapiro, L., & Spaulding, S. (2021). Embodied cognition. Edited by Edward Zalta. *The Stanford Encyclopedia of Philosophy*. URL=https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, 55, 349–74.

Singer, W; & Gray, C. M. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences USA*, 86(5), 1698–702.

Skinner, B. F. (1953). *Science and Human Behavior*. Macmillan.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: six lessons from babies. *Artificial Life*, 11(1–2), 13–29.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11, no. 1, 1–23.

Sterelny, K. (2010). Minds: Extended or scaffolded? *Philosophical Studies*, 157(3), 367–390.

Suchman, L. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

Thelen, E., & Smith, L. B. (eds.) (1993). *A Dynamic Systems Approach to Development: Applications*. MIT Press

Thelen, E., & Smith, L. B (eds.). (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.

Thompson, E. (2015). *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*. Columbia University Press.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

Toon, A. (2023). *Mind as Metaphor: A Defence of Mental Fictionalism*. Oxford University Press.

Turing, A. (1936). On computable numbers, with an application to the *Entscheidungsproblem*. *Proceedings of the London Mathematical Society*, S2–42(1), 230–265.

Tye, M. (1995). A representational theory of pains and their phenomenal character. In *Philosophical Perspectives. Vol. 9: AI, Connectionism and Philosophical Psychology*, pp. 223-239.

van den Herik, Jasper C. (2020). A twofold tale of one mind: revisiting REC's multi-storey story. *Synthese*, 198(12), 12175–12193.

van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92/7, 345–381.

Varela, F. (1986). *Experimental Epistemology*. Cahiers du CREA 9, 107–121.

Varela, F. J. (1979). *Principles of Biological Autonomy*. North-Holland.

Varela, F. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition*, 34(1), 72–87.

Varela, F., & Thompson, E. (2001). Radical embodiment. *Trends in Cognitive Science*, 4, 418–425.

Varela, F., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Varela, F.; Maturana, H.; & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*, 5(4), 187–96.

Varela, F., & Weber, A. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, 1(2), 97–125.

Virchow, R. (1860). *Cellular Pathology: As Based Upon Physiological and Pathological Histology*. (Translated by Frank Chance). John Churchill.

von Foerster, H. (1973). On Constructing a Reality. *Environmental Design Research*, 2, 35–46.

von Foerster, H. (2003). *Understanding Understanding: Essays on Cybernetics and Cognition*. Springer.

von Neumann, J. (1966). *Theory of Self-Reproducing Automata.* (Edited and completed by Arthur W. Burks). University of Illinois Press.

von Uexküll, J. (1934/2010). *A Foray into the Worlds of Animals and Humans, with A Theory of Meaning.* (Translated by Joseph D. O'Neil). University of Minnesota Press.

Ward, D., Silverman, D., & Villalobos, M. (2017). Introduction: The varieties of enactivism. *Topoi*, 36(3), 365–75.

Ward, D., & Stapleton, M. (2012). Es are good. Cognition as enacted, embodied, embedded, affective and extended. In Fabio Paglieri (ed.), *Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness*. John Benjamins Publishing.

Watson, J. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20, no. 2, 158–177.

Weiskrantz, L. (1997). *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford University Press.

Wheeler, M. (2005). *Reconstructing the Cognitive World*. MIT Press.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

Wiener, N., Rosenblueth, A., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10(1), 18–24.

Wilson, R. (1994). Wide computationalism. *Mind*, 103(411), 351–72.

Young, B. (2017). Enactivism's last breaths. In *Contemporary Perspective in the Philosophy of Mind*, edited by M. Curado and S. Gouveia. Cambridge Scholars Press.