RADICALIZING THE RADICAL:

TOWARDS A RADICALLY EXTERNALIST BIOLOGY OF COGNITION

MUHAMMED YUNUS ŞAHİN

BOĞAZİÇİ UNIVERSITY

2025

RADICALIZING THE RADICAL:

TOWARDS A RADICALLY EXTERNALIST BIOLOGY OF COGNITION

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Cognitive Science

by

Muhammed Yunus Şahin

Boğaziçi University

2025

Radicalizing the Radical: Towards a Radically Externalist Biology of Cognition

The thesis of Muhammed Yunus Şahin

has been approved by:

Assoc. Prof. İnci Ayhan ____________________________
(Thesis Advisor)

Haydar Oğuz Erdin ____________________________
(Thesis Co-Advisor)

Assoc. Prof. Lucas Thorpe ____________________________

Prof. Yağmur Denizhan ____________________________

Prof. Mark Michard ____________________________
(External Member)

May 2025

DECLARATION OF ORIGINALITY

I, Muhammed Yunus Şahin, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature……………………………………………...

Date………………………………………….............

ABSTRACT

Radicalizing The Radical: Towards a Radically Externalist Biology of Cognition

In this thesis, I outline a radically externalist approach to cognitive science, explain why such an approach is needed, and discuss the potential benefits it offers. I examine the dominant divisions in the field, i.e. internalism vs. externalism and representationalism vs. anti-representationalism, and track the argumentative steps that yield these positions. Along the way, I identify key background assumptions, particularly realism, multiple realizability, and the autonomy of the special sciences. I argue that these assumptions underlie most theoretical frameworks in cognitive science and that they ultimately lead to the problem of mental causation and the hard problem of content. I claim that these problems disappear once we abandon those assumptions. In their place, I defend an anti-realist stance, draw on the new mechanistic philosophy of science, and propose a new tool: the Naturalistic Coherence Criterion. I argue that this framework points us toward a radically externalist cognitive science, one that remains agnostic about the metaphysics of mind and avoids unnecessary theoretical burdens.

ÖZET

Radikali Radikalleştirmek: Radikal Dışsalcı Bir Biliş Biyolojisine Doğru

Bu tezde bilişsel bilime radikal dışsalcı bir yaklaşımın anahatlarını, neden gerektiğini ve ne gibi faydaları olabileceğini açıklıyorum. Bilişsel bilimdeki farklı çerçevelerin etrafında kümelendiği içselcilik-dışsalcılık ve temsilcilik-temsil karşıtlığı hatlarının argümentatif adımlarını takip ediyor ve varsayımlarını tespit ediyorum. Bu varsayımların başında realizm, çoklu gerçekleşebilirlik ve özel bilimlerin otonomisi geliyor. Bu varsayımların bilişsel bilimdeki hemen hemen tüm çerçeveler için zihinsel nedensellik ve içerik zor problemine yol açtığını gösteriyorum. Bu varsayımların yanlış olduğunu savunuyor, bu varsayımları terk ettiğimizde zihinsel nedensellik ve içerik zor probleminin, bilişsel bilimci için ortadan kalktığını iddia ediyorum. Bu varsayımlar yerine benimsediğim anti-realizm, bilim felsefesinde yeni mekanistikçilik ve bu tezin yeni bir önerisi olan Natüralistik Tutarlılık Kriterini açıklıyorum. Tüm bu değerlendirmenin bizi zihnin metafiziğine agnostik, radikal dışsalcı bir bilişsel bilime yönelttiğini savunuyorum.

ACKNOWLEDGMENTS

I am first and foremost grateful to my father, Cumhur Şahin and my mother, Selma Şahin. For they always supported me in pretty much every decision I took and never criticized me for not pursuing anything financially profitable, which was the common practice in our community, thereby enabling a free intellectual journey for me. I am sincerely thankful to my sister, Sümeyye, my partner in crime in life and laughter for standing by me through the highs and lows of this journey, and your quiet faith when I could not find mine.

I am thankful to my past teachers, especially to Alirıza Çakır, Yaşar Durmaz, Sinan Can Yücel and Ekrem Ayyıldız for they have been more than a teacher and more of a role-model, broadening a teenager's horizon and showing him what is in the world out there.

I am sincerely thankful to my advisor Esra Mungan, for she always believed in me, always encouraged me to express my radical and speculative views even when they were quite opposite of hers, and created a free and comfortable experience for me. She has been always there, ready to listen, validate and support even when in my quite grumpy times. I am aware that the experience I have been through is unique in today's academy and I am pretty sure, if it was not for her, this study could not come true. She is not only a great mentor but also a great human being, exemplifying with her personality what I should strive for in being a 'person'.

I express my heartfelt gratitude to H. Oğuz Erdin, my co-advisor, for presenting me new horizons in philosophy, supporting me in the academic organizations that we organize together, and encouraging me to think radically and take philosophy and science seriously. As himself being a case, he also showed me frivolous decisions are not that bad at all and not every academic is boring.

I want to express my special gratitude to Boğaziçi University as a community, as an institution, *as home* and to all my friends there who welcomed and helped me build my *home* with them. To the risk of forgetting someone, I want to thank Ateş Metin, Nevzat Ersoy, Gülfem Sarı, Halit Özsoy, Zehra Erkoç, Alihan Koçak, Sarp Kargı, Mehmet Ali Karataş, Hesna Yılmaz, Begüm Güven, Seyit Ünalan, Elif Canseza Kaplan and many more for they showed me, to my surprise, it is possible to find your home even as a grown up adult. I want to thank my friends from CogIST, especially to Maharrem Ayar, Mehmet Tekin Soyer, Furkan Avcı, Yunus Emre Karaman, İbrahim Ethem Deveci, Neslihan Turan, Mert Can, Helin Erden, Ongun Kılıç, Miray Morova, İlknur Eliş and more for they kept me believing in the value of science communication and creating a common good together. Without their companionship, I could not stay motivated to pursue such a research and without our endless discussions, my opinions would not mature in this way.

Lastly, I thank cognitive science program at Boğaziçi University, especially İnci Ayhan, Lucas Thorpe, Yağmur Denizhan and Emre Uğur for their invaluable contributions to my intellectual development. They became life-long role-models as academics. I am very glad we have met and I had the opportunity to learn from them.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Someone looking at cognitive science from a distance today would likely see a fragmented field. When trying to understand its theoretical structure, they might notice a rough division: on one side are computationalist approaches inspired by computer science, and on the other are embodied cognition frameworks rooted in biology. Some schools of thought would seem to fall somewhere in between, attempting to integrate both perspectives, while others appear to be in sharp contrast, positioned far apart from one another.

Many disputes between different schools of thought revolve around if a theoretical construct, an entity or a posit is real or not. For example, philosophers and cognitive scientists have been discussing if mental or neural representations are real (Fodor & Pylyshyn, 1981; Thompson & Piccinini, 2018; Piccinini, 2022; Ramsey, 2007; Şahin & Deveci, 2023; Van Gelder, 1995). If they are real, what is the nature of them? Is representational content a real entity (Schulte, 2023)? Is mind or/and brain really a computer (Chalmers, 2010; Richards et al. 2019)? Do our mind/brain really encode and store the things outside of it (Bickhard & Richie, 1983; Brette, 2019)? Are really there modules in our mind (Carruthers, 2006; Fodor, 1983)? How much of computation assumed or proposed in our models is *actually* taking place? For example, is there really activation function(s) in brain? Do our neurons really calculate probabilities or really take derivatives (Haefner et al. 2024; Marblestone, Wayne & Kording, 2016; Son et al. 2020; Richards et al. 2019)? Do all organisms really have agency or only animals, or humans (Moreno & Etxeberria, 2005; Walsh, 2015)? Is autopoiesis real (Maturana & Varela, 1973/1980; Varela, 1979)? Is free will real (Mitchell, 2023; Sapolsky, 2023)? Is consciousness real (Dennett, 1991; Frankish, 2016; Seth, 2021)? These are the questions of the sort that distinguish varying schools of thought, different frameworks, different ontologies. Cognitive scientists and philosophers can be said to be 'from' such and such a school of thought or framework depending on the answers they give to such questions. People usually choose some framework over another depending on what they take to be real and not.

Among these questions, one stands out as the most significant and influential: whether mental representations are real, and if they are, what exactly their nature is. Since Brentano, representations have been regarded as intrinsic to the nature of mentality, what distinguishes the mental from the non-mental. Given that cognitive science claims to study the mind, the answers offered to this question have shaped the broader assumptions and methods of various schools of thought, often leading to deep and fundamental differences.

Recently, though, Wheeler (2017) claims that the real dispute is not between representationalism versus anti-representationalism but rather between internalism and externalism, as externalism taken as the thesis that "the material realizers (or vehicles) of cognitive states and processes are distributed over brain, body and world, and so are located partly outside the skin." Expanding on Wheeler (2017), Villalobos and Silverman (2018) identifies the root cause here as *systemic asymmetry.* In their view, even in the anti-representationalist approaches like radical enactivism (Hutto & Myin, 2012), it has been the common practice to propose something as 'mark of the mental/cognitive' to distinguish it from the non-cognitive, i.e. to answer the question of what constitutes cognition. Only then, for the common practice, we could know what the relevant phenomena are and how cognition produces them whether it is embodied or computational. However, Villalobos and Silverman claims, this only resulted in the *systemic asymmetry* in which cognitively constitutive parts carry the explanatory burden and become primary against the non-constitutive causes, hence opens up a space for possible internalist interpretations where only the boundary shifts but the division itself remains. That undermines the coherence of embodied approaches and make them potentially vulnerable to pitfalls of other internalist accounts which they are critique of. To avoid this, a secure, a complete externalism is needed. Even though anti-representationalism is a prerequisite for such an account, it is not adequate by itself.

In this thesis, following the sequence of the discussion outlined above, I will address the question of whether it is meaningful to debate the “reality” of a given theoretical entity in science and, more specifically, in cognitive science. In this context, the second chapter will be dedicated to the debate between scientific realism and anti-realism. I will propose a form of radical anti-realism, one that applies not only to scientific theories but also to everyday or folk ideas about what is real. Of course, rejecting realism comes with challenges. If we give up the idea that science tells us what is truly real, how do we evaluate theories? How do we decide which concepts are worth using? To address these questions, I will turn to the New Mechanistic philosophy of science.

In the third chapter, I will examine how the new mechanist approach conceptualizes science. In doing so, I will distinguish between its broad and narrow interpretations. According to the broad interpretation, New Mechanism aims to account for science as a whole and provides a normative framework for how science ought to be conducted. In contrast, the narrow interpretation treats it as one of several alternative approaches, more suitable for specific domains of science rather than universally applicable. This thesis adopts a neutral stance with respect to this distinction. Following this, I will address the question of whether New Mechanism is compatible with anti-realism. Although many mechanist philosophers are realists, I will argue that no such commitment necessarily follows from the core assumptions of the new mechanist framework. On this view, what matters is not whether mechanisms are real in some metaphysical sense, but whether they enable us to give explanations that are both conceptually and empirically coherent.

In the fourth chapter, I will examine prevailing practices in cognitive science. In doing so, I will focus on functionalism and argue that, due to the influence of the ideas of the *autonomy of special sciences* and *multiple realizability*, nearly all existing schools of thought in cognitive science, despite their many differences, ultimately diverge in terms of their commitments to the metaphysics of mind. I will offer a critique of both the autonomy of special sciences and multiple realizability, contending that these views lack sufficient grounding. Building on the discussions developed in the second and third chapters, I will then introduce the concepts of *Conceptual Coherence* (CC) and *Naturalistic Coherence* (NC) as the core proposal of this thesis. Drawing from these, I will propose the *Naturalistic Coherence Criterion* (NCC) as a normative standard to be employed both in the introduction of new concepts and in the evaluation of existing conceptual frameworks within naturalistic sciences. I will argue that, when combined with the commitments of New Mechanism, this criterion leads us toward an approach to cognitive science that does not center metaphysical commitments about the mind, but instead treats them as secondary and *a posteriori* concerns. To support this claim, I will draw on both a historical case from developmental biology and a thought experiment.

In the final chapter, I will clarify what this thesis does and does not claim, in order to avoid potential misunderstandings. I will briefly highlight how the approach developed here diverges from existing schools of thought within cognitive science, and I will outline possible directions for further research.

With this thesis, I aim to outline both the necessity and the contours of what may be called a *radically externalist* cognitive science. By this, I refer to an approach that makes no intrinsic attribution and reference to the mind, remains entirely agnostic with respect to the metaphysics of mind, and is thus "externalist" in this broader sense. Although at such a radical level of externalism there may be no room left for any kind of "internalism" to the extent that the very distinction between internal and external collapses, though the label remains, in my view, a convenient and useful one. In this context, it may be helpful to provide an example of an internalism–externalism spectrum in order to clarify the position of the view defended here. On one end of this spectrum, we find phenomenology as the most internalist approach, while on the other end lies the radically externalist position. This is not because phenomenology is committed to locating the mind "in the head" or treating it as something exclusively tied to the brain. In fact, phenomenology explicitly avoids such commitments. Rather, its position on the internalist pole stems from its methodological orientation. Phenomenology centers the mental, intentionality, and first-person experience as the primary focus of analysis, and resists reducing these to spatiotemporal processes. This might be called the intensionalist strategy of phenomenology. By contrast, what makes radical externalism "radical" is not a straightforward reduction of the mental to spatiotemporal processes either, it also avoids grounding the mental in such terms. However, it does so by adopting an extensionalist methodology that shifts the focus entirely away from mental phenomena. Accordingly, when I refer to radical externalism, I am also taking into account a methodological criterion — intensional–extensional distinction — for evaluating and positioning views along the internalism–externalism spectrum. This axis distinguishes approaches not merely by where they locate the mind, but by how they treat the status and explanatory role of mental concepts within scientific inquiry.

CHAPTER 2

WHAT CAN WE EXPECT FROM SCIENCE:

THE CASE AGAINST SCIENTIFIC REALISMS

Shapin and Shaffer (1985), Latour (1990) and Stengers (2023) vividly tell the story of how sociopolitical crises played an important role in the so-called scientific revolution. According to their narrative, the need to create a consensus led people to consult an authority, if possible an authority that will be powerful enough to settle the disputes and convince everyone, so that crises like bloody wars may come to an end. Nature itself, the objective reality, devoid of any subjective opinion emerged as such an authority and scientists were to become the bearers of that authority as people who can speak with nature through their experiments, delivering nothing less than *facts* and *truth*. This is the basis of modern science's *epistemic sovereignty*.

The status of science delivering *facts* and *truth* has been dominant in academy and the public. Pretty much anything quickly becomes a matter of *evidence* or *data* and if not supported by them it is quickly disregarded as something speculative or even dogmatic. But of course, there were disagreements, too. Especially the twentieth century witnessed important debates around what science can tell us.

In this chapter, I will first summarize the realist stance and its arguments. Then I will explain the criticisms directed at the realist stance by the anti-realist one and its overall arguments, and I will propose that one should take an anti-realist stance. In doing that, I will tap into scientific realism, selective realisms, constructive empiricism, pessimistic argument, underdetermination, perspectival realism, sociological and historical approach to science and practice approach to science.

Note that, my aim in this chapter is mainly to answer the question of "what can we expect from science." The answer to this question will later constrain our more local questions and answers in particular fields of sciences, for the cases of this thesis, in cognitive science.

2.1 Setting the stage for scientific realism

The controversy as to how we should think about the reality of our theories goes at least back to the discussion between Copernicus and Bellarmine in 17th century (Feyerabend, 1975). Likewise, 19th century witnessed a heated debate around the similar questions and concerns, Duhem, Poincare and Mach being at the centre stage (Duhem, 1991; Mach, 2013; Poincare, 2014). The debate continued into 20th century especially with the drive of novel claims by logical positivism (Richardson and Uebel, 2007) and it is later taken up by Putnam, Boyd and others (Boyd, 1973, 1983; Putnam, 1962, 1975a, 1975b).

As is the case with most topics in philosophy, it has a long history that cannot be dealt adequately here in a single thesis, or even worse, in a single section of a thesis. Therefore, in this section I will try to confine the discussion in plausible limits and start off with the recent and more relevant formulations of the discussion.

Scientific realism is basically the idea that scientific theories are true (or approximately so) and/or they give us, to put a bit metaphorically, the 'picture' of a mind-independent reality (Chakravartty, 2017). At the first glance, one might think the fact that scientific theories are fallible contradicts this idea and it cannot be correct, but of course scientific realists acknowledge that fact, which is the reason for the mention of "approximate truth." Note that this is most importantly a claim about *theoretical terms* as opposed to *observational terms* which means the terms refer to things that we can observe — construed in a broad sense here — whereas the former means the terms that we employed to explain the *observational terms* (Chakravartty, 2017). This distinction itself has a long history and a literature full of various stances which is beyond the scope of this thesis.[1] However, suffice it to say that there are scholars who use *theoretical terms* as including any term that is employed by scientific theories whether observational or not (Chakravartty, 2017). Relatedly, some prefer *observable*/*unobservable* distinction as more apt to the situations and take the discussion as mainly revolving around unoservables, which are entities that we cannot 'directly observe.' Note that, though, there is no consensus on how we should construe the meaning of 'direct observation.' But for most people electrons or many components of String Theory are good and accepted examples of unobservables.

Scientific realism has three main commitments: *ontological or metaphysical, epistemological* and *semantic*. Ontological one is the commitment that *there is a mind-independent reality* whereas epistemological commitment states that *we can and do come to know that independent reality through science*. Lastly, semantically, realists commit to the view that *scientific statements are and should be construed literally (as referring to the mind-independent reality)*. Discussions of scientific realism have plenty of different stances depending how one formulates these commitments more specifically.

Hilary Putnam was one of the most important analytic philosophers of the 20th century, and his ideas on these are foundational for contemporary debates on the topic. Putnam takes the semantic commitment as crucial to secure the others and gives a causal-historical theory of reference. According to the model that will later

[1] For more details, see: (Churchland, 1985; Hacking, 1985; Teller, 2001; Van Fraassen, 1985; 2005)

come to be also known as Putnam-Kripke Model (Kripke, 1980; Putnam, 1973, 1975b) a name's reference, its extension, is fixed at the moment of 'baptism', namely the moment the name is given to the relevant object by convention. That name then becomes 'the rigid designator' of its object and it is passed onto other people and other generations which links all the later references causally to the first act of naming, so-called the moment of 'baptism'. For a better understanding of the underlying idea, we can consider the famous Twin Earth thought experiment proposed by Putnam (1973). Suppose that there is another earth which is exactly the same with that of ours except for just a small difference: 'water' is not constituted by H2O there but by some mixture that is unknown to us called, hypothetically, XYZ. Notice that this small difference does not change anything in terms of water's appearance or how it is sensed, perceived, etc. So, Putnam argues, if it were for the appearance of 'water', or how it is perceived, sensed or conceptualized by us that is determining the reference, then we would be referring to the same thing both in our earth and the twin earth even if their extensions are not the same. However, this poses a contradiction since different extensions lead to different truth-conditions, which in turn contradicts the mention of 'the same reference.' This would also pose a challenge for 'mind-independence' as here the reference would be mind-dependent (depending on how 'water' is sensed, perceived, conceptualized, etc.) One further consequence of this, for example, is that we would not be able to refer to the same thing throughout history or across different cultures when using the same 'names' which, in turn, obviously undermines the claims of scientific realism. Consequently, to solve this, Putnam argues that the meaning, the reference of a name must be fixed causally which makes it determined by the object's nature, it's mind-independent causal powers becaue only then would it be possible to refer the same thing with the same name throughout history and across cultures. Accordingly, for Putnam, science discovers the appropriate extensions of natural kinds. Water is a *natural kind*, in this sense, as it is fixed by H2O's causal powers.[2] What is true and false for H2O is also true and false for water. This is how science reveals a *true* picture of the mind-independent reality.

It may be useful to elaborate further on this point, especially since the ideas discussed here might appear incompatible with Putnam's functionalist views, which we will address later. Putnam might be expected to argue that the meaning of water emerges from its causal relations and emergent structure. In fact, some commentators have derived precisely this conclusion from the Twin Earth thought experiment (Searle, 1983). Putnam's main point here is an attempt to demonstrate that certain concepts used across different scientific theories may refer to "the same thing," even if the theories themselves, and the semantic frameworks that give rise to them, are different. In this way, Putnam seeks to show that we can talk about the same entities while subscribing to different theories and explanations concerning them, and that the meanings of these entities do not originate from the theories themselves. While explaining the thought experiment, Putnam notes that if an Earthling were to travel to Twin Earth and discover that the substance called "water" there is not H_2O but XYZ, they would report: "On Twin Earth, the word 'water' means XYZ." Symmetrically, a Twin Earthling visiting Earth would say, "On Earth, the word 'water' means H_2O." Similarly, from this perspective, drawing on Kripke, it can be argued that the sentence "Bertrand Russell is a famous chef" is false, even if the speaker mistakenly believes that Russell was a famous chef rather than a renowned philosopher. What makes the sentence false is not the speaker's internal

[2] I am aware that this is not the only account of natural kindhood. I do not have enough space to dive into that discussion, though suffice it to say that there are theories of natural kinds that are compatible with anti-realism.

conceptualization or belief, but the fact that the expression "Bertrand Russell" refers to the philosopher, regardless of the speaker’s mental concepts.

Two points need to be emphasized here. First, this view, much like Frege’s, draws a distinction between sense and reference, and it treats reference as constitutive of meaning. For example, people who, for years, did not know that the Morning Star and the Evening Star are the same celestial object, Venus, nonetheless used both expressions and the referent of these expressions remained Venus, irrespective of that ignorance. In a similar fashion, for readers less familiar with the philosophy of language, it may be helpful to briefly reference the intension–extension distinction. In this context, extension refers to the things an expression picks out, while intension refers to the manner in which it picks them out. In rough terms, extension corresponds to reference, and intension to sense. To illustrate, imagine a room containing 20 cube-shaped objects and 10 round ones. By coincidence, the round objects are colored green and blue, while all the cubes are red, and nothing else in the room is red. In this case, both “the red objects in the room” and “the cube-shaped objects in the room” are expressions with different intensions, but their extensions are identical. The second point I wish to stress is that reference, that is, on this account, meaning, is not individual, but collective and historical. This is one of the most commonly misunderstood aspects of Putnam’s and Kripke’s theories. The core claim here is that when a term is introduced, regardless of the reason behind it introduction, and begins to circulate within a linguistic community, either orally or in writing, its reference remains causally fixed by that initial usage, even if all associated senses later change. Precisely for this reason, someone using the term with a particular sense or intension may still succeed in referring to the original object, even if they are unaware of its reference. Likewise, even if the individual who first introduced the term did so with an incomplete or mistaken sense or intension, the referent remains fixed in a collective and historical manner.

However, this account has been criticized in different ways. We can count three of them shortly: 1) Metaphysics of chemistry suggests that we cannot claim water = $H2O$ since it is never that simple and there are many more factors and constituents (VandeWall, 2007). 2) Many scientific realists do not share this commitment of correspondence and find it unnecessary for scientific realism. For example, Dupre, makes a case against natural kinds using biological taxonomy as a counter-example (Dupré, 1993). 3) There is an entire different tradition in philosophy of language with a different account of language, especially since Wittgenstein's *Philosophical Investigations* (Hallett, 1991; Hanfling, 1984, 2000; Loomis, 2017). This tradition equates meaning with use, which is context-dependent, and rejects 'natural kinds' as neither necessary nor possible.

An elaboration for the third point is in order here. The concept of *language games* is central to Wittgenstein and Wittgensteinian philosophy. Accordingly, language is constituted by the act of speaking in a particular situation where the situation itself provides the relevant rules for the act. The act of speaking, in this sense, is continuous with other forms of action and the relevant rules are not explicit most of the time, just like many forms of *games* we encounter. Therefore, the act of speaking can be construed as a 'move' in a game, where the game itself, namely which game it is, its rules, etc. can only be apparent or emergent locally and contextually albeit only vaguely. Thus, *mind-independence* or *truth* becomes irrelevant here.[3] Perhaps, the following passage from *Philosophical Investigations* can be more explanatory [4] (Wittgenstein, 1953, aphorism 23):

> 23. But how many kinds of sentence are there? Say assertion, question, and command?—There are countless kinds: countless different kinds of use of what we call "symbols," "words," "sentences." And this multiplicity is not something fixed, given once for all; but new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten. (We can get a rough picture of this from the changes in mathematics.)
>
> Here the term "language-game" is meant to bring into prominence the fact that the speaking of language is part of an activity, or of a form of life. Review the multiplicity of language-games in the following examples, and in others: Giving orders, and obeying them, describing the appearance of an object, or giving its measurements, constructing an object from a description (a drawing), reporting an event, speculating about an event ... forming and testing a hypothesis, presenting the results of an experiment in tables and diagrams, making up a story; and reading it, play-acting, singing catches, guessing riddles, making a joke; telling it, solving a problem in practical arithmetic, translating from one language into another, asking, thanking, cursing, greeting, praying...

As mentioned above, this way of thinking about language opened up new ways and there is a rich literature growing day-by-day where Putnam's claims regarding natural kinds and its semantics are challenged.

On the other hand, causal-historical theory of reference or Putnam-Kripke model is not Putnam's only argument in favor of scientific realism. Putnam put forward another famous argument called "No Miracle Argument" or "Miracle Argument" which has been fundamental. The name refers to the Putnam's statement that realism "is the only philosophy that does not make the success of science a *miracle*" (Putnam, 1975). To put it briefly, the idea depends on the question of how

[3] This view that takes language as a form of action, as something used as a tool to change some things in the world rather than to attain the *truth* or *reality* has been defended more recently by the embodied trends in cognitive science (Baggs, 2015). Relevantly, Vygotsky is another prominent and historical figure that is being rediscovered lately (Vygotsky, 2012). In similar vein, see also Christopher Gauker (1994).

[4] Even though it will not be at the centre stage argumentatively, a Wittgensteinian kind of contextualism will be lurking behind throughout the present thesis. I do not commit some particular reading of Wittgenstein and broadly construe and borrow the idea of *games* or *language games* as a way of thinking.

come our best scientific theories are successful if they are not true or approximately true. And it continues, only the realist stance can explain it. However, for the realist to back this argument up they need to show how our best theories can and do approximate to truth. For this, *selective realism*s might have an answer.

2.2 Selective realisms: explanationist realism, entity realism and structural realism

Selective realism is the collection of proposals which maintain that theories change throughout the history of science but that selective parts of theories are retained thanks to which a particular kind of accumulation is achieved. In turn, the accumulation here might be thought of as approximation to truth. Another important factor here is realists' desire to merge scientific realism with the widely accepted idea that most theories are false even if not all. The aforementioned accumulation might provide the answer for which parts of theories are true and which parts are not.

There are different kinds of selective realism depending on what they take as being retained. For the present discussion, though, I will broadly tackle what their basic claims are, how they mainly diverge and what their common problem is.

As stated in the subtitle, there are mainly three selective realisms: *explanationist realism*, *entity realism* and *structural realism*. Explanationist realism holds that what are retained in theory changes are the essential and 'necessary' parts of the previous theory in explaining/predicting the relevant phenomena. In this vein, Kitcher proposes that there are "idle parts" or "presuppositional parts" of theories along with the "working parts" which do the 'real' work (Kitcher, 1993).

On the other hand, entity realists argue mainly that, through experimentation using different techniques and methods, it is possible to see if we can manipulate the relevant unobservable entity to create certain and predicted kinds of effects, and to the extent that we can do this, it is safe and reasonable to believe in the mind-independent existence of that entity (Hacking 1983, 1983; Miller, 2016). Notice that not every entity realist commits to the causal-historical theory of reference, it is relatively more compatible with it than other approaches.

Lastly, structural realism, as historically known as relationism, is the view that even though unobservable entities proposed in scientific theories cannot correspond to the mind-independent reality, the structure between them can do so. And just like other selective realisms, we can see how some ‘structure’ is retained through the theory changes (Chakravartty, 2007; Frigg & Votsis, 2011; Psillos, 2006).

Every selective realism faces a particular kind of challenge which are voiced and levered against each other by the proponents of each realism. For the present purposes, though, the important thing is what is common to all selective realisms: they all infer approximation to truth, or truthlikeness if you like, from some form of continuity in the history of science. Whether this inference is justified, or whether it is problematic in terms of approximation and truthlikeness, will be discussed below. However, they are not the only ones making inferences from history. Anti-realists are also famous for their historical inference in the exact opposite direction: pessimistic meta-induction or shortly, meta-induction.

2.3 Pessimistic meta-induction

Pessimistic meta-induction or pessimistic induction (sometimes called *disastrous* induction (Psillos, 2022) is famously proposed by Larry Laudan in his 1981 paper[5] (Laudan, 1981). Laudan argues that history of science is a graveyard of theories,

[5] For a similar argument called “Principle of No Privilege,” see: Hesse (1976)

what is more, it is also a graveyard of successful theories; even worse, a graveyard of successful theories that are thought to be the best theories of their own time as well. Laudan, generalizing from this observation, claims that just as our past successful theories turned out to be false, we do not have any compelling reason to think that our successful theories now will not turn out to be false in the light of new theories in future. Thus, Laudan does not share the 'optimism' of realism and argues against No Miracle Argument.

The argument has received different criticisms and led to much controversy. One such criticism takes pessimistic argument as a fallacious reasoning in that it commits *base-rate fallacy*. The idea stems from how induction is normally calculated and claims that without knowing the distribution of successful theories and unsuccessful theories among the totality of theories — which is impossible for obvious reasons — it is mathematically impossible to do such an induction (Lewis, 2001; Mizrahi, 2013, 2016). In fact, this argument is also valid in the case of No Miracle Argument when understood as an induction from historical record in favor of the claim that successful theories are true or approximately true (Howson, 2000). Thus, *base-rate fallacy* objection would better be seen as an argument against historical inductions where the base-rate is unknown.

Independently of the base-rate fallacy objection, Chang (2003) criticises the scientific realist's alleged inference from the continuity in theory changes to approximate truth or truthlikeness as there is no logical link between the two. Being aware of these objections, Psillos (2009) takes the realist stance as constituted by two distinct steps; first one concerns itself with showing there is a continuity in theory changes and since this is not enough to establish approximation to truth or truthlikeness, the second one comes to rescue in the form inference to the best explanation (IBE). More clearly, the second step is the argument that the continuity at hand, which is shown by the first argument, cannot be explained without assuming approximation to truth or truthlikeness. However, whether IBE at this philosophical context is a legitimate argumentative move or not is a further controversial issue for various reasons. For instance, one might reject that the first step of the argument has been successful and established the continuity in theory change explanation-wise, structure-wise or entity-wise. Or someone, like Van Fraassen (1989), could argue that theory changes might be susceptible to some form of Darwinian selection where the selection dynamics e.g. some factors that are important to scientific community, lead to such continuity, thus removing even the need to develop a further explanation. Or in parallel with socio-historical approaches to science that I will mention below it is possible to argue that alleged continuity in theory changes is actually an empirical phenomenon that can and should be explained socio-historical processes in science. One can also claim that scientific realism's assumption of aproximate truth or truthlikeness is not "explanatory" either and even if we did an inference to the 'best' explanation it would not be in favor of scientific realism.

On the other hand, when the argument construed in two steps and the link between the continuity in theory changes and approximation to truth or truthlikeness becomes allegedly unnecessary, another logical possibility might appear. If it is not the continuity in theory changes that gives us the realism, then maybe it might be the discontinuity in theory changes. This is the view defended by Devitt (2007), Doppelt (2007) and Fahrbach (2011). We can summarize it in three branches: 1) Our current science is in some sense incomparably better than the past and our best theories are confirmed to a degree that we can say there is a qualitative difference. 2) Last century has witnessed an exponential growth both in the numbers of papers published and in the numbers of scientists. Thus, for the current successful theories in science, it is safe to say that they are much more confirmed by evidence. 3) It would be difficult for the anti-realist to find a theory which is accepted in the last century and also abandoned later. For Fahrbach (2011), this shows that our current best theories are qualitatively different than the previous ones as it would take a shorter time to discover they were false and abandon them.

However, the first point can be criticized on the grounds that the past theories will always look worse through the glasses of current theories. For the second point, though, Wray (2013) states the argument from exponential growth would be true at any point — or most points, or at least at some past points — in time, which undermines the argument itself, showing that exponential growth cannot show any qualitative change for the benefit of current theories. As for the last point, it can be argued that such an argument can only show that science has matured in some ways, but nothing more. Moreover, when considering how ever more easy it is for anyone to publish something compared to a century or more ago, it becomes even less suggestive of approximate truth or truthlikeness and becomes an empirical subject of sociohistorical factors. Alai (2017) also asks the question that if there were a qualitative difference between the theories of current science and that of past, then how could we explain that past theories had been successful even to some degree? Consequently, the argument from discontinuity requires that the proponents should show how there is a qualitative and conceptual difference between the failed theories and successful ones of our current science. The burden of proof is on their shoulders and it seems there is not any compelling case yet.

Let us get back at the pessimistic meta-induction. The problem that got us to this sidetrack was mainly the criticism against inductive inferences from history of science whether realist or anti-realist. However, might it also be possible, or even more accurate to construe pessimistic induction not as an inductive argument but as a deductive one? Wray (2015) does exactly this and claims pessimistic induction is actually not an induction but a deduction and, in fact, just one single successful but false theory either in history of current would be enough to reject the scientific realism's claim. On this view, the historical part plays only a supportive role in showing that there is of course one, and possibly more than one, successful but false theory which challenges the claim that there is a link between success and approximate truth or truthlikeness.

2.4 Underdetermination and Stanford's New Induction

More recently Kyle Stanford (2006) proposed a new inductive argument from history of science which is aptly called by him as "*New Induction.*" Since New Induction makes reference to good-old issue of underdetermination, we should tackle it first.

The problem of underdetermination can be traced back to 19th century physicist Pierre Duhem. Duhem observed that there is no and cannot be a 'critical experiment' that could either confirm or refute a given theory once and for all. This is because all hypotheses and all experiments contain auxiliary assumptions and background beliefs that might be confounders in the result of experimentation, which means that it is never possible to test a single hypothesis or a theory in isolation. Even if one would try to isolate further possible confounders and test them, those further tests would also rely on other auxiliary assumptions and background beliefs resulting in an endless chase. It has become clear that experiment plus logic is never conclusive and cannot determine our choice of a theory over another. The term ‘underdetermination’ refers to this situation where theories are *underdetermined* by the evidence. This poses a challenge for realism because it shows that there might not be a rational or logical and in this sense ‘universal’ way of theory choice which opens up a space for possible factors of an arbitrary, or at least of a mind-dependent nature. Even though Duhem initially proposed this idea only for theoretical physics, scholars realized it is applicable to pretty much every field of science.

Later, in mid 20th century Quine (1951, 1975a, 1975b) famously extended this to the totality of our knowledge with his concept of *web of beliefs*. Accordingly, not just our scientific hypotheses but all of our beliefs, all of our knowledge constitutes an interconnected web. The following quote from Quine himself maybe helpful (Quine, 1951):

> The totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only along the edges (p. 42)

Thus, although we can somehow intuit there is something false with this web of belief, we cannot know what exactly it is since ‘it only impinges on experience along the edges’, both in the context of scientific epistemology or epistemology in general. Of course one can argue against this by pointing out that in both science and other contexts of knowledge we can revise and change our web of beliefs according to some criteria, some rule, etc. This is what Laudan basically claims (Laudan, 1977, 1984). However, as a counterpoint this would miss the main claim of Quine which is that even those criteria or rules are already within the web of beliefs. Moreover, those criteria or rules might not be able to determine the exact nature of revision in that alternative revisions might be ‘reasonable’ at the same time.

This is a radical stance. The challenge for realism here is mainly the idea that this web of belief is subjective, it varies from person to person, from culture to culture or from historical period to another historical period and it seems that there is no way we can construct or discover a procedure, algorithm, criterion or a meta-constraint of another kind that will be applicable to all possible web of beliefs. If we did find such a meta-constraint, it would help us bridge the gap between the web of beliefs and a mind-independent reality and select certain types of web of beliefs over others. This radical scepticism has been taken up by socio-historical approcahes in philosophy of mind that we will mention below. Proponents of these approaches claim that the web of beliefs are socio-historically contingent whereas Quine would take up another route and state that nature of human psychology is the definitive factor in this context.

However, this radical holism is not the only way to interprete the problem of underdetermination and much of the discussion has been revolved around a more modest reading. In fact, Stanford (2023) distinguishes between the two readings as *holist underdetermination* and *contrastive underdetermination*. Holist underdetermination is when we cannot test a hypothesis in isolation and determine if we received the result due to our hypothesis or some other confounding variable such as auxiliary assumptions or background beliefs. On the other hand, contrastive underdetermination is the situation where a given set of data or evidence can be accounted for more than one and possibly conflicting alternatives. Even though these two are obviously related to an important degree in that they differ in their emphases. The former emphasizes the challenge to single out a factor when faced with a data, or some evidence whereas the latter emphasizes the difficulty to choose alternative theories or 'web of beliefs' given some data or evidence. One implication of the distinction is that even if we did single out a relevant factor in determining whether it led to a given result (data/evidence) or not, still there would be possibly more than one hypothesis to lead that given result. Notice that ‘contrastive underdetermination’ is closely related with the notion of ‘empirically equivalent alternatives’ that has been central to recent debates in philosophy of science.

The issue of empirically equivalent alternatives or rivals is, as the name suggests, there might always be some other alternative which is empirically equivalent to the theory we believe in, undermining its truth yet being empirically equivalent. However, critics contend that this issue is mostly proposed as a possibility in principle. In actuality, on the other hand, it is either the case that there are no alternatives available at a given time, or the proposed alternatives are not strong enough to be challenging.

Stanford (2006) in this context claims that the underdetermination must not only be about empirical adequacy, it would still be problematic if there were empirically unequivalent but equally confirmed alternatives. Notice that the difference is all about *unobserved observables*. Two theories might have different predictions regarding unobserved observables yet they can both be equally confirmed by the evidence at hand. Stanford claims that we almost always witness such a weak form of underdetermination both currently and throughout the history of science. What is more, on Stanford’s view, this is a recurrent situation in that, say, we chose an alternative theory over the one at hand, then this weak underdetermination –another alternative- would then occur for the newly chosen theory as well. He further claims that there might also be presently unconceived such alternatives (empirically unequivalent but equally confirmed) which can only later occur to researchers. And his New Induction makes the case that The Problem of Unconceived Alternatives has always been there throughout the history of science from Aristotle to String Theory. Scholars could not literally conceive equally confirmed alternatives at the time and when they conceived the new alternative replaced the past theory. Hence, he thinks, we are most probably in the same situation while dealing with our current theories.

2.5 Socio-historical approaches

There are both sociological and historical turns in the philosophy of science. Since they are related in and similar in some of their views, I will present them in a single section.

The so-called historical turn in philosophy of science is taken to start with Kuhn (1962/1970) In his monumental work, *The Structure of Scientific Revolutions*, Kuhn claims that when we take history of science seriously and look closer to investigate how science works, we see that there are actually two different phases: normal science and the crisis. During normal science there is nothing peculiar and scientists work on local problems without the need to consider theoretical or philosophical questions behind them. However, during a crisis phase, most of the time triggered by some contradictory evidence piling up, scientists grow more concerned about the theoretical and philosophical aspects which would later result in a paradigm shift where a new philosophical and theoretical framework, i.e. paradigm, replace the older one. This cyclic theme in history of science is actually the norm that we encounter all the time, Kuhn claims. A paradigm brings with its own ontology, own methodology, own techniques, own phenomena, own explanations and even own conceptualizations and perceptions. For Kuhn, not just phenomena are theory-laden, but everything is everything else laden which could remind us Quinian holistic underdetermination. In this regard, one of Kuhn’s central claim was incommensurability between the paradigms in the sense that two scientists working in different paradigms are said to be living in different worlds, they cannot even communicate properly since they mean different things even when they use the same terms.

The issue of incommensurability between paradigms has been a controversial topic and there are different stances. After Kuhn, contributions of two imporant figures became prominent in keeping the valuable insights of Kuhn's work and providing a considerably less radical view. For example, Lakatos proposed the notion of *research programmes* instead of *paradigms* which contain a *hard core* and *auxiliary hypotheses*. Roughly the idea is that the hard core is the essential part of a programme which cannot be abandoned without abandoning the entire program as a whole, whereas the *auxiliary hypotheses* is susceptible to change. Newton's three laws of motion is an example to hard core in this context. The three laws of motion and the law of universal gravitation do not, by themselves, specify what one will observe. To generate testable predictions, these laws must be supplemented with a range of auxiliary assumptions about the positions, masses, and relative velocities of celestial bodies, including that of the Earth. On the other hand, Laudan proposed *research traditions* where the components are only loosely related in comparison to both Kuhn and Lakatos in that elements might enter or leave the hard core from time to time, or it is okay that a later update on theory can only account for less phenomena than before for various reasons which is sometimes called *Kuhnian loss*. Laudan also maintained that there might be more than one research traditions in a given field at a given time, as opposed to Kuhn's paradigms. Moreover, these traditions might be in competition and can become dominant in the field in different periods. Notice that Laudan's picture still admits that research traditions have their own ‘way of doing things’ in terms of metaphysical understanding, methodology or some other institutional practice etc. even though they are loosely related here. Hence the problem concerning mind-independent reality arises here again since it is an open question how to attain some mind-independent truth within a particular research tradition or a set of research traditions that often conflict one another at a given moment. Laudan only refers to the criteria that a scientifici community holds at a given time, but that criteria themselves would be under a given paradigm or a given web of beliefs, to speak in a Kuhnian or Quinian sense. This is pretty much the same issue that we mentioned above in the section on Quine’s holistic underdetermination.

Notice that the trend we are talking about here takes the ‘conventions in science’ seriously. In parallel with that, the questions of what are the conventions in science and how does scientific knowledge come about conventionally gained prominence. Thus sociological approach emerged. Sociological approach focuses on how social settings, political factors, the different practical conventions in different instititutions or competition between individuals, institutions or entire schools of thought affect or even determine scientific knowledge. [6] Even though this can be compatible with realist commitments, the real controversy arises when one considers the Strong Programme in sociology of science (Barnes, Bloor & Henry, 1996; Bloor, 1976). Proponents of the Strong Programme have two main commitments: 1) They should be totally neutral as to the success, truth or falsity, or epistemic status of the theories, findings, interpretations, claims, beliefs, practices etc. to avoid the possible biases that might stem from intra-theoretic or intra-conventional leanings. That is to say, to proponents of the Strong Programme, one should investigate what is taken to be true

[6] As we mentioned above, one might think this can also be formulated as ‘how the web of beliefs are constructed via social factors and conventions’ as opposed to Quine’s appeal to nature of human psychology as the ground.

and false all alike. 2) Social factors not only affect the scientific processes including theory choice over another, it determines them.

The second commitment can be seen as related to the first one because when one explains how both so-called true and false beliefs are produced by the same social principles and same social dynamics the truth becomes irrelevant in a mind-independent sense. This stance also undermines the controversial distinction between the scientific knowledge or knowledge in general, including the everyday knowledge or knowledge-related practices of different cultures and social settings, in line with contextualist views that we tackle in this thesis a couple of times.

In its claims Strong Programme makes use of case studies to an important extent. This view is often criticized on the grounds that case studies cannot support general conclusions about science, or that the case studies employed as examples are frequently misrepresented. However, even though the criticisms are important methodologically, the claim that some or all case studies conducted by sociologists of science turn out to be false cannot give an a priori reason that social factors have not much of an importance. This is especially so considering most scholars even in the realist side of the debate acknowledge that conventions are important to a varying extent. Therefore, to think that the conventions taking place in social situations might be affected or even determined by social and/or political factors seems reasonable. It is enough to show it only once to raise suspicion. Hence, the burden of proof is on the shoulders of the realist or anyone claiming that knowledge-related dynamics are not affected by such factors. Provided there are too many decisions and practices or even theory changes throughout the history of science this might be practically impossible, which would again be a challenge for the realist side of the debate.

2.6 Perspectival realism

Admitting the problems and the controversy surrounding the topics in philosophy of science and the realist claims, that we also tried to cover to some extent, Massimi (2018, 2022) offers a minimal account of realism bypassing most of the issues about theory change, continuity in philosophy of science, etc. She claims that despite the alleged problems of the realist stance we should not drop the concept of 'truth' and 'reality' and try to preserve them in a new framework called '*perspectival realism*.' She grounds truth on 'perspectival cross-evaluation.' On Massimi's view, of course there are different perspectives, different cultures, different conventions in terms of 'knowledge' and yes; they often conflict with each other. However, she maintains, they do not simply conflict with each other but, at the same time, they are also compatible with one another and can establish some agreement. This part is actually what we can take to be truth albeit in a perspectival fashion. For her, confrontation between different perspectives may lead to the discoveries of such truth.

Admittedly this position has an advantage over the previous ones in that it takes as its central point the fact that there are many and often conflicting perspectives, belief systems, etc. in the world. However, it has several problems in tandem with previous ones as well. First of all, perspectival realism cannot give us a mind-independent truth or reality, or an approximation thereof. It can only show that we can establish consensus to some extent between perspectives, which might also be cumulative in nature.[7] Massimi herself admits that and she further states perspectival realism rejects the very conception of *objectivist*, mind-independent truth. I take this to be a problematic move as the chosen conception of truth seems to

[7] Notice that the arguments in favor of scientific realism from the continuity in history of science are also of this nature. They can only show that there is a consensus throughout history that is also cumulative to some extent. We mentioned this above with reference to Psillos (2009).

be chosen *ad hoc*. It is certain we are and always have been able to define or conceptualize any concept as we wish. However, the controversy between scientific realism and anti-realism seems to stem from the historical consensus in the field as to the relevance of mind-independence to realism. Without doing justice to this observation, it seems to me an arbitrary move to make the problem appear resolved and reduce it to a mere semantic convention. On the other hand, if the perspectival realist chooses to take this route, they must at the very least explain why we should not simply abandon the traditional concept of (mind-independent) truth altogether and replace it with an alternative. This would be what philosophical integrity and linguistic hygiene would require. Moreover, truth and objective reality have always been put forward as a meta-perspectival or non-perspectival authority to appeal when faced with a conflict. If it’s not going to serve this function, why insist on keeping it?

One last problem with perspectival realism occurs in cross-perspectival evaluation. For cross-perspectival evaluation to work, a given statement should be possible to assess from both perspectives and it should satisfy the *performance-adequacy standards*[8] of them. For this, an account of translation between the perspectives must be established. Because, even though translation and cross-perspectival evaluation seem possible between *near* perspectives or narrow contexts, it might be nearly impossible, or even outright impossible with perspectives that are *far from each other*. What I mean by these two expressions, the distance between perspectives, might be best exemplified with different examples. For example, the Atkinson-Shiffrin model of memory (Atkinson & Shiffrin, 1968) and distinctiveness theory of memory (Glenberg & Swanson, 1986; Glenberg, 1997; Nairne, 2006; Schmidt, 1991) can be said to be *near* perspectives whereas ecological theory of

[8] More information on the working details of perspectival realism see, Massimi (2018, 2022).

visual perception (Gibson, 1979) and accounts from deep learning approach to neuroscience (Richards et al. 2019) might be very *far* from each other. On the other hand, say, Aristotelian conception of soul and current cognitive science might constitute very distinct and distant perspectives. To take it a bit more extreme, the distance between the perspectives of a non-industrialized society, say, an Amazonian people, and the perspective of a particle physics would be to an extreme degree. Therefore, I argue that translation and thus cross-perspectival evaluation is a problematic step in the argument that needs to be cashed out.

2.7 Constructive empiricism

Before going into the concluding section of this chapter, I want to explain why I do not adopt a constructive empiricist approach. This is important since constructive empiricism might be the most renowned stance in the anti-realist camp (Van Fraassen, 1980, 2001, 2002).

Constructive empiricism is proposed by the philosopher Van Fraassen. His main motivation in the framework is to propose an empiricist view of science which does not fall into the traps of logical positivism. That is, he tries to give an account with minimal metaphysical commitments while also giving most possible credit to scientific knowledge through grounding it in *experience* as typical of empiricism. Unobservable/observable distinction plays a key role here. Because for Van Fraassen, science does not aim to give a true picture of mind-independent reality but *it aims to give empirically adequate theories, it aims to save the phenomena*. And on his view, theories are empirically adequate to the extent that their statements about observable phenomena are true. Then the question arises: what is observable and what is unobservable?

Van Fraassen's observability criterion is a central aspect of his constructive empiricism, and it is defined with respect to the limits of human perceptual capabilities. According to Van Fraassen, an entity is observable if it is, in principle, accessible to the human perception without the aid of instruments. For instance, distant stars or planets qualify as observable entities, not because we currently perceive them directly, but because if they were close enough, we could. Their observability is tied not to current accessibility but to the modal condition of potential unaided observation. In contrast, microscopic entities such as bacteria do not satisfy this condition. We can see them only through technological instruments like microscopes, whose reliability and function rely on further scientific assumptions. Thus, the observability criterion draws a line between what can be said to fall within our natural reach and what depends on theoretical constructions and instruments whose epistemic status is not theory-neutral. This distinction has been one of the most debated aspects of constructive empiricism, raising multiple objections, particularly those of relativity and circularity. Before addressing these concerns, it is worth emphasizing an important clarification Van Fraassen himself offers. He acknowledges that there is no sharp a priori boundary between observable and unobservable phenomena. Instead, what counts as observable ultimately depends on the actual capacities of epistemic agents and the context in which observation is situated. In other words, the line between observable and unobservable is drawn not from some metaphysical or purely conceptual demarcation, but from a contingent and empirical consideration of what some particular community of observers are able to perceive, given their biological, technological, and social circumstances,. This point becomes clearer when Van Fraassen introduces the famous thought experiment involving an epistemic community whose members are born with electron microscopes biologically integrated into their sensory systems (Van Fraassen, 1985). In such a world, microscopic entities like molecules or bacteria would be observable for them, just as trees and rocks are for us. Thus, the observability criterion is not fixed across all contexts or agents. It is, in a sense, relative to the epistemic community. However, this dependence brings us to the first major objection, the problem of relativity.

The relativity critique suggests that if observability is relative to the epistemic community or the contingent capabilities of observers, then it risks becoming theory-dependent. That is, any scientific theory could stipulate its own observability criterion compatible with its commitments, and thereby justify or exclude entities accordingly. In such a scenario, the observability criterion no longer functions as an independent constraint but collapses into the very theoretical framework it was meant to regulate. This undermines one of the main attractions of constructive empiricism, namely its attempt to avoid metaphysical commitment while retaining epistemic humility. To address this issue, Van Fraassen asserts that the observability criterion is theory-independent. He claims that it is not the theory that determines what is observable, but rather our biological capacities and perceptual modalities. Yet this response invites a second, closely related objection, circularity. If our knowledge about what is observable is itself derived from scientific theories such as our understanding of how microscopes work or how visual perception operates then using that very knowledge to define a supposedly theory-independent criterion seems circular. That is, we use theoretical commitments to draw the line between observable and unobservable, only to then justify those same commitments based on what we define as observable.

This circularity becomes particularly pressing when we consider the role of instruments. For example, microscopes are only trustworthy epistemic tools if we accept the theories that explain how they work and validate the reliability of their outputs. But those theories are, in turn, justified by scientific practices that rely on observable outcomes. In rejecting small-scale entities as observable because they require theory-laden instruments, Van Fraassen appears to be engaged in a circular maneuver: appealing to theory to limit what theory can be committed to. As such, the claim of theory-independence becomes increasingly difficult to sustain. One way to interpret Van Fraassen's insistence on theory-independence is to suggest that he sees the observability criterion as belonging to a meta-scientific or philosophical domain, something like a conceptual framework external to scientific theory itself. On this reading, the criterion functions as a philosophical constraint on scientific practice, and thus its independence is secured by its non-scientific nature. However, Van Fraassen explicitly rejects this interpretation. He states clearly that he does not propose the observability criterion as a philosophical or conceptual tool external to science (Van Fraassen, 1980, Ch. 3, Sec. 7). He insists that his empiricism is grounded within scientific practice, not above it. Therefore, this first interpretation fails to capture his intent. The alternative interpretation is that the observability criterion is scientific in nature and that science itself determines what counts as observable. Yet this reintroduces the circularity problem again. If the scientific framework is used to define what is observable, then observability can no longer constrain science from the outside, it becomes one of its internal assumptions. Moreover, if Van Fraassen maintains that observability is theory-independent and that theory itself determines it, he seems to be offering two incompatible claims.

Either observability is external and constraining, or it is internal and relative. Holding both simultaneously appears contradictory.

This tension has been acknowledged by Van Fraassen himself. In his later exchange with Morton (Morton & Van Fraassen, 2003), he concedes that the issue of circularity cannot be dismissed so easily. He admits that the observability criterion's claim to theory-independence does not, in fact, resolve the charge of circularity. As such, he concludes that some level of circularity may simply be unavoidable if one adopts the constructive empiricist view. That is, if we accept that science should be committed only to what is observable, and yet our knowledge of what is observable is itself shaped by science, then we must accept this feedback loop as a structural feature of our condition.

In light of these considerations, I argue that the observable/unobservable distinction, so fundamental to constructive empiricism, cannot be maintained in a principled way. Therefore, I reject the distinction and, by extension, the foundational role it plays in constructive empiricism. [9]

In line with concerns already noted in the literature, I want to raise an additional worry about the observable–unobservable distinction. As discussed earlier, this distinction appears arbitrary, particularly in light of the relativity and circularity problems. What further strengthens this concern is that, especially in relation to the relativity problem, the distinction between what counts as observable and unobservable is troubled by a continuity between scientific and folk ontologies or, more broadly, between scientific and natural languages. It seems difficult, if not impossible, to draw a clear line and say, "this part is scientific," and "beyond this point is just folk ontology." Philosophers like Quine, Sellars, the Churchlands, and

[9] Note that I do not think this is particularly *that* problematic since any philosophical account broad enough to cover the entirety of science or anything on a par with that would probably fall into some sort of circularity at the end. I view relativity problem as a bigger concern here.

Rorty have pointed out this continuity. Natural languages themselves fall under the category of theoretical languages, and the folk ontologies they express are not theory-independent, they are themselves theoretical constructs. Therefore, the anti-realist arguments we have discussed in relation to scientific realism also apply to these domains. This undermines any attempt to say that observations made through unaided perception are 'real' while those involving theoretical instruments or concepts are 'unreal.' Observations via unaided perception are still mediated by natural language and folk ontology. In this light, it seems difficult to sustain a non-ad hoc position that is anti-realist about science but realist in a broader metaphysical sense.

Constructive empiricism, in fact, is much broader than the part we covered here. I just wanted to present the part essential for our debate about realism and my criticism thereof. For a thorough understanding of the position one can see the original resources cited.

2.8 Conclusion

In this chapter I gave an overview of the debate between scientific realism and anti-realism along with my comments and stances with it. In doing that, I aimed breadth over depth since my motivation was to show even though there has been a heated debate and various arguments, most of them collapse into some particular version which ultimately either does not give us an account of mind-independence or trivializes the concept of *realism* and *truth* by outright rejecting mind-independence as a prerequisite for realism.

We mentioned how Putnam-Kripke model of referentiality is problematic from a Wittgensteinian tradition in philosophy of language, noting that the concept of language games and Wittgensteinian contextualism will be implicit for our further discussions. Then we saw how selective realisms all collapse into some particular form of consensus, as they aim to show a historical continuity that might very well be a result of certain type of conventions, and this is itself not enough to establish mind-independence or approximate truth. Likewise, we dealt with perspectival realism and found it problematic on three grounds: as (1) trivializing the concept of truth as an ad hoc move, (2) being only capable of establishing a particular kind of cross-perspectival consensus[10] and (3) having an important gap in the argument about translatability of statements especially between perspectives that are *distanced* from each other.

On the other hand we found discontinuity thesis about current science and its success problematic for two reasons: (1) that can only show science has matured in some ways and (2) there does not seem to be a compelling case in favor of a conceptual difference between the past science and the current one. Likewise, on the realist camp, constructive empiricism suffers from a contradiction in its observable/unobservable distinction.

Quinian holism, Stanford's New Induction and Problem of Unconceived Alternatives and the sociohistorical approaches continue to challenge realist stance. In parallel with this, I take science as a particular kind of cultural/social practice[11], an emergent institution developed throughout centuries and embedded into other practices of our modern world. Even though this makes it unable to legitimately claim mind-independent truth or an approximation to it, contrary to first common

[10] This might be called *synchronic consensus* whereas continuity theses like selective realisms that try to show how science keeps some specific parts of theories in theory changes throughout history are *diachronic consensus*.
[11] For some formulations of sociohistorical /cultural practice approach to science, see: (Bourdieu, 2004; Collins, 1981; Collins & Pinch, 1993; Knorr-Cetina, 1981; Latour & Woolgar, 1979; Pickering, 1984) I did not include them in this chapter since their crucial commitments for the realism-antirealism debate are pretty similar to sociohistorical approaches.

reaction of an individual facing this claim, it is not the case that such a conception of science renders it equal to other practices like religion, spiritualism, etc. Every practice is a particular kind of *game*, with its own implicit and explicit rules, with its own functions, own results and implications. Therefore, science can still be science without claiming the hold of mind-independent truth in any non-trivial sense. Because we know and see what we can do with science in medicine, in technology, in education, etc... How science is serving its alleged function without realism is a big question for history, sociology and philosophy of science altogether. An important stance in this regard is that science gives us explanations of *our* worlds and a particular kind of explanation at that.

CHAPTER 3

WHEN MECHANISTIC EXPLANATIONS WENT UNREAL

During the last few decades, the New Mechanism tradition in philosophy of science has gained prominence. This tradition departed from the mainstream strands in 20th century philosophy of science, eschewing the discussions about nomological explanations, theories that are composed of lawlike statements and inter- or intra-theoretic reduction. The New Mechanists took *mechanisms* or *mechanistic explanations* at the centre stage and built a new way to look at how science explains.

One might interprete the New Mechanism framework in two ways, broad or narrow. In its broad interpretation, the mechanistic view of science is an overarching framework, both descriptive and normative of how scientific explanation works and how it should work, covering science in its entirety. The narrow interpretation is a modest one in that it takes the mechanistic framework only as a particular way of doing science among other alternatives (Glennan, 2017). I am committed to the broad interpretations for the reasons that will be clear at the end of this chapter. However, the claims of the present thesis are surely compatible with the weak version as well. Because it is both compatible to read the thesis as a bold normative claim relevant to all scientists or as only a contribution to mechanistic way of doing science for those who take the narrow interpretation of mechanistic outlook.

In this chapter, I will explain the basic claims and commitments of the New Mechanism. I will first show what mechanisms are or what mechanistic explanations are taken to be. Then I will explain how mechanistic explanations come about, how we arrive at them. Later, I will focus on how mechanistic *levels* can be illuminating in terms of the relations between scientific disciplines and emergence. Lastly, I will clarify the main misconceptions about mechanistic explanations which are mostly about the compatibility of abstraction/idealization, linearity and the distinction between mechanisms and machines. At the end of the chapter, I will adress how mechanistic explanations are compatible with an anti-realist view on science.

3.1 Introducing mechanisms and mechanistic explanations

Machamer, Darden and Craver's 2000 article has been very influential at the resurgence of New Mechanism. Even though they do not give a formal definition of mechanisms, they characterize it as "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." However, this has later been updated by other mechanists mostly on the grounds that it is too restrictive and there is no a priori reason for a mechanism to be *regular* or to have set-up and termination conditions[12] (Bechtel, 2011, 2022; Glennan, 2009; Krickel 2014; Leuridan, 2010; Machamer, 2004). For example, Bechtel and Abrahamsen (2005) argue that "a mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena." However, this characterization has later been found to be problematic as the concept of *function* is a very loaded term metaphysically and there is no consensus as to how to situate it in the mechanistic

[12] There are different conceptualizations of *regularity.* Following (Craver & Tabery, 2023) I take it that a mechanism is regular when it works the same way given the same background conditions. This also might relate to a counterfactual understanding of causality, though not necessarily. However, it does not follow from this that when we cannot provide the same conditions more than once for the contingent reasons, say some event in history occuring only once, we cannot attribute a mechanism to it. It is only a contingent complication of that specific event that it does not happen more than once. This contingency does not pose any problem to the claim that the event is a result of some mechanism or that there might be some mechanisms accounting for that event.

science[13] (Craver, 2013; Garson, 2013). Thus I will take Glennan's *minimal mechanism* (Glennan, 2017, p.92) as the most recent and widely accepted characterization of mechanisms: *"A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon."* So there is a phenomenon and a mechanism for that. Before going into the details, it will be better to look into that distinction.

At this point, Glennan (2017) makes use of Giere's account on models (Giere, 2004, 2006) and mostly follows him**.** Giere proposes a liberal conception of model where there is not a definitive characteristic or property of being a model except that something is used by the modeler as a model to represent something else with a particular purpose. This conception stems from the practice that both in philosophy and in the sciences there are countless ways something is taken to be a model. Equations, algorithms, simulations, diagrams or even living creatures as in model organisms or model animals.[14] Note that the representational relation stated here crucially relies on the modeler's agency and purpose. [15] Accordingly, there are only models of models of models ad infinitum... Glennan goes on and states that mechanistic explanations are models as well, sometimes called *mechanistic models* or *mechanical models*. Moreover, they do have two parts: phenomenal *description*

[13] A caveat about functions: I am sympathetic towards Craver's perspectival account (Craver, 2013) and I take functions to be heuristic attributions to systems/mechanisms. The main reason is that even though functional thinking may appear to enrich our theorization and modelling processes, at least only by itself, it has no contribution to whether some model is explanatory or an explanation is successful or not. Because description of a mere causal chain without any attribution of function would not lose explanatory power. Consider explaining some functional phenomena or translating a functional description into description of a set of particular events without any normativity (as opposed to function-talk since functions always come with malfunction and normativity) that where we have enough knowledge. It would be still possible albeit practically very difficult or impossible. Moreover, one might even say functions themselves, when put forward as explanatory, become further explanandums in mechanistic science. This way of thinking can be traced back to the well-known distinction between *the context of discovery* and *context of justification* which is conventionally attributed to Reichenbach (1938). I will not focus on the debate about functions in this thesis since it is not relevant to the main claims of the present thesis.

[14] Natural language sentences, words, idioms or metaphors etc. might also be taken as models in this picture.

[15] Notice that *the modeler* here might be more than one person, it can even be an entire community.

and mechanism *description* (Glennan 2005, 2017). Phenomenal description only gives us the cause and effect pairing, namely the explanandum in a sense.[16] The mechanism description, though, gives us how that cause and effect come about as a result of particular kind of components, their relation to one another, their activities and the overall organization of them. It is important to realize, though, that these two descriptions cannot be separated much clearly in most mechanistic models and can only serve as an idealized picture here. In this context, mechanism descriptions can come in different ways. For instance, some phenomenon might be the result of a temporal process whereas some others might be synchronically realized or constituted by the mentioned mechanism (Craver and Darden, 2013). In this sense, mechanisms are said to be producing, underlying or maintaining the phenomenon. Producing refers to situations where the mechanism produces the result as an end-point of a temporal sequence, e.g. protein synthesis. On the other hand, a mechanism can be underlying or maintaining the phenomenon as is the case with organelles and their activities that make up a cell or a metabolismic mechanism to maintain a certain regulatory processes in an organism's body. Another parallel distinction is between *etiological* explanations and *constitutive* explanations. Etiological explanation pretty much corresponds to when the phenomenon is produced whereas constitutive explanation is when the mechanism constitutes the phenomenon. Processes such as protein synthesis or a neuron's reception of current through its dendrites, its loading and firing, the transmission of the signal across the synapse, the subsequent release of neurotransmitters at the terminal point, and the resulting activation of another neuron via a different dendrite can serve as examples of etiological mechanistic explanations. In contrast, the constitutive mechanism can be illustrated through the

[16] I use the word explanandum with a grain of salt here just to illustrate that phenomenal description gives us the thing or process to be explained. It should not be confused with the original meaning of the word that comes from covering-law explanations.

earlier example of cells and organelles. This last point has led to much debate about how to demarcate constitution and causality in mechanistic explanations. Since this debate has no definitive relevance to the present discussion I will not much focus on it except for the few sentences below in the passages about levels. A quote from Glennan (2017) can summarize the upshot of the present discussion:

> Mechanistic models are the vehicles for mechanistic explanation. A mechanistic model characterizes both the phenomenon to be explained and how the organized activities and interactions of some set of entities produce or underlie that phenomenon. So mechanistic models show how the phenomenon is caused and constituted by a mechanism. It is for this reason that I call mechanistic explanations *how* explanations.[17] (p. 68)

I said above that it may not be too easy to distinguish between the phenomenal description and a mechanism description of a mechanistic model. Another related issue is that mechanistic models are always incomplete to differing degrees.[18] For example, there might be some models where a subset of components or activities corresponds to mechanistic parts whereas other details do not correspond to anything as such and only posited for some theoretical or formalistic reasons. This is especially apparent when one considers how scientists come to posit mechanism descriptions. There are two important distinctions that can be read as two orthogonal[19] to each other. The first distinction is between *how-possibly models* and *how-actually models* while the second is between the *mechanism-sketches* and *mechanism-schemas* (Craver, 2007). How-possibly models are proposed as conjectures or hypotheses[20] about how a mechanism works. If enough empirical tests

[17] In Bechtel and Abrahamsen's (2005) words "mechanistic explanations explain *why* by explaining *how*."
[18] An example of partial and incomplete model is Hodgkin-Huxley model discussed by Craver (2007) and Glennan (2017).
[19] Note that we owe this remark on orthogonality to Glennan (2017).
[20] Glennan (2017) notes that how-possibly model can be interpreted more broadly where they are models of logical possibilities that may or may not be conjectures. One benefit of this interpretation is that it allows some scientific practices like exploratory simulations or testing hypothetical scenarios

or the background knowledge at hand confirms the model then it becomes and how-actually model. Of course it is not decided at a moment by the relevant epistemic and/or scientific community but as the evidence grows the community increasingly builds consensus. In this context, it is always possible for another how-possibly model with more details or with a different content to emerge and become the how-actually model at present. It is also a possibility that a scientist may not know if the stated components, activities, parts and the organization is correct at the beginning, before any substantial process of empirical testing.[21] On the other hand, *mechanism-sketches* and *mechanism-schemas* are about how precise and/or detailed a mechanistic model is. Craver tends to view these two distinctions in a parallel fashion in that he claims the more a model approaches to become a how-actually model the more it also moves from being mechanism-sketch to mechanism-schema. It becomes more detailed and fine-grained. However, even though that might be a common tendency for models, it does not have to be so since we could easily think of thoroughly fine-grained mechanistic models that are false and coarse-grained, sketchy models that are *more* correct. Therefore, as Glennan suggests it is more accurate to conceive these distinctions as orthogonal to each other. However, remember, these are all about giving a mechanism description for the phenomenal description. So, where this phenomenal description comes from? In fact, it comes from some previous mechanism description probably[22] proposed to account for another phenomenon. As Glennan suggests, all mechanistic models contains gaps, components, activities that need to be further accounted for, hence every mechanistic

---

that are known to be factually false just to make idealized inferences. Glennan then adds another category in between, *how-roughly models* that are conjectures or hypotheses.

[21] Notice that it is not much different in theorization in other kinds of explanatory frameworks. It is still possible and legitimate to posit unobservable or yet unobserved constructs into the model as well.

[22] I say *probably* because that is usually the case but one can imagine situations where one just observes or imagines something as an explanandum out of blue.

model creates new *known-unknowns* to be explained (Glennan, 2005). Mechanistic models are being continually refined and modified. Appropriately, Craver even argues that this constant filling in the gaps found in mechanistic models is what drives the scientific progress.

The emphasis in filling in the gaps and always approximating more details in mechanistic models, especially made by Craver and Darden (2013) led to much criticism (Levy & Bechtel, 2013). The idea rests on problematizing the blurred line between a complete and incomplete model with emphasizing the role and importance of abstraction and idealization in science. Opponents ask whether we should go until the atomic level or even below just to explain some phenomena in, say, social cognition. Even though the whole debate contains various arguments and to fully capture it here is impossible, suffice it to say that the New Mechanists do not have any problem with using explanations that are incomplete in differing degrees whenever it suits the researcher's pragmatic concerns. For them an explanation is *completely* complete when science is complete, i.e. when all explanations are complete. At this point it is obvious that this is an impossible ideal but an ideal consistent with the practice of science and scientists since it always aims to broaden its scope and updates itself forever. Hence, it is no problem for them to admit incompleteness.

Another issue is that Craver and Darden, sometimes with David Kaplan, stated multiple times that their views situate abstraction and idealization in a place no less important than that of opponents (Craver, 2013; Craver & Darden 2013; Craver & Kaplan, 2020). They maintain that every mechanism is a mechanism of some phenomena, which would require the researcher to ignore irrelevant details for the phenomena at hand. But their emphasis on filling in the gaps also stems from the concern that if we did take functional explanations, black boxes and other explanations with many gaps and unexplained subcomponents as *complete*, then there would not be any iterative or progressive way to advance science in a coherent and consistent way.

Agreeing with Craver, Darden, and Kaplan, I would add that every explanation is, by its nature, a model, and every model is necessarily an abstraction—because it is not the thing it represents. We can only understand the system in question through concepts that are already abstracted, along with the relations and activities we propose between its parts. In this sense, abstraction is not something we can avoid; it is an essential part of how we explain anything at all. "The more detail the better" is true in the context of science as a whole but it is always possible to use an explanation at a preferred level according to the context or the relevant theoretical or practical needs. For instance, in neuroscience, one might explain the transmission of a signal across a neural network by referring to action potentials and synaptic transmission without invoking the molecular details of ion channel behavior or neurotransmitter synthesis. In some contexts, this high-level mechanistic account is not only sufficient but preferable such as when modeling large-scale neural activity in computational neuroscience. Thus, while further detail might enrich the explanation in principle, abstraction at a certain level can be pragmatically more fruitful, depending on the context.[23]

The hierarchical structure of mechanistic explanation has become apparent in the above discussion. This is an aspect of the New Mechanism that contrasts with its antecedents in that they almost exclusively focused on etiological mechanisms since

[23] I used the terms idealization and abstraction interchangeably here, following Craver and Kaplan (2020), because everything that is said about one of them also applies equally to the other in this context.

the *levels* are implicit in the new mechanistic science.[24] We explain some phenomenon by either providing an etiological or a constitutive mechanism - and ideally both. Then we explain further the components within that very mechanism. This iterative process goes on as far as our epistemic capacities enable us. However, this process does not only run through from upper *levels* to lower ones, we also *look up* to see if our mechanism is a part of another phenomenon at some other *level* (Bechtel, 2009). This intuitively and intrinsically follows from the characterization of mechanisms and how they explain.

In this context, we can clarify what we mean by "level" with an example. If we take the cell as our example again, the components that make up a cell, namely, the organelles and their activities, constitute one level, whereas the interactions in which the cell, as a whole, participates with other cells constitute a different level. To describe these intercellular interactions, it is not necessary to refer to the internal subcomponents of the cell, at least, this is not a logical necessity. Whether it is an empirical necessity may depend on the specific phenomenon in question. The intercellular interactions of a cell may indeed be influenced by and even depend on its internal processes. However, someone who only studies the organelles without also investigating the cell's intercellular interactions will not be able to establish the relation between the two. In other words, studying organelles alone may not be sufficient for predicting the intercellular behavior of the cells they constitute. This discussion may remind those familiar with the topic of the concept of emergence. The debate over emergence versus reduction has a long history and, roughly speaking, stems from the claim that all sciences can ultimately be reduced to physics. Against this view, it is often argued, especially in fields such as biology and

[24] There is also the issue of whether something has an ontic aspect i.e. a claim about the structure of reality itself. However, in line with my argument on the anti-realism in the previous chapter, this will no longer be relevant for this discussion. I obviously do not commit to such an ontic view.

psychology, that many processes cannot be reduced to physics, and that no matter how well we understand the physical level, this understanding will not suffice to predict what happens at higher levels. A crucial distinction here is whether this limitation is ontic or epistemic in nature. The former is associated with the notion of strong emergence, which holds that the structure of the universe is genuinely multilayered, and that causal relations at levels above physics are just as real as those within it. The latter, known as weak emergence, maintains that irreducibility stems from our cognitive or methodological limitations, rather than from the nature of reality itself.

Notice that even though not entirely incompatible, there is little room for strong emergence here. Mechanistic explanations are mostly taken to be suggestive of a minimal account of emergence where organizational emergence is possible but strong emergence is something to be explained further, if possible. Some mechanist philosophers argue against *spooky emergence* on the basis that the resulting properties of the emergent entity should be the result of the explaining lower-level mechanism even though they might be unpredictable from it. However, spooky or strong emergence opens up a space for new properties that are not the result of the underlying[25] mechanism. On a related note, this is where mechanistic approach departs from mainstream reductionism in that it has no claim regarding the predictability of higher-levels from the lower ones as reductionism would have it. Instead it is open to the idea of *epistemic emergence* as well where the reason we cannot predict the novel properties of the emergent level might be our restricted ability of knowing be it by our technology and/or methodologies or our cognitive capacities.

[25] I am using the word colloquially here.

Notice that the mechanistic levels have also implications for the integration of different disciplines. Disciplines working at different phenomena at different levels might be integrated in an inter-level way naturally via such mechanistic outlook (Bechtel, 2009; Craver & Darden, 2013; Tabery, 2014). A caveat would be that different disciplines might explain the same or related phenomena via different mechanisms which would be problematic for an anti-pluralist but not problematic for a pluralist. However, even if you are an anti-pluralist it is a theoretical and empirical problem that might be solved eventually by further research and testing. This hierarchical structure of mechanistic explanation and the mechanistic conception of emergence (*mechanistic emergence* henceforth) and "levels" will be crucial for the argument I will develop in the next chapter.

3.2 Misconceptions and criticisms about mechanisms

There are a bunch of misconceptions about the mechanistic framework. Some of them are outright novel misunderstandings whereas others stem from the commitments of historical precedents of the New Mechanism. The word *mechanism* and *mechanistic* have certain type of connotations due to that historical background. However, the New Mechanists repeatedly claimed and dismissed most of those problematic commitments. Especially since those misconceptions also feed much controversy and possibly arguments against the New Mechanism, it is important mention them and make clarifications.

One such prominent misconception refers to the alleged incompatibility of abstraction/idealization with mechanisms. This concern mostly stems from some mechanists' emphasis on *more detail is always better.* We already mentioned in the above discussion why this is and cannot be the case provided that mechanistic models are always models of some phenomena and always ignores some components *out there* whereas retaining others.

Another visible misconception is taking mechanisms as being identical with machines (Craver & Tabery, 2023; Darden, 2006). This leads to the criticism that most things, for example organisms, are not machines and hence they cannot be treated mechanistically (Nicholson, 2018). However, the New Mechanists especially eschewed this interpretation and stated that even though the concept of mechanism and machines are both etymologically and historically related they are not necessarily the same thing. Something that is not a machine may still contain or contribute to some sort of mechanisms. To defend a weaker thesis, something that is not a machine can still be modeled in virtue of its component parts, activities and organization. In fact, we can witness this kind of a modelling practice throughout the entire history of science. A salient example is the cardiovascular system in humans. While the heart is not a machine in the literal sense, it is often modeled mechanistically as a pump that drives the flow of blood through a system of vessels, regulated by valves and pressure gradients. Such a model allows for causal explanation and intervention without presupposing that the system is a machine in the strict engineering sense. What matters is not mechanical construction but the presence of organized components and their coordinated activity.

One other controversial issue in this regard is the claim that mechanisms must be stable. Stability here means that the arrangement of parts, activities and the organization be fixed and cannot change. However, scholars like Glennan (2009) talks of *ephemeral mechanisms* where the parts, activities of parts and organization of them can change through time. Notice that this does not lead to a contradiction with those parts, activities and organization and the change of them take place in space and time. How and why does that change occurs in a given mechanism is also something to be explained by some mechanistic models.

Another widespread misconception which sometimes becomes a central line of critique is the claim that mechanistic explanations, as characterized by the New Mechanism, are inherently linear and thus incapable of accounting for phenomena involving feedback, circular causality, or dynamic regulation. This critique, however, rests on a misreading. The New Mechanism does not commit to linearity as a necessary feature of mechanisms. Rather, it emphasizes the organized interaction of entities and activities that jointly produce a phenomenon. In fact, feedback loops are fully compatible with mechanistic explanations, as long as the loop itself can be decomposed into or explained in terms of underlying organized components that sustain the behavior in question. Servomechanisms, for instance, constitute a classic case where feedback is inherent to the system's capacity to actively maintain certain variables within bounds, producing apparent stasis or equilibrium as a result of continuous internal regulation (Bechtel, 2011, 2022). What may seem like a static output is in fact an emergent outcome of dynamically interacting mechanistic components. The New Mechanism is well equipped to capture such dynamics, provided that the relevant explanatory models remain committed to uncovering the constituent parts and their modes of interaction that give rise to such regulated patterns. Therefore, criticisms grounded in the assumption that mechanistic models must be linear or cannot accommodate feedback overlook the flexibility and scope of the New Mechanism.

As a last remark in this context, considering the above discussion, one might then ask, what is it that incompatible with mechanisms? Is the concept of mechanism so trivial that it is compatible with pretty much everything? If mechanism covers all these then what does it not cover? The primary clarification for such a concern is that to be comprehensive to a great extent is not surprising for a framework that claims to explain — at least for some proponents — how science *actually* works. We mentioned that there is a broad interpretation that takes New Mechanism as both descriptive and normative framework capturing the entirety of science and the ontology of explanation. In this reading, other kinds of models are explanatory only in so far as they can capture and take into account the mechanistic activities and components. For example, Kaplan and Bechtel (2011) claim dynamical models are illuminating in terms of temporal organizations of interactive mechanisms whereas others think that they are only descriptive models and do not provide an explanation (for an application of this argument in cognitive science, see; Erdin, 2020). This is also the case for other so-called explanations like topological explanations and network explanations. For mechanists, they fail to distinguish between good but entirely fictional predictive models and when they do not fail to distinguish them it is only possible through some implicit or explicit mechanistic model. One could argue that this is precisely what distinguishes an explanation from a simulation. Otherwise, just as in the case of deep learning models, we could endlessly increase the predictive power of our model by adding countless elements whose correspondence to any structure or causal organization in the target system or phenomenon (e.g., the brain or a given cognitive function) remains unspecified or unknown. Yet this would not suffice for explaining the system. As the mechanists argue, in order to legitimately claim that we have explained a system, the components or processes used in the model must correspond to actual entities or processes within the system itself. In this sense, the model must be medium-dependent (Bechtel & Abrahamsen, 2010; Milkowski, 2011, 2016).

Accordingly, to clarify the discussion it is safe to say an explanation must be medium-dependent and what makes a model explanatory is its content rather than its form or predictions. Therefore, network models, topological models, dynamical models, etc. are all the cases of medium-independent models.

However, since scholars who take a narrow interpretation of mechanism mostly do not think a mechanistic explanation is compatible with any of those we mentioned above, this triviality objection would not arise in their case. Nevertheless we can refer to a short list given by Craver and Tabery (2023) as examples of what are not mechanisms: entities, objects, correlations, inferences, reasons, arguments, symmetries, fundamental laws, relations of logical and mathematical necessity.

For the present discussion, the following quotation from the same source might be helpful as well:

> The idea of mechanism is a central part of the explanatory ideal of understanding the world by learning its causal structure. The history of science contains many other conceptions of scientific explanation and understanding . . . Some have held that the world should be understood in terms of divine motives. Some have held that natural phenomena should be understood teleologically. Others have been convinced that understanding the natural world is nothing more than being able to predict its behavior. Commitment to mechanism as a framework concept is commitment to something distinct from and, for many, exclusive of, these alternative conceptions. If this appears trivial rather than a central achievement in the history of science, it is because the mechanistic perspective now so thoroughly dominates our scientific worldview (Craver & Tabery, 2023).

3.3. New Mechanism goes unreal

I argued for an anti-realist conception of science in Chapter 2. However, most mechanists commit to a realist view in that they take it that in science we discover mechanisms (out there) *in nature* (Bechtel & Abrahamsen, 2005; Glennan, 2005). That might possibly lead to a confusion that needs to be dealt with here.

First of all, I should make something clear: the discussion about realism versus anti-realism is mainly concerned with science in its entirety, meaning that it does not say anything particular for some modeling practice in science. It concerns itself with what any modeling practice in science can achieve in terms of mind-independent truth and/or reality.

On the other hand, mechanist philosophers' commitment to some form of realism might not stem from their commitments to mechanistic modelling but from realism being the dominant philosophical stance in the philosophy of science academia. In this regard, Colombo, Harmann and Van Iersel's article (2015) takes up the task of showing why and how modelling practice in mechanistic science does not guarantee realist commitments just by themselselves as there is no necessary relation between the two. In doing that, they first show why mere empirical adequacy cannot give us realist commitments in a similar vein as we did above. Then they argue that the fact that a *true* mechanistic model cannot be *true mind-independent* does not prevent the models from being useful, insightful and explanatory according to the criteria set by the relevant scientific community at the time. They maintain that those criteria are also already used in determining whether a model is *true* in the first place — whether construed realistically or anti-realistically. Therefore, they claim that what makes a mechanistic model, *true* or in more accurate terms, *correct* or *accepted,* is the empirical *coherence* of that model with the background knowledge of the scientific community. So far so good... Yet, they mostly construe coherence in an empirical way, which is in terms of beliefs and evidences. Even though this is not problematic for their paper's aims and scope because they want to show how mechanistic modelling is in principle equivalent to that of mechanism with a realist belt, it has a missing part. We know that science is something more than *belief plus* *evidence* as most clearly portrayed by the underdetermination discussion. What kind of beliefs do we form, namely, what kind of concepts do we employ, what is our *ontology*, our “conceptual toolkit” are also matters of *coherence.* Because obviously we do not form beliefs about angels or fairies and test them in science but do so about particular things and events. We only take particular kinds of concepts relevant to scientific investigation. Might there be a coherence criterion that can help us to clarify this observation further?

CHAPTER 4

INTEGRATING SCIENCE AT ITS JOINTS

Most debates within research traditions in science, or more specifically in cognitive science, revolve around the questions of realism, i.e., the realism of different entities or theoretical posits. Hence throughout the earlier chapters we looked into whether science could give us an answer concerning the reality of anything at all. My answer was no. One consequence of the discussion about realism versus anti-realism we presented here is the practice approach to science that takes science as a cultural practice, as some sort of a cultural institution. Just like other cultural practices, it cannot grant a logical link to mind-independent reality or truth per se, but it may very well have its own distinct rules — either implicit or explicit — that distinguish it from other and at the same time gives it its uniqueness. This reference to explicit or implicit rules was inspired by a Wittgensteinian contextualism which is already implicit in the sociohistorical and cultural approaches to science. But contrary to most takes in philosophy of science we refrained from appealing to a mere instrumentalism according to which those rules are determined by local contexts and needs of the given epistemic community, instead we stated that one rule as to how science *explains* its subject matter might be the mechanistic explanations construed by the New Mechanism. There are two ways to interpret New Mechanism one of which is the narrow reading where the New Mechanism is both a descriptive and normative framework of a particular way of doing science albeit not the only one. The other reading is the broad reading which takes the New Mechanism as both a descriptive and normative framework aimed to cover the entirety of science. I argued that even though I employ a broad interpretation here, it is not necessarily so in terms of the scope of the present thesis and an explanatory pluralist might still find the present argument for a relatively local context useful or inspiring. One possible concern taking mechanistic explanation as a rule in an anti-realist conception of science would be the dominance of scientific realism among the New Mechanist philosophers. Relying on the work of Colombo, Hartmann and Van Iersel (2015), we saw that there is no logical and necessary incompatibility between the two. In this treatment though, the concept of *coherence*, as opposed to truth/reality, became prominent in accepting a particular explanation in science. As opposed to Colombo et al. (2015)'s broad way of construing coherence by only referring to evidence and background knowledge, which makes it only a matter of empirical investigation, I proposed that there is a need to another coherence criterion concerning conceptual aspects of research traditions because what kind of beliefs we form, what kind of tests we employ upon them, and as a result, what to accept and what not to accept are all determined, or at least, importantly influenced by such conceptual constraints.[26] We do not ask empirical questions about fairies and angels but about particular kinds of things and events, so what leads to the distinction at hand here?[27]

In this chapter, I am going to deal with this concern and investigate what would be conceptual criterion of coherence in an anti-realist science in which the

[26] Notice that this is in the vicinity of both Laudan's distinctions between the theoretical/conceptual, the empirical/methodological and the institutional. A somewhat parallel distinction can be found in (Sanches de Oliveira & Baggs, 2023). Though both accounts make a tripartite distinction there is a case to be made in favor of a quadripartite layering including the distinction between *the theoretical as in empirical theories* and *the theoretical as in philosophical theories*, e.g. local ontologies, conceptual frameworks, etc. A useful example might come from cognitive science where two opposing models of memory, Atkinson-Shiffrin Model of Memory (Atkinson & Shiffrin, 1968) and Distinctiveness Theory of Memory (Glenberg & Swanson, 1986; Glenberg, 1997; Nairne, 2006; Schmidt, 1991) share the same philosophical/ontological/conceptual background, i.e. computationalism, but differ in their empirical contents and predictions. Likewise, the debate between classical cognitive science and connectionist cognitive science would lie somewhere between the two as it is partly a philosophical/ontological/conceptual matter as well as partly an empirical matter.
[27] Notice also that this concern is in line with Sider (2011) or with the discourse around *carving nature at its joints* in general. However, as the present anti-realist stance prevents us talking about *carving nature* objectively, the title of the chapter is "integrating science at its joints." We will talk abpout this furthermore later.

New Mechanism will prove helpful. But before that, I will first tackle the mainstream view in cognitive science and the problems it faced according to the framework presented here. That discussion will further help us navigate through the conceptual criterion subject as well. I will argue that there are three foundational commitments that historically and logically gave rise to the current fundamental debates in cognitive science: 1) multiple realizability of mental states, 2) autonomy of special sciences, 3) naturalizability of the mental. I will argue that these three are in fact interdependent and lead to same kinds of thorny problems for pretty much every framework in the field. Then I will propose and explain how conceptual coherence criterion can bypass these problems

4.1 Foundations of the science of mind

When you open up a textbook, a dictionary or a common internet source to see what cognitive science is, you would likely see something like this: "Cognitive science is the interdisciplinary study of mind (and/or intelligence.)" George Miller, one of the founding figures of cognitive science, even says that they were reluctant to use "mind" because of the dominance of behaviorism, so they used "cognition" in its stead (Miller, 2003). So, cognitive science has been defined over the concept of mind/cognition.[28] But what is mind? What was mind back at the time when cognitive science emerged?

In the early and mid twentieth century, there were two main strands in philosophy of mind answering the very question: behaviorism and brain-identity theory. Roughly, behaviorism was the claim that mental concepts mean nothing more and beyond the dispositional characteristics of an entity. Just like a glass is

[28] Throughout this thesis, *mind* and *cognition* are used interchangeably if not stated otherwise.

predisposed to break when it falls from a certain height, when we say someone believes something we only refer to their disposition to do something or to act some particular way in certain contexts, namely we take their probability to do that or act in some particular way more higher than the alternatives[29] (Ryle, 1949). On the other hand, brain-identity theory refers to the idea that mental concepts refer to brain states (Feigl, 1958; Place, 1956; Smart, 1959). Notice that that does not mean that mind and the brain are exactly the same things as clearly exemplified by our ability to talk about how many neurons there are in the brain in contrast to our inability to do the same about mind. There are no neurons in the mind. One may construe this as a constitution relation like Place (1954, 1656) or as different senses with the same reference (Feigl, 1958; Smart, 1959).

In a series of influential papers Putnam (1960, 1963, 1967) criticized these views. His criticisms were grounded in a couple of interrelated arguments. In his 1963 article, although he initially approaches behaviorism through the concept of translatability, he quickly dismisses this perspective by noting that virtually no one still defends it. He then reformulates the view, proposing that behaviorism rests on two fundamental claims: 1) While mental terms and behavioral terms may not be fully translatable into one another, there exists a certain type of entailment relation between them. Even if these entailments are not strictly analytic, they still emerge in some way from the meanings of the terms involved. 2) The lack of one-to-one translatability stems from the fact that mental terms are of lower resolution compared to behavioral terms. Behavioral terms are more precise and varied, making them capable of capturing distinctions that mental terms cannot completely illustrate (Putnam, 1963, p. 327). Then he goes on to criticize behaviorism.

[29] Notice that this is different from other kinds of behaviorisms sometimes called psychological and methodological. Here we mention only the philosophical or logical behaviorism as it is the subject of philosophy of mind and what mind is.

Putnam's response builds on an example and a thought experiment. He draws his example from medicine, arguing that whether something counts as a disease is determined by the relevant symptoms, which may have different causes. Moreover, it is possible to discover that what was once thought to be the cause of a disease is not its actual cause, and that other factors are responsible. However, changes in the identified causes do not affect the definition of the disease or the determination of whether a person has the disease. He uses multiple sclerosis (MS) as an example to illustrate this point. In essence, we do not equate causes with effects. The effect and the cause are distinct entities, and a disease is simply the name we give to a particular effect, regardless of its underlying causes.[30]

Based on the first example, Putnam argues that a mental state, such as "pain," is not merely a cluster of symptoms or responses but rather the cause of a cluster of symptoms or responses. He supports this claim by appealing to our ability to conceive of possible worlds where a mental state like pain causes no observable responses at all. In the thought experiment he proposes, we can imagine "super-Spartans" who, due to cultural norms, never express their pain outwardly, regardless of how much they feel it. Moreover, we could even imagine that after many generations, this trait evolves to become entirely biological rather than culturally conditioned. In this scenario, super-Spartans could still verbally report feeling pain when asked, but they would exhibit no other external responses associated with pain. If such a scenario does not seem "nonsensical" or inherently self-contradictory, then

[30] In fact, Putnam seems to have already assumed the conclusion he intends to argue for. It remains unclear why the mind should be compared to diseases or whether different examples from history of medicine could challenge this metaphor, a question best left to medical experts rather than to tackle here. However, at least in some cases, it does not seem entirely implausible to think that a disease might be identified not only by its symptoms but also by its causes. When we learn that the causes differ, it is conceivable that we might introduce a new taxonomic category based on the cause rather than the symptom alone. In fact, as we will see later, this objection touches on the same point raised by Jaegwon Kim with his example of "jade." However, let us set this issue aside for now.

behaviorism's claim cannot be correct. According to Putnam, this possibility shows that mental states could, at least in principle, be entirely independent of observable behaviors. Therefore, behaviorism's fundamental claim that mental terms are in some sense entailed by behavioral terms becomes untenable.[31]

Before turning to Putnam's positive argument about how the mind should be conceptualized, let us consider his critique of identity theory. In fact, rather than developing a detailed critique, he makes a few brief remarks in his 1967 paper (Putnam, 1967). For example, he suggests that saying "heat is mean molecular kinetic energy" is more plausible than asserting "pain is…" or "being in pain is being in a particular brain state." This is because, in the latter case, the spatiotemporal relation changes, making the identity claim less plausible since mind is not something spatiotemporal. We cannot say mind is of such and such height, length or weight, or we cannot measure it by any other spatiotemporal measure.[32]

In addition, there are arguments suggesting there may be creatures that we might reasonably attribute *pain* or other mental categories to, even if the attribution do not correspond to any particular type of reductionistic relation. Evolution shows significant variation in this respect. For instance, it seems plausible that there could be creatures showing similar behaviors or engaging in seemingly identical mental activities while sharing little to nothing in terms of evolutionary or physiological mechanisms. Given the diversity of life and the different evolutionary trajectories, such possibilities seem likely. Following this consideration, Putnam shifts from

---

[31] The status of conceivability arguments and how we should evaluate them is a controversial meta-philosophical issue. For further discussion, see: (Gendler & Hawthorne, 2002; Hallett, 1991)

[32] I find this reasonable as well. For precisely this reason, I will defend *exceptionalism of the mental* in the context of mental causation and multiple realizability below. This argument aligns with that of Feyerabend's (Feyerabend, 1963) and, in a certain sense, with the phenomenological tradition. However, adopting this perspective is not a prerequisite for the present thesis. One could accept the argument presented in this thesis without considering the mental as exceptional in this regard. Though, I believe doing so would lead to other problems independent of the thesis's central issues. These concerns will be addressed in the conclusion.

offering an a priori critique of identity theory to presenting his own alternative, *functionalism* or more specifically *machine functionalism*, which he defends as a more reasonable "hypothesis." According to this view, a mental state is neither a behavioral state nor a brain state but rather a functional state. What does this mean?

To explain and better illustrate this point, Putnam refers to the Turing Machine. Just as the states in a Turing Machine are functional in the sense that each state directs us to another state based on specific instructions in the machine's instruction table without being dependent on the physical substrate that the Turing Machine runs, mental states are also functional states. In this framework, the type of hardware on which the Turing Machine operates or the materials from which it is made is irrelevant. What matters is that the machine implements the relevant rules, regularities, and relations between states. This notion, in fact, marks the emergence of the computer metaphor and the computationalist approach to mind (Chalmers, 2011; Milkowski, 2018; Newell, 1980; Pylyshyn, 1980, 1984; Searle, 1990; Shagrir, 2006). Each state in the system is interrelated and causally interdependent, yet "independent" from lower-level physical substrates in a certain sense. According to Putnam, the critical point is that if a mental state M1 leads to another mental state M35 in response to a specific input and there exists a causal relation of this sort with a particular output, then attributing the corresponding mental ascription to any system exhibiting such a relationship is justified. Note that, these relations need not be entirely deterministic; they could be probabilistic, but that is a trivial issue, one that can be addressed during the process of modeling or theorizing. This causal structure remains entirely independent of the system's physical substrate. As readers familiar with the subject would recognize, this is precisely what the concept of *multiple realizability* corresponds to.

Notice that while we have discussed functionalism and machine functionalism, it is not a logical necessity for a functionalist to refer to Turing Machine. The relevant functional relations can be conceptualized in different ways. In other words, at least in principle, one can be a functionalist without being a computationalist.[33] In fact, many non-computationalist or even anti-computationalist frameworks in cognitive science are examples of this as long as they allow the *multiple realizability* of mind.

It is not possible to deal within the scope of this thesis with the entire history of functionalism and the extensive debates surrounding it over the years. Rather, my aim here is to highlight some of the fundamental concepts, approaches, and tendencies that played a central role in the emergence of cognitive science. I will then mention how these ideas persist, even in opposing camps within contemporary debates, and explain what the thesis developed here contributes to these conceptual axes. I hope that the discussion so far has conveyed the main motifs of functionalism. However, cognitive science also involves other important issues that run parallel to or stem from functionalism, which must be addressed as well. For instance, around the same time with the developments mentioned above, Jerry Fodor, building on these ideas, developed the notions of *the autonomy of the special sciences* and the language of thought hypothesis. As for the latter one, more specifically, he argued for the need to integrate *the computational theory of mind* with *the representational theory of mind*, laying the foundation of *cognitivism*. These two ideas, rooted in functionalism, have become some of the most influential views in cognitive science. Even in later embodied and radically embodied approaches that

[33] In fact, even Putnam himself acknowledges that saying "the mind is a machine" is somewhat trivial, as pretty much anything can be considered as a machine when viewed through this lens. This perspective helps explain the rise of pancomputationalist and panmentalist views in the field, particularly over the past two decades which is an unsurprising development given that such stances follow naturally from functionalism.

oppose computational and representational theories, we see that much of the debate still unfolds on this conceptual ground. After addressing these issues, we will proceed to the core propositions of this thesis.

As we just mentioned, another important and related idea is *the autonomy of special sciences* (Fodor, 1974; 1997). The autonomy of special sciences refers to the idea that specialized scientific disciplines, such as biology, psychology, or economics, operate independently of more fundamental sciences like physics or chemistry. This independence is rooted in the unique entities, properties, and regularities that these sciences study, which cannot always be fully reduced to or explained by the principles of the more basic sciences. For instance, while physics underpins all physical phenomena, the concepts and explanations used in biology, such as evolution or ecosystems, are specific to living systems and cannot be entirely derived from physics alone. So far, so good... But just as pretty much every idea in philosophy, it seems completely plausible when explained in just a short paragraph as this. So let us explore it a bit more.

Fodor (1974) begins his article with the statement, "*Every science implies a taxonomy of the events in its universe of discourse.*" According to him, there are natural kinds to which lawlike generalizations apply. For example, while a predicate like "moving an object over a distance of 3 kilometers" might apply to many objects and instances, it is nevertheless not part of any scientific vocabulary because the things it applies to are arbitrary, contingent, and disjunctive. Its relationship to its domain, to the objects and instances it applies, is not lawlike. However, generalizations in psychology or economics, for Fodor, do not share this characteristic. In these fields, we can formulate highly effective generalizations that hold and yield new predictions and allow us to discover new regularities as well.

This implies that the taxonomies in these fields employ natural kinds. Notice that this view poses a strong relation between lawlike regularities and the natural kinds, as Fodor himself also acknowledges. Even though this might be something dubitable, it is beyond the scope of this discussion.

These natural kinds, in turn, determine the vocabulary and taxonomy employed by the special sciences. Although Fodor does not use this term, for the purposes of our discussion, we might refer to these taxonomies and/or vocabularies as *special ontologies*. Special ontologies are the ontologies posited by the special sciences. As the name implies, special ontologies are special, that is they cannot be reduced to lower levels. On the other hand, where Fodor diverges from reductionism is in the requirement, according to reductionism, that the natural kinds of the special sciences must be coextensive with physical natural kinds i.e., the natural kinds in lower-level sciences like physics[34]. This would be analogous to the case of heat being equivalent to mean molecular kinetic energy. However, Fodor argues that such coextension is impossible for psychology and cognitive science. For example, he claims that many regularities in psychology are not coextensive with neurological processes as in the case of heat and mean molecular kinetic energy. For any given psychological function, there is potentially an unlimited, or if not unlimited, a highly complex and disjunctive set of neurological correlates. As a result, the laws of special sciences cannot be mapped onto the laws of physical sciences, as the coextension criterion required for such mapping cannot be fulfilled. Fodor takes this argument so far as to list certain disciplines and assert that the work being conducted within them is problematic:

[34] This distinction between lower-level and higher-level sciences is of course not unproblematic. Most simply because it is obvious that physics is not just concerned with small-scale phenomena but it also aims to explain astrophysical events. But the distinction, at least for the discussion at hand, is conventional.

"If psychology is reducible to neurology, then for every psychological natural kind predicate there is a co-extensive neurological natural kind predicate, and the generalization which states this co-extension is a law. Clearly, many psychologists believe something of the sort. There are departments of 'psycho-biology' or 'psychology and brain science' in universities throughout the world whose very existence is an institutionalized gamble that such lawful co-extensions can be found. Yet, as has been frequently remarked in recent discussions of materialism, there are good grounds for hedging these bets." (Fodor, 1974, p. 105)

When considered in the context of cognitive science and psychology, concepts such as *belief*, *attention*, *memory*, *perception*, *mental representation,* and, depending on the context, *encoding* and *computation* are part of the special ontologies of these fields. Note the connection between this discussion and Putnam's functionalism and the thesis of multiple realizability. While Putnam critiques reductionism in the context of philosophy of mind and advocates for functionalism and the independence of mental terms from their physical bases, Fodor adopts a similar strategy within the philosophy of science. Fodor emphasizes that reduction is not appropriate from the perspective of the philosophy of science and he argues that higher-level phenomena must be addressed on their own terms, with their own ontologies and methodologies. Furthermore, the regularities that special sciences enable us to discover, and the predictions they allow us to make, should be evaluated based on their own internal criteria. For instance, later in *Psychosemantics* (1987), Fodor defends folk psychology as fundamental to cognitive science and psychology. For him, if we agree with a friend to meet at a specific place, at a specific time, on a specific day a year from now, *ceteris paribus*, this agreement will hold, folk psychology works remarkably well in such contexts. While I find this perspective compelling, particularly regarding the relationship between reduction and the interrelation between sciences, I also have certain concerns that make me somewhat skeptical of this view.

First, for a reductionist, as Fodor himself acknowledges but ultimately rejects, there is a common reductionist narrative according to which progress in science, including so-called special sciences, often occurs through the increasing association of the phenomena in special sciences with sub-disciplines and their eventual reduction to them. Namely, as the science progresses the special sciences are expected to disappear since they are being reduced to the lower, reducing sciences through this progress. Notice that here the backbone of scientific explanation is reduction. In Fodor's case, however, it remains unclear how progress is to be achieved or what meta-scientific criteria we should adopt in theory evaluation. This becomes particularly evident when we consider Fodor's claim that appealing to lower, 'reducing' disciplines to explain exceptions and anomalies in special sciences is both normal and necessary.[35] This opens the door, in principle, to *ad hoc* arguments against any given counter-evidence in the disputes and competition between theories. After all, any theory or framework, when evaluated by its intra-theoretical criteria, can appear highly explanatory. What exposes the explanatory limits of a theory is often a potential new theory, sometimes

---

[35] Fodor supports this claim by pointing out that there are exceptions to the laws in special sciences but the laws of physics or chemistry do not have exceptions. So, when an exception is observed in the special sciences it is normal to refer to the more powerful and exceptionless laws of the lower levels. However, this seems to be a false dichotomy as whether there is a clear-cut and categoric distinction between the natural sciences and higher-level special sciences in terms of laws and exceptions is dubitable. Physical laws seem exceptionless because we deal with them only under strictly controlled conditions. The problem with the laws of the special sciences is that we often lack the tools to provide such strict controls, and these tools may, in principle, be impossible to develop due to ethical and epistemological constraints. Therefore, making a categorical distinction between explanatory practices in science on this basis is unjustified and problematic. Once this distinction is removed, various pathways for bridging special sciences and natural sciences become apparent, pathways that bypass typical categorizations between reduction and emergence. As we will explore below, mechanistic interlevel integration may offer a promising pathway in this context.

accompanied by a significantly revised ontology that provides a reference for comparison. Fodor’s proposed framework creates a significant gap in this regard. There might even be a case to argue that the current theoretical and methodological crises in cognitive science and psychology are related to these meta-scientific views (Gigerenzer, 2010; Hommel & Colzato, 2015; Sanches de Oliveira & Baggs, 2023). On the other hand, whether the proper relationship between sciences in this context should indeed be reduction is a distinct matter, which we will return later. However, before doing so, we will address a major critique of Fodor’s position and his response to it, thereby concluding the discussion on the autonomy of special sciences.

Famously, Jaegwon Kim disagrees with Fodor (Kim, 1989; 1992). For Kim, non-reductive physicalism[36] is untenable. Such a view must either collapse into a form of physicalism or lead to eliminativism. Otherwise, it would require positing non-physical causes, which is problematic since all causal processes depend on physical[37] events. Without this dependence, we lose any justification for excluding "spooky" concepts and explanations. Thus, explaining something through non-physical means when physical explanations are available becomes arbitrary and violates the principle of parsimony. For instance, Kim argues that to explain a psychological and/or mental phenomenon, such as pain, we must engage in local reductions. While Fodor is correct in asserting that what we call "pain" could operate on entirely unrelated physical mechanisms or principles for a Martian, an octopus, an insect, and a human—making it impossible to reduce all instances of pain to a single, unified explanation or reduction—we can still reduce each instance separately. That

[36] Functionalism has been proposed as a stance both non-reductive and still physicalist, hence the naming. However, later, non-reductive physicalism has gained its own trajectory.
[37] Notice that that does not mean that every phenomenon belongs to the domain of phsyics as a science. It just means that in naturalistic explanations we do not refer to super-natural or extra-natural elements.

is, we could reduce Martian pain, human pain, octopus pain, and so on individually, treating them as distinct types of pain. Why should we not do this? If we choose not to reduce these phenomena in this way, then we might decide not to call them "pain" at all. Instead, we could address the behavioral phenomena using different and newly formulated concepts, eliminating mental concepts altogether. However, this approach would no longer be non-reductive physicalism—it would amount to eliminativism.

According to this view, special sciences are not autonomous, and in any case, explaining phenomena within these disciplines requires progressively appealing to lower-level regularities and that is entirely common scientific practice. For Kim, the fact that this does not yield a lawlike regularity between lower-level and higher-level phenomena is not a problem. A common criticism is that it would prevent us from reaching generalizations about psychological and mental phenomena, leaving our explanations confined to local instances. However, according to the critiques, one of the primary goals of science is to arrive at generalizations. Whether the aim of science is to achieve generalizations is a long-standing and complicated debate in the philosophy of science. Even though the debate lies outside the scope of this thesis, suffice it to say that the dominance of local explanations rather than broad generalizations in many areas of biology often serves as a common counter-example in these discussions. On the other hand, this criticism is closely related to the notion of multiple realizability. According to Kim's views, as outlined above, there is no single, multiply realized instance of *pain* or any other mental state for that matter. Consequently, it would be unreasonable to expect observations about one instance of, say, Martian pain, to lead to generalizations about another, say, octopus pain.

Therefore, even if the goal of science is to achieve generalizations, Fodor's claims do not provide a plausible framework for reaching such generalizations.

Kim illustrates his critique of multiple realizability most famously with the example of jade. Kim wants us to consider that with the progress in science, we discovered that what we commonly refer to as jade is actually composed of two distinct substances: jadeite and nephrite. This means that, despite their striking similarity in macro-physical properties, nearly identical in appearance and most other macrophysical observations, they differ at the microphysical level. If someone attempts to sell nephrite as jade, they would be committing fraud because it is not genuine jade. Kim argues that, for the sake of discussion, let us assume that all our observations of jade thus far have been based either on jadeite or nephrite, without any observations of the other. Given that we now know these two substances are distinct, would it be justified to make predictions about one based on observations of the other simply because they share significant similarities in macro-physical properties and have been falsely classified under the same kind, "jade"? Similarly, how can we make predictions about a Martian's experience of pain or an octopus's pain based on findings from studies on human psychological phenomena, such as pain? For Kim, There is no singular, unified "thing" being realized across these cases. Consequently, we cannot project from one onto the other. Kim goes even further, arguing that Fodor's argument cannot lead to the conclusion that psychology is autonomous but rather to the conclusion that psychology lacks a proper subject matter. Ultimately, based on this argument, Kim concludes that multiple realizability is a problematic conception.

Fodor's response to this critique is somewhat interesting (Fodor, 1997). He actually agrees that jade is not multiply realizable. However, he argues that this is not due to whether jadeite and nephrite are projectible but rather because jade is not multiply realizable in the first place. He claims that jade has a nomological relationship with jadeite in all possible worlds, much like water's relationship to H_2O[38]. In contrast, this is not the case for pain or other mental states. For Fodor, it's the other way around. We do not arrive at multiple realizability because making inferences from one instance to another is justified; rather, if something is multiply realizable, then making inferences from one instance to another is justified. He goes even further and asserts that, for a functionalist, mental states already multiply realizable. Therefore, for a functionalist, it is justified to infer from human pain to octopus pain or Martian pain, precisely because mental states are multiply realizable in the first place. In contrast, jade is not multiply realizable. However, that means that Fodor has not actually proven anything, nor has he provided any argument to support the claim that the mental states are multiply realizable in the way and to the extent he suggests. In fact, by presupposing this from the outset, he appears to be drawing a circular picture.

Let us recall that Putnam also presented functionalism not as a grounded theory but as a "possible alternative hypothesis." Fodor makes the same move here, and after years of debate, we find ourselves back where we started. Thus, in its current state, functionalism remains an arbitrary theoretical preference.The debate over multiple realizability is an exceptionally long and complex one and it is impossible to cover its entirety here. It spans across a vast field of conceptualizations at the intersection of philosophy of language, philosophy of science, metaphysics, and philosophy of mind. Moreover, there are more recent approaches, as well as alternative perspectives and variations that we cannot delve into here but which

[38] I mentioned Putnam's microessentialist views in the second chapter and found them problematic.

certainly require separate discussion.The aim of this thesis is not, by any means, to make a definitive claim about whether multiple realizability is true or false. However, the discussion so far suggests the following: some things may indeed be multiply realizable, while others may not. In some cases, the material and/or physical substrates may be indispensable for defining a phenomenon or a concept. Yet, claiming that mental states are multiply realizable and defining it functionally is neither a neutral nor a given perspective in cognitive science, it is a theoretical choice and should be recognized as such.

We have discussed above how *the autonomy of special sciences* and *multiple realizability* are closely associated.[39] While we refrained from adopting a general stance on multiple realizability, we concluded that functionalism is an ungrounded and unjustified thesis. Similarly, we noted that these two theses are not theory-neutral propositions for cognitive science; rather, they are theoretical choices and, as such, can be criticized or rejected in the context of theoretical disputes. Let me clarify that I remain entirely agnostic about the status of multiple realizability or functionalism in the context of other disciplines or metaphysical debates. Within the scope of this thesis, I focus exclusively on their relevance to cognitive science. This is because, even if multiple realizability and functionalism are entirely reasonable and valid for many objects and concepts, they may not hold for *mind*. It is possible and likely that even though one day the relationships between higher-level and lower-level sciences may be well-grounded for various objects and concepts at the human scale, mind could represent an exception in this regard. For instance, we do

[39] While these two positions are often found together and have historically developed in close parallel, it is important to stress that there is no necessary or logical entailment between them. Rather, they have been mutually suggestive. The plausibility of one often lends support to the other.

not speak of a "table-atom problem,[40]" yet we do discuss the "mind-body problem." In this context, I argue that even if we resolve how to ground and scientifically explain all other objects and concepts at the human scale, understanding the relationship between some higher-level and lower-level phenomena without any problematic issue, it would still be unjustified to assume that that framework also applies to mind. This stance, which I would term *exceptionalism about mind*, is of course rooted in *the mark of the mental* debate. To further engage with this issue, let us explore another topic arising from the two foundational pillars of functionalism discussed above, and its implications for cognitive science and philosophy of mind. Ultimately, this discussion will lead us to *the hard problem of content*. [41]

Let us remember Putnam's machine functionalism, a.k.a. computational theory of mind. Putnam conceptualizes mental states in terms of functional roles, analogous to the states of a Turing machine. He argues that mental states are not reducible to specific physical or biological structures but are instead defined by their causal roles in a system. Just as the states of a Turing machine are defined by the functional relations between inputs, outputs, and transitions governed by its instruction table, mental states are defined by their functional interrelations within a broader system of inputs, behavioral outputs, and other mental states. This perspective is also endorsed by Fodor.

---

[40] Similar issues are discussed in the context of metaphysical identity. However, there is a categorical difference between grounding a table or other objects at that scale and grounding mind, particularly in terms of causation and the philosophy of science. The former is merely a matter of scale, whereas the latter involves fundamentally different problems beyond the issue of scale. Even if we were able to fully understand the relationship between higher-level and lower-level phenomena, we would still be left with an unanswered question: what exactly sets apart ordinary objects, like tables and chairs, which exist at a scale comparable to human behavior, from things that are considered mental? Notice that this is directly related to the *mark of the mental a.k.a intentionality,* and *the hard problem of content.*

[41] The discussions on multiple realizability, higher-level and lower-level phenomena, the unity of science, and similar topics often involve realist assumptions, such as those concerning the *fundamental* level of reality. Therefore, while the sketch presented here is as concise and aligned with the mainstream narrative as possible, there are numerous points that I might disagree with upon deeper consideration. Consequently, the reader should approach this sketch with a grain of salt.

Later, however, Fodor finds this view inadequate and proposes necessary add-ons (Fodor, 1975, 1981, 1983, 1987, 2008). He argues that, if approached solely in this way, it would be scientifically trivial, as it would allow for the proposing an if-then instruction table for every single behavior without any criteria to constrain it. Moreover, he contends that the goal of a scientific inquiry into mind cannot be an endless effort to provide if-then explanations for every single behavior. It is worth noting the parallelisms between this critique and the earlier discussion of local explanations versus generalizations. For Fodor, any theory of human thought must account for two key features: *productivity* and *systematicity*. According to this view, human thought and behavior are potentially infinite and exhibit systematic relationships. A classic example of systematicity is the relationship between the sentences "John loves Mary" and "Mary loves John." Anyone capable of forming or understanding the first sentence should also be able to form or understand the second, demonstrating a systematic relationship between the two. Fodor emphasizes the need for a principle or constraint that accounts for this systematicity and introduces the Language of Thought (LoT) hypothesis. Assuming that LoT is well-known in this context, suffice it to say that Fodor’s view holds that thought is regulated by language-like systematic and productive relationships (See for more on LoT: Quilty-Dunn, Porot & Mendelbaum, 2023; Fodor,1975, 2008; Gomila, 2011; Lupyan, 2023; Rescorla, 2024). At the core of this are semantic representations that the "machine" performs computations on. These semantic representations possess combinatorial properties that establish the rules and constraints governing their permissible combinations. These rules and constraints, in turn, underlie both systematicity and productivity. The best example of this can be found in natural language: despite having a finite set of elements, natural language allows for the creation of an infinite number of systematic expressions and sentences due to similar rules and constraints.

This idea which is often referred to as Fodorian Cognitivism has led to extensive debates and has branched into numerous subfields. However, at its core lies a critical question that has posed a serious challenge for cognitive science and the philosophy of mind: where and how do these representations acquire their semantic content? Remember, according to the functionalism, mental representations *qua mental* can be realized by different substrates hence making different physical substrates with the identical content possible. These physical substrates called *vehicles* of representations. So, the question is how do these different sorts of vehicles can acquire their content? This issue has led to significant theoretical crises, as exemplified by the Chinese Room argument (Searle, 1980) and the symbol grounding problem (Harnad, 1990), both of which are closely related to this question. While various theories of content have been proposed, the rise of embodied approaches, particularly since the 1990s, can be seen as a response to this crisis. These approaches, which emphasize the role of the body and the environment in cognition, have generated a wide array of responses across different schools of thought and theoretical traditions. Notice that the problem of content historically goes back to the question of intentionality and *the mark of the mental* (Brentano, 1995; Crane, 1998). So it means that what we take to be contentful is minded and what we take to be minded is contentful. If we can answer this question then we can know the categorical difference between a table or a chair and a human or some other creature's mind. In this context, some favoured something in the vicinity of *embodied representations* that take their content through sensorimotor information as opposed to classical cognitivism's amodal representations (Barsalou, 1999, 2008;

Glenberg, 1997; Goldman, 2013). On the other hand, some approaches altogether rejected the notion of representation if conceptualized in the cognitivism's way (Van Gelder, 1995; Varela, Thompson & Rosch, 1991). Among such views are enactivism (Chemero, 2009) and radical embodied cognitive science (REC) (Varela, 1979; Varela, Thompson & Rosch, 1991). Enactivism proposes life-mind continuity thesis and grounds *the mark of the mental* in terms of *autopoiesis*.[42] For an enactivist, every living system is also an autopoietic system and every living system is minded. Even though, enactivism is most commonly classified as an anti-representational stance, most enactivists are open to some conceptualizations of representation that depend on *autopoiesis* and/or organism-environment interactions. On the other hand, proponents of REC reject any need for any kind of representation. There are content eliminativists like Chomsky and Egan that accept representations but only reject content which gives representations only a heuristic status (Chomsky, 1995; Egan, 2014).[43] Or there are also separate theories of mental content that usually appeals to information-theoretic (Dretske, 1981; Eliasmith, 2005; Lloyd, 1989; Rubert, 1999, 2008; Usher, 2001) and teleological approaches (Millikan, 1984, 2004, 2017; Neander, 2013, 2017; Shea, 2018; Stampe, 1977). These approaches have been viewed as most successful and promising theories of content in mainstream. However, they are criticized on the grounds that they all employ information-as-covariance and thus, as the consensus in the field acknowledges, cannot account for the content since covariance does not constitute content (Hutto & Myin, 2012). The main trouble arises only in the context of content that is semantic, normative and/or truth-bearing which has been taken to be crucial for mindedness. No kind of

[42] Autopoiesis is an important and heavy term that is central to enactivism. I cannot give a full review of relevant discussions here, but suffice it to say that it roughly means sum of self-organization and autonomy, in the sense that organism is autonomous and it can act despite the environment.
[43] Note that neither Chomsky nor Egan is a proponent of embodied cognitive science. Their positions rest on different arguments than embodied cognitive scientists.

covariance, no kind of correlation can give us semantics, truth or normativity by itself. The hard problem of content is pretty much this idea that natural phenomena, natural explanations are extensional whereas semantics, truth and normativity are intensional, hence causing them to be separate in a principled way. In other words, according to hard problem of content, content cannot be naturalized.[44] Most cognitive scientists or philosophers would argue against such a claim by saying that semantic content can be grounded in some natural explanation or natural concept. The various embodied cognition approaches or different theories of mental content mentioned above would, of course, claim that they have either resolved the issue or are on the verge of doing so. However, the problem lies in the fact that choosing one or more regularities found in nature and labeling them as "semantics," “content” or “normativity” is neither compelling nor sufficient to convince an impartial observer to accept this attribution. All these stances remain self-attributive and relatively arbitrary.

In this short review, I have outlined the contours of these complicated debates within cognitive scientce and to illustrate the primary issues and challenges along the major conceptual axes. What I have aimed to show thus far is that the two foundational functionalist views central to the establishment of cognitive science, i.e. multiple realizability and the autonomy of special sciences, have given rise to *the problem of mental causation* and *the hard problem of content*. Moreover, I have argued that these two problems cannot be addressed in isolation from one another for a conceptually complete and coherent cognitive science. Finally, I have shown that

[44] Even though I am on the same page with Hutto and Myin on the hard problem of content, I am skeptical towards both their distinction between basic minds and linguistic minds, and their claim that language and representationality are intrinsically related, issues both of which are beyond the scope of this thesis. See, for example Van den Herik (2014) for the latter issue. Also, as Villalobos and Silverman (2018) point out, asserting a semiotic relation in basic minds instead of representationality does not do much difference in terms of metaphysics of mind and the relevant philosophical problems.

different frameworks within cognitive science offer varied responses to these issues, but all of them work under functionalist assumptions, trying to provide cognitive science, assumed by them to be a special science, with a corresponding special ontology. Accordingly, nearly all schools of thought in cognitive science begin by establishing their metaphysics of mind and their approach to understanding it. Based on these primary answers, they then proceed to determine how cognitive science should be conducted. This is true not only for functionalism and the cognitivism but also for the post-cognitivist approaches grouped under the umbrella term embodied cognition.

At this point, we are gradually approaching the central argument of this thesis. My question is: can we not take a couple of steps back and think of a different trajectory for cognitive science, one that is entirely agnostic regarding the metaphysics of mind and the problems associated with such ontological stances? Building on our discussions of anti-realism and the New Mechanism in science, I will argue below that such a trajectory is indeed possible and even needed.

4.2 Naturalistic Coherence and sidestepping the mental

It has been a long journey, so let us take a moment to recall where we started. Beginning with the observation that most debated topics among different schools of thought in cognitive science is the reality of certain theoretical entities, we questioned whether prioritizing the reality of any entity or focusing on this question during theoretical conflicts is justified. At the conclusion of that discussion, we found an anti-realist stance to be inevitable, treating (scientific) concepts as non-real constructs. Within this framework, we conceptualized science as a social and cultural practice that produces these some particular subset of these constructs. However, just as every social and cultural practice has its own unique boundaries, rules, and forms, we acknowledged that science too has its distinct form, procedures, and limits. Given that there is no theory-neutral method to resolve disputes between theories and frameworks regarding the reality of entities, we explored in the second chapter how such conflicts might be addressed and sought to understand the practical boundaries of science. This led us to examine the New Mechanism approach in the philosophy of science. We noted that this approach can be interpreted in two ways: under the *broad interpretation*, New Mechanism describes the entirety of science and offers a normative framework for it; under the *narrow interpretation*, it provides a descriptive and normative account only for certain fields or aspects of science. For the purposes of this thesis, we argued that both interpretations are compatible. Those who adopt the broad interpretation might view the proposal here as saying, "Cognitive science must follow this model because science as a whole is such and such." In contrast, those who find the narrow interpretation more plausible might see it as "a coherent and potentially new way to address certain issues in cognitive science." Although many proponents of New Mechanism are realists, we briefly argued that New Mechanism is, in principle, compatible with anti-realism. We emphasized that for an anti-realist perspective, *coherence* is more important than *reality*. In this context, we distinguished between *empirical coherence* and *conceptual coherence*. We outlined that the central focus of this thesis is to introduce the concept of conceptual coherence and to explore its implications for cognitive science. Along the way, we took a detour to summarize the major debates and foundational issues in cognitive science that are relevant to our discussion. Now, having reviewed these key issues in cognitive science, we are finally in a position to turn our attention to the concept of conceptual coherence.

Let's remember the question of the present discussion: *Can we not take a couple of steps back and think of a different trajectory for cognitive science, one that is entirely agnostic regarding the metaphysics of mind and the problems associated with such ontological stances?* Given that a) we think that the autonomy of special sciences rests on controversial assumptions and thus is yet unconvincing; b) we differ from functionalism regarding the goals of science and what constitutes a scientific explanation; c) we argue that the *autonomy account* fails to provide a framework for how scientific progress occurs and leaves room for *ad hoc* justifications in theoretical disputes; d) we find the *hard problem of content* argument convincing and approach the naturalizability of the mental with skepticism; e) we do not take it granted that there is a categorical difference between sciences with different domains at different scales; f) we do not consider it necessary to view cognitive science as a special science that must first be provided with a special ontology: How can we proceed from here?

As discussed in the third chapter, according to New Mechanism, when we encounter a phenomenon in science, what we observe is essentially Event A and Event B. Let us assume that the available evidence and the scientific framework suggest that these two events are causally related. To explain this relationship, we can adopt two approaches. The first approach is to *zoom in* and examine the components of Event A and Event B, focusing on both their internal interactions and organization and the possible relations between the two events at this subcomponent level. This involves using relevant methods and techniques to formulate hypotheses based on existing theories and evidence, which are then tested. The second approach is to *zoom out* and examine how Event A and Event B, as components, relate to the larger events or mechanisms they are part of, or how those larger mechanisms interact with the subcomponents of the mechanisms underlying Event A and Event B systems. This, too, requires developing and testing hypotheses using similar methodologies and techniques. For New Mechanism, neither of these approaches is inherently superior to the other. Science progresses by moving in both directions: the questions we pose lead to new questions, and we navigate between scales or levels according to the demands of these inquiries. The resulting descriptions that emerge at different though interdependent levels or scales are what we refer to as *mechanistic explanations*.

In doing that, namely starting with Event A and Event B while our knowledge of their relations to higher and lower levels remains limited, in other words when these territories are *known-unknowns*, the explanations we develop, however localized, necessitate introducing new elements, processes, and events that were previously outside the scope of scientific knowledge. Notice that I am making no claims about whether these entities are observable or unobservable by any specific criterion. This distinction is not relevant to the current discussion. Within the natural progression of scientific inquiry, the new "things" we propose to explain various phenomena are inherently tied to ontology and incrementally expand our conceptual framework, both as a scientific community and individually. For instance, centuries ago, concepts like microbes, electrons, or quarks were absent from our discussions. Today, despite their content, definitions, and different "language games" that vary according to context, these entities have secured places in both scientific and, to some extent, folk ontologies. This is a natural outcome of the scientific progress. As late 20th-century philosophy of science and philosophy of language have shown, there is no categorical boundary between the empirical and the conceptual. Notice the contradiction regarding special sciences and their special ontologies. For a special science to exist in the first place, it must rely on a shaky, one whose origins are unclear, whose foundations may be unstable, whose reliability is uncertain, and upon which we cannot impose any criteria to determine if it misleads or wrongfoots us. Moreover, as a natural consequence of its autonomy, the explanatory criteria of this special science are tied to its special ontology, which, as noted earlier, will always appear successful and useful when evaluated by intra-theoretical standards. When faced with explanatory challenges, such as an inability to account for certain phenomena or the counterexamples to its explanations, these special disciplines, as Fodor himself acknowledges, tend to appeal to lower-level regularities to explain away such issues. As a result, the discipline, regardless of its particular content, is perpetually regarded as successful and sufficiently scientific, drifting according to trends without any clear criteria for progress. In contrast, the process outlined above, grounded in mechanistic explanation, suggests that there are no clear-cut divisions between levels of explanation. Instead, science progresses by filling in gaps and increasing the resolution of its understanding, just like an ever-sharpening image. As it does so, the localized explanations we develop, along with the new theoretical entities and constructs we propose, contribute continuously to the development of our conceptual frameworks.

At this point, two crucial questions arise for a researcher or a scientific community: 1) When dealing with *known-unknowns* and when theoretical and empirical needs require us to introduce a new construct, what should we consider? What are the relevant constraints? For instance, why, in our explanatory practices in science, say, when trying to explain a highly localized, low-level phenomenon in biochemistry, do we not invoke "spooky" concepts like spirits, ghosts, or fairies? Why do we instead even prefer to name these new constructs using sequences of letters and numbers that sometimes might seem entirely meaningless (e.g., H5N1, FOXP2, HD 209458, OGLE-2005-BLG-390Lb) 2) Beyond the methodological and epistemological constraints of a given scientific community at a particular time, are there any ontological, or more precisely *ontic*, constraints that compel us to stop at a certain scale, level of explanation, or point which will lead us to conclude, *"Beyond this, no further explanation is possible; we must stop here"?*

This is precisely where conceptual coherence becomes relevant. In light of the discussion above, let us define *conceptual coherence* as follows:

Conceptual Coherence (CC): When introducing a new concept in a given domain for any reason, the concept must, by its definition, be coherent with the conceptual framework of the domain and it must be consistent with the rest of it.

Conceptual Coherence can be applied to different conceptual frameworks and need not necessarily be limited to scientific or naturalistic[45] contexts. However, since this thesis focuses on naturalistic science, we will also define a specific case of conceptual coherence:

Naturalistic Coherence (NC): In a naturalistic domain and/or (a naturalistic scientific) discipline, when introducing a new concept for any reason, the concept must, by its definition, be naturalistic[46] and consistent with the rest of the domain.

Finally, in cases where this coherence fails which may either obvious in some cases or subtle in others, I will refer to the resulting mismatch and issues as

[45] Notice that I do not say 'natural science' as it has connotations regarding the distinction between the natural science and the social science. What I mean by 'naturalistic science' here is in fact the entirety of disciplines that are both scientific and naturalistic which may or may not cover social disciplines and/or some other so-called 'special sciences'.
[46] Here I will shortly refer to the Rorty's definition of naturalism: "I define naturalism as the claim that (a) there is no occupant of space-time that is not linked in a single web of causal relations to all other occupants and (b) that any explanation of the behavior of any such spatiotemporal object must consist in placing that object within that single web" (Rorty, 1998).

conceptual discordance.[47] For example, the mind/body problem is a clear case of such conceptual discordance.

Let us return to the questions. The answer to the first question lies in people's tendency to adhere to conceptual coherence (CC). However, since this tendency is implicit, it largely depends on the intuition of individuals and researchers, and as with any complex matter, it is unlikely to function without any issues or ambiguities. Therefore, even though such a tendency exists, we can still observe incompatibe instances. The answer to the second question is that there are no inherent ontic constraints in nature itself. Some may claim otherwise, but one wonders how they overcome the anti-realist arguments presented in the second chapter, as our anti-realist commitments render such ontic attributions implausible. However, in a conceptual sense, *conceptual discordance* may nevertheless arise. Such limitations and constraints, rather than stopping research or imposing restrictions on science, can and should only push us to reconsider and potentially revise the conceptual framework we employ in our scientific and philosophical inquiries. This is because, as long as the conceptual framework remains unchanged, the relevant issues will persist. The distinction between the *intensional* and *extension*al in the context of *the hard problem of content* serves as an example as well. If we cannot separate mental/normative concepts from semantics, or from notions of truth and falsehood in a non-trivial and non-ad-hoc way, which is something we actually cannot, then there is no point in insisting on using these concepts either. Instead, it would be more reasonable to propose a new conceptual framework or that is already compatible with the existing naturalistic framework. In other words, rather than insisting on

[47] In fact, the particular word I have in mind is Turkish, "*cızırdama*." It is at the intersection of *discordance, scratch* and *sizzling*.

naturalizing what is "unnatural" or attempting to force our way through these kinds of conceptual impasse, *we should expand the domain of what is already natural*.[48]

Both Conceptual Coherence (CC) and Naturalistic Coherence (NC) are, in fact, both descriptive and normative. However, as previously discussed, their normative aspect has remained implicit due to their intuitive nature. I aim to make this implicit tendency explicit and postulate it as a normative criterion: the *Naturalistic Coherence Criterion* (NCC). I propose this as a conceptual criterion both for evaluating theories or conceptual frameworks in science and as a constraint in proposing such theories or conceptual frameworks. By adopting this criterion, we can continue to conduct science without leading to conceptual discordance, sustain scientific progress, enhance resolution, and expand the naturalistic domain.

To better illustrate my point, let us briefly consider an example from the history of developmental biology.[49] The central question in developmental biology is: *How is it that cells with the same genetic base "know" what to do, evolving into different tissues with different functions to form distinct body parts, resulting in what we call "development"?* At the dawn of developmental biology in the late 19th and early 20th centuries, Hans Driesch reached an important finding while researching this question. He observed that when certain parts of sea urchins were cut away, the separated cells were capable of forming entirely new and healthy sea urchins. To explain this phenomenon, Driesch introduced a concept he called *entelechy*, inspired by Aristotle's notion of *telos*. According to Driesch, entelechy is an immaterial entity that cannot be reduced to physical processes or explained mechanistically. It directs the cells by providing them with the "knowledge" and "representation" of how and what to do in the process of development. In his view, the cells carried the

[48] Feyerabend (1963), in a two-page commentary, proposes something very similar in outline.
[49] This example and historical narrative are taken from Weber (2022).

"purposeful" knowledge of the organism's developmental process (hence *telos*). Entelechy was directly related to "life" itself and it had no connection or relationship to lower-level processes.

Notice that the perspective here is strikingly similar to the arguments developed by Putnam and Fodor regarding *multiple realizability* and *the autonomy of special sciences*. In fact, it seems plausible that if Driesch's debate had occurred a few decades later, it might have been framed as an argument in favor of treating developmental biology as a special science. Consequently, the necessity and inevitability of a special ontology that would include irreducible entities like entelechy for this field might have been proposed. However, contrary to the vitalist view advanced by Driesch at the time, the history of developmental biology has shown that these developmental processes could indeed be explained mechanistically, even though we had to wait until almost a century later. Initially, it was thought that morphogens[50] within cells directed developmental pathways in a straightforward, unidirectional manner. Over time, however, it became clear that morphogens, whose exact nature and functioning were not well understood at the time, actually interacted through complex feedback mechanisms with the cells and tissues they helped create. These interactions revealed that the developmental process was far from unidirectional. The vaguely defined space where these mechanisms were thought to occur was termed the *morphogenetic field*, inspired by physics. In the early 20th century, there were still vitalists who maintained that morphogenetic fields were irreducible and could not be explained mechanistically. However, Huxley and De Beer (1934) speculated that certain metabolic factors

[50] The concept *morphogen* was originally introduced by Alan Turing in the 1930s. However, even earlier, there were other terms used in a similarly vague sense to refer to "entities or molecules present in the cell, whose presence, structure, or various regularities influence or determine the developmental process." In this context, I am using the term morphogen in this general sense, albeit somewhat ahistorically.

might regulate these fields, pointing toward a potential mechanistic explanation. By the late 1990s, with developments in molecular biology, we learned how proteins such as Bone Morphogenetic Proteins (BMP) and Wnt influence gene expression through some specific receptors, thus playing crucial roles in developmental processes. Mechanistic explanations of these processes have enabled the development of mathematical models. Research in this field continues, and future discoveries may of course change our understanding of these processes. But the point I wish to emphasize, and which I believe is apparent here, is that scientific progress, regardless of how long it takes or which extra-scientific factors (e.g., technological, political, or economic) it depends upon, has been possible only through a mechanistic assumption and adherence to the Naturalistic Coherence Criterion (NCC). It is mechanistic because we consistently operate within constraints regarding inter-level relationships, which guide our investigations and prevent arbitrary drifts. It adheres to NCC because, in the process of advancing scientific explanations, we do not propose concepts that, by their definition, extend beyond the domain in question or introduce theoretical and philosophical gaps that hinder rather than advance scientific explanation. The concepts we propose are already defined or approached within a naturalistic framework. I suggest that the acceptance of vitalist views during that period, relying on assumptions about irreducible, immaterial, and mechanistically inexplicable phenomena could have undermined progress. In doing so, I leave it to the reader to consider what the history of developmental biology might have looked like if a view like *the autonomy of special sciences* had been widely accepted in the developmental biology of that era.

Let us revisit our two questions mentioned earlier. In the first question, we asked: When dealing with a low-level scientific question, why do we not invoke "spooky" entities, propose such entities, or feel the need for them? Recall my responses referring to Conceptual Coherence (CC) and Naturalistic Coherence (NC). Introducing a concept upon observing some *known-unkown* phenomena and defining it as irreducible and mechanistically unexplainable is a theoretical and philosophical choice. Conversely, as Huxley and De Beer did, considering what a potential mechanistic explanation might look like, proposing *how-possibly* explanations and continuing the research is also a theoretical and philosophical choice. We cannot know in advance whether such research will yield any meaningful results, or whether we will reach those results in fifty, one hundred, or even five hundred years. However, as far as I can see, this is the only way for science to exist, expand the naturalistic domain, and contribute to its explanatory power. [51] Specifically for cognitive science, I argue that following the assumptions of *the functionalist dogma* and trying to understand the mental via *multiple realizability* and *the autonomy of special sciences* is a choice similar to the first one above. Returning to the second question: When faced with a phenomenon we cannot yet explain, i.e. a known-unknown, do we have any ontic (as opposed to epistemological or methodological) reason to conclude that it cannot be explained or further connected to lower levels? Setting aside our anti-realist commitments, which we have already discussed, I argue that there is no such reason and that making such an assumption contradicts the very nature of scientific explanation and progress. In this example, I fail to grasp how accepting that entelechy is irreducible, mechanistically unexplainable, and, in a sense, "*ontologically special*" could possibly contribute to scientific progress.

[51] In saying this, I adopt a similar view with Feest (2010) who is also cited by Weber, albeit not entirely. The view in question suggests that concepts in science are research tools. I believe this naturally follows from my defense of anti-realism, my references to Wittgensteinian contextualism, and my view of science as a socio-cultural practice.

To even further the point being made here, we can appeal to a thought experiment similar to classic ones found in literature. As usual, consider a curious alien, or even a Martian scientist, who seeks to understand the behavior of Earth's living beings. Let us further imagine that in the scientific community of these Martians, and indeed across their entire species, including their everyday and social interactions, there are no mental or even teleological concepts. Assume that their language is entirely causal, and they conduct all their communication this way. While their science or everyday language may include concepts at different levels of abstraction, resolution, vagueness, or concreteness, they are completely unfamiliar with any intensional, semantic, or contentful concepts, neither their language nor their thinking incorporates such things. Now suppose further that these Martian scientists are centuries or even millennia ahead of us in terms of their mathematics, observational tools, and experimental methods. With this in mind, imagine that these scientists set out to study and explain the behavior of Earth's living beings, from the simplest to the most complex. As they come across various phenomena and mechanisms that require them to propose new concepts, what would compel these scientists, who have no prior knowledge of mental or teleological concepts to posit such notions? What would force them to invent these concepts, making it otherwise impossible for them to explain these phenomena? If we cannot demonstrate a categorical necessity, which I think we cannot, then the view I advocate here is a reasonable one.

Before closing this chapter, I want to address two potential misunderstandings. The first one is that the notion of coherence discussed in this thesis should not be conflated with epistemic coherentism, which is often the first association evoked by the term within analytic philosophy (for a concise review, see:

Olsson, 2022). Although there are some similarities in terms of structural outline and emphasis, the two issues serve different functions and belong to distinct domains. Coherentism in epistemology is a theory of justification, often presented as an alternative to foundationalism. It holds that a belief is justified if and only if it coheres with a system of mutually supportive beliefs. Epistemic coherentism typically rejects the idea of basic, self-justifying beliefs (as in foundationalism) and instead emphasizes holistic evaluation, mutual support, and the systemic integrity of doxastic networks. On the other hand, NCC is a meta-theoretical tool for constraining the introduction of new concepts and assessing the internal compatibility of theoretical frameworks. The purpose here is pragmatic and methodological, it is to guide the development of scientific theories that avoid conceptual discordance (such as the mind–body problem) by ensuring that their components integrate into a broader, naturalistic explanatory web. In other words, NCC is not doxastic. More clearly, in epistemology, coherence is about justification that is it tells us when a belief is warranted. Here, coherence is a (meta-)methodological criterion for conceptual legitimacy, if I may say so, it tells us when a scientific concept fits within the broader explanatory scheme. Lastly, traditional coherentism has been criticized for its relationship to truth in that coherence does not guarantee correspondence with reality. The proposal here sidesteps this problem by rejecting the realist assumptions altogether. Instead of claiming that coherence somehow relates to truth, I claim that truth is not the right criterion for science to begin with.

The second one is that what I am proposing here is *not* that we should stop practicing "philosophy of mind." Rather, I claim that when certain concepts in our current framework, potentially due to various reasons, do not align with one another, fail to progress cohesively, and create problems for a specific scientific discipline (in the context of this thesis, cognitive science), it may be a better approach to develop a new conceptual framework. This framework should be consistent with the rest of naturalistic science and, in a sense, with the "better-functioning" parts of it. Such concepts should also be proposed in light of various phenomena, considering local theoretical and empirical needs. To give an example from biology again, as seen in the literature, attempts to ground "life" or "living systems" might refer to DNA, far-from-equilibrium thermodynamics, or who knows what else (Mariscal, 2021). However, neither DNA nor far-from-equilibrium thermodynamics was proposed explicitly to "ground" or explain "life" or "living systems," nor do they gain their explanatory justification from this question. What is more, biology as a discipline existed long before any of these were introduced and it has progressed for centuries without a properly grounded concept of "life." (For the argument that grounding the concept of "life" is unnecessary for biology, see: (Tirard et al. 2010; Machery 2012; Cleland & Chyba 2002). Similarly, in cognitive science, to ground "mind" or to define the field based on a particular metaphysics of mind is neither necessary nor desirable. Cognitive scientists should not be required to frame their field around such metaphysical commitments (For similar but somewhat different views: Allen, 2017; Rubert, 2013; Villalobos & Palacios, 2021). Of course, philosophers of mind, just as philosophers of biology and life, may learn from cognitive science and debate whether concepts and findings from cognitive science can be useful for their philosophical discussions. They may use cognitive science to engage in science-informed philosophy. However, these questions should not form the core theoretical distinctions in cognitive science itself. In summary, the mind-body problem should be a concern only for the philosopher of mind, not the cognitive scientist. The cognitive scientist should be able to conduct their work solely within the framework of *res extensa*.

To do this, a critical question remains: what kind of conceptual framework could be proposed for cognitive science that would allow us to address phenomena traditionally considered mental without necessitating the attribution of any mental concepts to nature itself? More importantly, what conditions must such a framework satisfy, what components must it include, and what must it exclude? These are very important questions, and I believe that further research in this area will be definitive for the future of cognitive science. The answers to these questions, I suppose, could provide a way to unify the various conceptual traditions within cognitive science without relying on the metaphysical debates that currently divide them and establish a form of epistemological ground-zero for these schools and their associated research.

So far, I have attempted to summarize my claims. In the upcoming conclusion, I will briefly review the discussion thus far and clarify what I am *not* claiming, in order to preempt potential misunderstandings.

CHAPTER 5

CONCLUSION AND FURTHER DIRECTIONS

In this thesis, I tried to adopt a synthetic approach. By this, I mean an effort to integrate philosophical stances that the researcher independently advocate, allowing these views to constrain the space of acceptable answers to a given problem and, as a result, seeking and, if possible, presenting a holistic and unified solution. For instance, at the outset, I observed that theoretical debates in cognitive science often focus on the *reality* of various entities, which led me to engage with the debate on scientific realism. This discussion, for independent reasons, pushed us toward a scientific and metaphysical anti-realist position. Similarly, I outlined the new mechanistic philosophy of science, which I endorse independently of the realism debate. I concluded that there is no categorical incompatibility between the two. I included that discussion to provide a coherent and unified meta-constraint for the philosophical stance we adopt toward a particular scientific discipline, here cognitive science in particular. In light of these, I considered the relevant implications for cognitive science. In doing so, I gave attention to respect the meta-constraints established through our previous discussion. As a result, I have attempted to present as comprehensive and consistent a framework as possible.

In this context, I argued that anti-realism is inevitable when considering any non-trivial definition of mind-independent reality and/or truth. I contended that the arguments put forward in favor of realism are inadequate to justify it. I critiqued Putnam’s early notion of *natural kind semantics* and pointed out that approaches like *entity realism*, *structural realism*, or *explanationist realism*, which often appeal to a certain kind of historical continuity, can only merely highlight a continuation of historical consensus and tell us nothing substantive about mind-independence. Similarly, I argued that views like *perspectival realism*, which reject the objectivist assumptions of realism, are ad hoc, as it remains unclear what purpose a non-objectivist conceptualization of reality would serve beyond attempting to resolve the realism debate itself. In other words, I suggested that abandoning the criterion of mind-independence requires a non-ad-hoc justification independent of this debate, and that such a move trivializes realism. In contrast, I emphasized that we cannot escape *Quinian holism* and socio-historical and cultural theses regarding theory change in science, which remain relevant. Additionally, unlike some philosophers of science, I argued that the distinction between observable and unobservable phenomena is untenable. Furthermore, I maintained that the anti-realist arguments in question also apply to folk ontologies, making it arbitrary to adopt a scientific anti-realist stance while refraining from metaphysical anti-realism. Consequently, I defended a form of *radical anti-realism*.

Then, considering that the anti-realism left us without any answer to the questions of reality in the theoretical and philosophical conflicts of cognitive science, I supposed that looking at how science works might give us a lead. In this context, New Mechanistic philosophy became relevant. I outlined the framework of new mechanistic philosophy, focusing on mechanistic inter-level integration and demonstrating how it overcomes many of the limitations traditionally attributed to classical mechanistic approaches. I also addressed objections that new mechanistic philosophy trivializes the concept of mechanisms, arguing that, unlike causal-functional analysis, it provides a cohesive framework on science and this is its important and characteristic aspect. Whether one chooses to call it mechanistic or something else is only a *stylistic* dispute. Finally, while many mechanists are realists, I noted that mechanistic philosophy is not, in principle, incompatible with anti-realism. I later concluded that *coherence* could serve as a criterion instead of mind-independent reality. However, I argued that *empirical coherence* (EC) alone is not enough and that a criterion of *conceptual coherence* (CC) is also necessary.

I claimed that schools in cognitive science are fundamentally divided based on their responses to the metaphysics of mind. I argued that this division stems from functionalism, reviewed its connections to multiple realizability, the mental causation problem, the autonomy of special sciences, and the hard problem of content. I explained the relationships between these issues, showing how they reinforce one another, all rooted in functionalism. Specifically, I argued that the autonomy of special sciences leads cognitive science to be treated as a special science, hence requiring a distinct and autonomous ontology which seem to have received considerable support to this point. I then highlighted how the very nature of the so-called autonomy of special sciences gives rise to certain problems, particularly those regarding the concept of scientific progress. I argued that the distinction between special and natural sciences on the grounds how laws work differently in each of them is problematic and ungrounded. Accordingly, I further claimed that to *a priori* assume some scientific constructs are irreducible and unexplainable, and even unrelatable in any way to lower levels is at odds with the goals of scientific inquiry and scientific explanation. In response, I introduced the concepts of *CC* and *NC*. While conceptual coherence can be applied to any conceptual framework, naturalistic coherence is specifically applicable to naturalistic domain. I suggested that, when thinking in terms of mechanistic levels, we can use a variety of constructs according to scale, phenomena, levels of abstraction, or degrees of concreteness/vagueness depending on the phenomena and explanations at hand.

However, I argued that choosing a construct which is already defined “irreducible,” “unexplainable,” or in a way incoherent with the existing naturalistic framework over an alternative is ultimately a theoretical choice of the particular theorist or scientific community. This, as such, might severely hinder genuine explanatory efforts. To illustrate this, I provided an example from the history of developmental biology and a thought experiment. Lastly, I noted that people are intuitively inclined to attend to coherence, but like all intuitions, this can sometimes fail. To prevent this as much as we can, I suggested making this implicit intuition explicit. In this context, I proposed the *Natural Coherence Criterion (NCC)* as a constraint for the appraisal of various constructs and conceptual frameworks in science or for proposing new ones.

Misunderstandings are inevitable given that I might have undertaken a task far larger than my reach in the attempt to address these complicated issues within the scope of a master’s thesis. Therefore, I should now attempt to answer the question of what have I not claimed as opposed to what I have claimed?

First and foremost, I have not claimed that functionalism is wrong. On this matter, I am agnostic. Functionalism may very well be correct, but that is an issue for the philosophers of mind. My argument is simply that cognitive science, as a scientific discipline, does not need to rely on any particular metaphysics of mind to establish itself or ground its theories. In principle, one could even be a functionalist and still support the thesis presented here. There is no inherent contradiction in doing so. Although admittedly I would not understand why someone would even bother to do that.

I have not argued that concepts like *content*, *intentionality*, *normativity*, or *mindedness* cannot be grounded and should no longer be discussed even though I because of the reasons I have outlined above I am skeptical towards such attempts. On the other hand, philosophers of mind are, of course, welcome to continue debating these topics, and in doing so, they can very well draw upon a non-mentalistic cognitive science just as they draw upon a non-vitalist biology in trying to answer the question of "what is life."

I do not claim that anti-realism entails New Mechanism, nor that New Mechanism entails anti-realism. I defended each independently, giving distinct justifications for both, and argued that they are compatible with one another. I further suggest that bringing these two together, particularly when considering the contextualist stance of the kind of anti-realism defended here, can provide a foundation for a normative criterion to guide the creation and appraisal of new concepts in science.

I think that the framework presented here differs significantly from most, perhaps all, forms of reductionism. The primary reason for this is that our radical anti-realism prevents us from considering any explanatory level as more "real" or “fundamental” than another. In this framework, atoms or neurons are no more "real" or “fundamental” than human behaviors or vice versa. All levels are interdependent fictions. However, this topic warrants its own separate treatment.

The proposal here differs from most, perhaps all, forms of eliminativism. Unlike common forms of eliminativism, it does not make claims about the existence or non-existence of particular entities. Rather, it argues that certain constructs or purported entities should be eliminated from the naturalistic domain if they are incoherent with it. Furthermore, due to the commitment to anti-realism and contextualism, "existence vs. non-existence" questions become meaningless. When these questions are meaningless, it is possible that many and even most conceptual problems in science could be addressed through *conceptual* and *linguistic hygiene*. This is because, under this framework, only the concepts and terms we use as tools to "work with" remain relevant. However, this issue also deserves a separate treatment.

I have not claimed that one must be an anti-realist to defend the *Naturalistic Coherence Criterion (NCC)*. In principle, someone could defend NCC while being a realist of some sort. However, in that case, they would need to either "naturalize" or reduce everything they consider real. This is the common practice currently observed in the field. While it is difficult to make a definitive statement, my pessimism on this strategy is evident. Notice also that for this strategy, even for the strong emergentist, there is a necessity to show how and why the (strongly) emerged new entities or properties correspond to the ones we take to be real in our folk-ontologies, e.g. mind, life, memory, etc.

Similarly, I have not claimed that one must adopt Mew Mechanism or a mechanistic framework of any kind to defend NCC. In principle, someone could oppose mechanism and still defend NCC. However, in doing so, assuming that they are not anti-realist, it would lead to functionalism and hence to facing all the problems that functionalism entails. For instance, they would face nearly all the issues we have found problematic about *autonomy of special sciences*. Without mechanistic levels and their interrelations, we lack constraints for proposing, connecting, explaining, or grounding concepts. In such a case, anything goes. This, as noted above, ultimately leads to the kinds of crises related to scientific progress that we have already highlighted. But an anti-realist and non-mechanistic defense of NCC is possible. Though probably it would still lead to problems related to scientific progress and the formulation of non-ad-hoc criteria for theoretical and philosophical conflicts between conceptual frameworks.

I have not claimed that defending the NCC requires accepting the broad interpretation of New Mechanism, i.e. the view that all of science is mechanistic. For the purposes of the present discussion, it is adequate to accept that only relevant parts of science, say, biology, cognitive science, neuroscience, are mechanistic. I am also not claiming that it is currently possible to explain or formulate all the phenomena that cognitive science considers and formulates as "mental" in a non-mentalistic way. Whether this is possible cannot be known a priori. The only way to determine this is to follow the progress of a non-mentalistic cognitive science with cautious and deliberate steps.

From where can we begin to develop such a conceptual framework then? What are the relevant clues at hand? How might further research proceed, and how could the foundations for a non-mentalistic cognitive science be established? Two potential approaches come to mind. The first is to start with the brain, following something in the vicinity of Buzsáki (2019). The second is to begin with evolution, drawing on something akin to Cisek's *phylogenetic refinement* (Cisek, 2020). Note that, though, neither of these perspectives explicitly aims to offer a non-mentalistic ontology and, in fact, they both accept certain mentalistic concepts. However, both are skeptical of some of the categorizations inherited from folk psychology and target them through two distinct routes. Of course, there is no space for a thorough review of these views in this chapter or even in this thesis. However, the first challenge faced by approaches like Buzsáki's, which start with the brain, is simply the vast complexity of the brain itself. The brain is incredibly complex so much so that without a solid conceptual framework, we can find correlations for virtually anything we choose to investigate (Meijer, 2021). Thus, it is unclear where to even begin. What is more, the purpose of a conceptual framework is precisely to provide this kind of constraint. Moreover, our experimental studies of the brain must rely on behavioral taxonomies. Yet, to my knowledge, there is no existing non-mentalistic behavioral taxonomy. (For similar critiques of Buzsáki, see: Poeppel & Adolfi, 2020.) Note that, that means without using mentalistic concepts, there is currently no foundation to categorize brain activity. In this context, relying on behavioral taxonomies raises two additional issues: First, behavior is as complex as the brain. Without a framework like folk psychology, it is unclear how we could carve up the behavioral domain, particularly at the human scale, in a way that would both provide a non-mentalistic ontology and avoid arbitrary and ad-hoc frameworks. Second, complex behaviors are observed in organisms without brains (Gomez-Marin, 2021). This undermines the motivation for starting cognitive science with the brain, calling into question the plausibility of a brain-centric approach.

On the other hand, this final point might suggest a novel and interesting direction. We could begin constructing our conceptual framework and taxonomy by focusing on simpler phenomena, using the taxonomies we develop for these behavioral phenomena as clues and constraints for approaching more complex phenomena, building incrementally. This approach parallels Cisek's strategy of relying on evolutionary and comparative studies for classification of functions and mechanisms in neuroanatomu rather than experimental research grounded in folk psychology. Admittedly, as I previously noted, Cisek is not categorically opposed to mentalistic concepts and actively employs them. Nevertheless, his strategy could provide important insights for developing a non-mentalistic cognitive science.

Notice the parallelisms between this view and the approaches like enactivism (Di Paolo & Thompson, 2014; Froese & Di Paolo, 2011; Gallagher, 2017; Noë, 2021; Steward, Gapenne & Di Paolo, 2010; Varela, Thıompson & Rosch, 1991; Ward, Silverman & Villalobos, 2017), autopoietic theory (Maturana & Varela, 1973/1980; Villalobos, 2015; Villalobos, 2022; Villalobos & Silverman, 2018; Villalobos & Ward, 2015), ecological psychology (Gibson, 1979; Lobo, Heras-Escribano & Travieso, 2018; Segundo-Ortin & Raja, 2024), the Free Energy Principle (FEP) (Friston, 2010; Friston, Fortier & Friedman, 2018; Parr, Pezzulo & Friston, 2022), and Technological Approach to Mind Everywhere (TAME) (Levin, 2022) in advocating a form of *life-mind continuity*[52] even if significant differences remain. The question of how the framework proposed here overlaps or diverges from these schools of thought in cognitive science certainly warrants a dedicated treatment. For the purposes of this discussion, suffice it to say that starting with the behavior of so-called simpler organisms or, more precisely, organism-environment interactions and progressively developing our conceptual framework toward more complex phenomena appears to be an epistemological ground-zero that most of us can agree upon. Approaches like enactivism, TAME, FEP, and ecological psychology are far from being non-mentalistic. However, autopoietic theory, with its anti-realist commitments rooted in second-order cybernetics and its proximity to a mechanistic framework (Lee, 2023), represents the closest view to the framework proposed here. That being said, significant differences remain. Most fundamentally, in autopoietic theory, mind is treated as a *conventional kind* as opposed to a natural kind (Villalobos & Palacios, 2021). On this point, I remain entirely agnostic in the present thesis, as I make no claims about the nature of the mind. Rejecting such a

[52] Of course, within the framework discussed here, referring to the proposed continuity as *life-mind* is no longer possible for obvious reasons. Terms like *evolutionary scaffolding*, *evolutionary coherence* or *evolutionary continuity* seem more fitting.

stance also complicates autopoietic theory's proposal to use mental concepts instrumentally, making it problematic within the framework I have outlined here.

Of course, the program I just outlined covers actually a vast amount of scientific area that will inevitably bring with it numerous questions, challenges, and puzzles. While some speculations about potential issues and how they might be addressed should have become obvious after my discussions in this thesis, I cannot explore them here due to space limits. Moreover, it is impossible at this stage to see beforehand where such a research program might lead, or how many researchers, master's theses, doctoral dissertations, articles, books, institutions, and years or decades it will require to achieve its goals. Nevertheless, this is science. We have always owed our success in science to its hard ways. Once again, we should bite the bullet and take the hard way.

I want to end this thesis with a quote by Lichtenberg, who reminds us something we all seem to have forgotten long ago: "There is no greater impediment to progress in the sciences than the desire to see it take place too quickly."

REFERENCES

Alai, M. (2017). Resisting the historical objections to realism: Is Doppelt's a viable solution? *Synthese, 194*(9), 3267–3290. https://doi.org/10.1007/s11229-016-1087-z

Allen, C. (2017). On (not) defining cognition. *Synthese*, *194*(11), 4233–4249.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press.

Baggs, E. (2015). A Radical Empiricist Theory of Speaking: Linguistic Meaning Without Conventions. *Ecological Psychology*, *27*(3), 251–264. https://doi.org/10.1080/10407413.2015.1068655

Barnes, B., Bloor, D., & Henry, J. (1996). *Scientific knowledge: A sociological analysis*. University of Chicago Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577–660. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology, 22*(5), 543–564. https://doi.org/10.1080/09515080903238948

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science, 78*(4), 533–557. https://doi.org/10.1086/661513

Bechtel, W. (2022). Levels in biological organisms: Hierarchy of production mechanisms, heterarchy of control mechanisms. *The Monist, 105*(2), 156–174. https://doi.org/10.1093/monist/onab029

Bechtel, W., & Abrahamsen, A. A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in history and philosophy of science*, *41*(3), 321–333. https://doi.org/10.1016/j.shpsa.2010.07.003

Bickhard, M. H., & Richie, D. M. (1983). *On the nature of representation: A case study of James Gibson's theory of perception*. Praeger.

Bloor, D. (1976). *Knowledge and social imagery*. Routledge & Kegan Paul.

Bourdieu, P. (2004). *Science of science and reflexivity* (R. Nice, Ed. & Trans.). University of Chicago Press.

Boyd, R. (1973). *Realism, underdetermination, and a causal theory of evidence*. *Noûs, 7*(1), 1–12. https://doi.org/10.2307/2216179

Boyd, R. N. (1983). *On the Current Status of the Issue of Scientific Realism*. Erkenntnis (1975-), 19(1/3), 45–90. http://www.jstor.org/stable/20010835

Brentano, F. (1995). *Psychology from an empirical standpoint*. London: Routledge. (Original work published 1874). ISBN 0-415-10661-3.

Brette, R. (2018). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*, e215. https://doi.org/10.1017/S0140525X19000049

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press. https://doi.org/10.1093/oso/9780190905385.001.0001

Carruthers, P. (2006). The case for massively modular models of mind. In R. Stainton (Ed.), *Contemporary debates in cognitive science*. Wiley-Blackwell.

Chakravartty, A. (2007). *A metaphysics for scientific realism: Knowing the unobservable*. Cambridge University Press.

Chakravartty, A. (2017). Scientific realism. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Summer 2017 ed.). Stanford University. https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/

Chalmers, D. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, *12*(4), 323–357.

Chang, H. (2003). Preservative realism and its discontents: Revisiting caloric. *Philosophy of Science*, *70*(5), 902–912. https://doi.org/10.1086/377376

Chemero, A. (2009). *Radical embodied cognitive science*. MIT Press.

Chomsky, N. (1995). Language and nature. *Mind*, *104*, 1–61.

Churchland, P. M. (1985). The ontological status of observables: In praise of the superempirical virtues. In P. M. Churchland & C. A. Hooker (Eds.), *Images of science: Essays on realism and empiricism* (pp. 35–47). University of Chicago Press.

Cleland, C., & Chyba, C. (2002). Defining “life”. *Origins of Life and Evolution of the Biosphere*, *32*(4), 387–393.

Cisek P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, perception & psychophysics*, *81*(7), 2265–2287. https://doi.org/10.3758/s13414-019-01760-1

Collins, H. M. (Ed.). (1981). *Knowledge and controversy*. *Social Studies of Science, 11*(1).

Collins, H., & Pinch, T. (1993). *The Golem: What everyone should know about science*. Cambridge University Press.

Colombo, M., Hartmann, S., & van Iersel, R. (2015). Models, mechanisms, and coherence. *British Journal for the Philosophy of Science, 66*(1), 181–212.

Crane, T. (1998). Intentionality as the mark of the mental. In *Intentionality*. *The Routledge Encyclopedia of Philosophy*. Taylor and Francis.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press. https://doi.org/10.1093/acprof:oso/9780199299317.001.0001

Craver, C. F. (2013). Functions and mechanisms: A perspectivalist view. In P. Huneman (Ed.), *Functions: Selection and mechanisms* (pp. 133–158). Springer.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Craver, C., & Tabery, J. (2023). Mechanisms in science. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023 edition). Stanford University. https://plato.stanford.edu/archives/fall2023/entries/science-mechanisms/

Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfield relations, and anomaly resolution* (Cambridge Studies in Philosophy and Biology). Cambridge University Press. https://doi.org/10.1017/CBO9780511498442

Dennett, D. C. (1991). *Consciousness explained*. Penguin Books.

Devitt, M. (2007). Scientific realism. In F. Jackson & M. Smith (Eds.), *The Oxford handbook of contemporary philosophy*. Oxford University Press.

Di Paolo, E. A., & Thompson, E. (2014). The enactive approach. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 68–78). Routledge.

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.

Doppelt, G. (2007). Reconstructing scientific realism to rebut the pessimistic meta-induction. *Philosophy of Science, 74*(1), 96–118. https://doi.org/10.1086/520685

Duhem, P. (1991). *The Aim and Structure of Physical Theory* (P. P. Wiener, Trans.). Princeton University Press. (Original work published 1906)

Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science*. Harvard University Press.

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

Erdin, H. O. (2020). Appraisal of certain methodologies in cognitive science based on Lakatos's methodology of scientific research programmes. *Synthese, 199*(Suppl 1), 89–112. https://doi.org/10.1007/s11229-020-02654-3

Fahrbach, L. (2011). How the growth of science ends theory change. *Synthese, 180*(2), 139–155. https://doi.org/10.1007/s11229-009-9602-0

Feest, U. (2010). Concepts as tools in the experimental generation of knowledge in cognitive neuropsychology. *Spontaneous Generations, 4*, 173–190.

Feigl, H. (1958). The 'mental' and the 'physical'. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories and the mind-body problem* (Vol. 2). Minnesota Studies in the Philosophy of Science. University of Minnesota Press. (Reprinted with a postscript in Feigl, 1967)

Feyerabend, P. K. (1963). Comment: Mental events and the brain. *Journal of Philosophy, 60*(11), 295–296.

Feyerabend, P. K. (1975). *Against Method*. New Left Books.
Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115. http://www.jstor.org/stable/20114958

Fodor, J. (1975). *The language of thought*. Thomas Y. Crowell.

Fodor, J. (1981). *Representations*. MIT Press.

Fodor, J. (1983). *The modularity of mind*. MIT Press.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press. (Edited by M. A. Boden)

Fodor, J. (1997). Special sciences: Still autonomous after all these years. *Philosophical Perspectives, 11*, 149–163. http://www.jstor.org/stable/2216128

Fodor, J. (2008). *LOT2: The language of thought revisited*. Clarendon Press.

Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's "ecological approach." *Cognition, 9*(2), 139–196. https://doi.org/10.1016/0010-0277(81)90009-3

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, *23*(11–12), 11–39.

Frigg, R., & Votsis, I. (2011). Everything you always wanted to know about structural realism but were afraid to ask. *European Journal for Philosophy of Science, 1*(2), 227–276. https://doi.org/10.1007/s13194-011-0025-7

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

Friston, K. J., Fortier, M., & Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness.

Froese, T., & Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmatics and Cognition*, *19*(1), 1–36.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press.

Garson, J. (2013). The functional sense of mechanism. *Philosophy of Science, 80*(3), 317–333. https://doi.org/10.1086/671173

Gauker, C. (1994). Thinking out loud. Princeton University Press.

Gendler, T., & Hawthorne, J. (Eds.). (2002). *Conceivability and possibility*. Oxford University Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science, 71*(5), 742–752.

Giere, R. N. (2006). *Scientific perspectivism*. University of Chicago Press.

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology, 20*(6), 733–743. https://doi.org/10.1177/0959354310378184

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences, 20*(1), 1–19. https://doi.org/10.1017/S0140525X97000010

Glenberg, A. M., & Swanson, N. G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(1), 3. https://doi.org/10.1037/0278-7393.12.1.3

Glennan, S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 443–464. https://doi.org/10.1016/j.shpsc.2005.03.011

Glennan, S. (2009). Productivity, relevance and natural selection. *Biology & Philosophy, 24*(3), 325–339. https://doi.org/10.1007/s10539-008-9137-7

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Goldman, A. I. (2013). The bodily formats approach to embodied cognition. In U. Kriegel (Ed.), *Current controversies in philosophy of mind* (pp. 91–108). Routledge.

Gomez-Marin, A. (2021). If the brain is the answer, what was the question? *Current Biology, 31*(5). https://doi.org/10.1016/j.cub.2021.01.101

Gomila, A. (2011). The language of thought: Still a game in town? [Review of *LOT2. The language of thought revisited*, by J. Fodor]. *Teorema: Revista Internacional de Filosofía, 30*(1), 145–155. http://www.jstor.org/stable/43047509

Hacking, I. (1982). Experimentation and scientific realism. Philosophical Topics, 13(1), 71–87. https://doi.org/10.5840/philtopics19821314

Hacking, I. (1983). Representing and intervening. Cambridge University Press.

Hacking, I. (1985). Do we see through a microscope? In P. M. Churchland & C. A. Hooker (Eds.), Images of science: Essays on realism and empiricism (pp. 132–152). University of Chicago Press.

Haefner, R. M., Beck, J., Savin, C., Salmasi, M., & Pitkow, X. (2024). How does the brain compute with probabilities? *arXiv preprint* arXiv:2409.02709. https://arxiv.org/abs/2409.02709

Hallett, G. L. (1991). *Essentialism: A Wittgensteinian critique*. State University of New York Press.

Hanfling, O. (1984). Scientific realism and ordinary usage. *Philosophical Investigations, 7*(3), 187–205.

Hanfling, O. (2000). Philosophy and ordinary language. Routledge.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*, 335–346.

Hesse, M. B. (1976). Truth and growth of knowledge. In F. Suppe & P. D. Asquith (Eds.), *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 2, pp. 261–280). Philosophy of Science Association. https://doi.org/10.1086/psaprocbienmeetp.1976.2.192385

Hommel, B., & Colzato, L. S. (2015). Learning from history: The need for a synthetic approach to human cognition. *Frontiers in Psychology, 6*, 1435. https://doi.org/10.3389/fpsyg.2015.01435

Howson, C. (2000). Hume's Problem: Induction and the Justification of Belief, Oxford University Press, New York.

Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content*. MIT Press.

Huxley, J. S., & De Beer, G. R. (1934). *The elements of experimental embryology*. Cambridge University Press.

Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science, 3*(2), 438–444.

Kim, J. (1989). The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association, 63*(3), 31–47. https://doi.org/10.2307/3130081

Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research, 52*(1), 1–26.
Kitcher, P. (1993). The advancement of science: Science without legend, objectivity without illusions. Oxford University Press.

Knorr-Cetina, K. (1981). *The manufacture of knowledge*. Pergamon Press.

Krickel, B. (2014). *The metaphysics of mechanism* (Doctoral dissertation). Humboldt-Universität zu Berlin.

Kripke, S. (1980). Naming and Necessity. Oxford: Wiley-Blackwell.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed., with postscript). University of Chicago Press. (Original work published 1962)

Laudan, L. (1977). Progress and its problems. University of California Press.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science, 48*(1), 19–49. https://doi.org/10.1086/288975

Laudan, L. (1984). *Science and values: The aims of science and their role in scientific debate*. University of California Press.

Latour, B. (1990). Postmodern? No, simply amodern! Steps towards an anthropology of science. *Studies in History and Philosophy of Science Part A*, *21*(1), 145–171. https://doi.org/10.1016/0039-3681(90)90018-4

Latour, B., & Woolgar, S. (1979). *Laboratory life: The construction of scientific facts*. Sage Publications.

Lee, J. (2023). Enactivism meets mechanism: Tensions & congruities in cognitive science. *Minds and Machines*, *33*(1), 153–184. https://doi.org/10.1007/s11023-022-09618-6

Levin, M. (2022). Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience, 16*, 768201. https://doi.org/10.3389/fnsys.2022.768201

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science*, *80*(2), 241–261.

Lewis, P. J. (2001). Why the pessimistic induction is a fallacy. *Synthese, 129*(3), 371–380. https://doi.org/10.1023/A:1013139410613

Leuridan, B. (2010). Can mechanisms really replace laws of nature? *Philosophy of Science, 77*(3), 317–340. https://doi.org/10.1086/652959

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology, 9*, 2228. https://doi.org/10.3389/fpsyg.2018.02228

Loomis, E. (2017). Necessity and apriority. In H.-J. Glock & J. Hyman (Eds.), *A companion to Wittgenstein* (pp. 346–358). John Wiley & Sons.

Lloyd, D. (1989). *Simple minds*. MIT Press.

Lupyan, G. (2023). Is language-of-thought the best game in the town we live? *Behavioral and Brain Sciences, 46*, e281. https://doi.org/10.1017/S0140525X23001814

Mach, E. (2013). The Science of Mechanics: A Critical and Historical Exposition of its Principles (T. J. McCormack, Trans.; 6th ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107338401

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25. https://doi.org/10.1086/392759

Machery, E. (2012). Why I stopped worrying about the definition of life… and why you should as well. *Synthese, 185*, 145–164.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience, 10*, 94. https://doi.org/10.3389/fncom.2016.00094

Mariscal, C. (2021). Life. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021 edition). Stanford University. https://plato.stanford.edu/archives/win2021/entries/life/

Massimi, M. (2018). Four kinds of perspectival truth. *Philosophy and Phenomenological Research, 96*(2), 342–359.

Massimi, M. (2022). *Perspectival realism*. Oxford University Press. https://doi.org/10.1093/oso/9780197555620.001.0001 (accessed April 2, 2025).

Maturana, H. R., & Varela, F. J. (1973/1980). *Autopoiesis and cognition: The realization of the living*. Springer

Meijer, G. (2021). Neurons in the mouse brain correlate with cryptocurrency price: A cautionary tale. *Peer Community Journal, 1*, Article e29. https://doi.org/10.24072/pcjournal.30

Miłkowski, M. (2011). Beyond formal structure: A mechanistic perspective on computation and implementation. *Journal of Cognitive Science, 12*(4), 359–379.

Miłkowski, M. (2016). A mechanistic account of computational explanation in cognitive science and computational neuroscience. In V. C. Müller (Ed.), *Computing and philosophy* (Synthese Library, Vol. 375). Springer. https://doi.org/10.1007/978-3-319-23291-1_13

Miłkowski, M. (2018). From computer metaphor to computational modeling: The evolution of computationalism. *Minds & Machines, 28*(4), 515–541. https://doi.org/10.1007/s11023-018-9468-3

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences, 7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Miller, B. (2016). What is Hacking's argument for entity realism? Synthese, 193(3), 991–1006. https://doi.org/10.1007/s11229-015-0789-y

Millikan, R. (1984). *Language, thought, and other biological categories*. MIT Press.

Millikan, R. (2004). *Varieties of meaning*. MIT Press.

Millikan, R. (2017). *Beyond concepts*. Oxford University Press.

Mitchell, K. J. (2023). *Free agents: How evolution gave us free will*. Princeton University Press.

Mizrahi, M. (2013). The pessimistic induction: A bad argument gone too far. *Synthese, 190*(15), 3209–3226. https://doi.org/10.1007/s11229-012-0138-3

Mizrahi, M. (2016). The history of science as a graveyard of theories: A philosophers' myth? *International Studies in Philosophy of Science, 30*(3), 263–278. https://doi.org/10.1080/02698595.2017.1316113

Monton, B., & van Fraassen, B. (2003). Constructive empiricism and modal nominalism. *British Journal for the Philosophy of Science, 54*, 405–422.

Moreno, A., & Etxeberria, A. (2005). Agency in natural and artificial systems. *Artificial Life, 11*, 161–175. https://doi.org/10.1162/1064546053278919

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In *Distinctiveness and memory* (pp. 27–46). Oxford University Press.Machamer, P. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science, 18*(1), 27–39. https://doi.org/10.1080/02698590412331289242

Neander, K. (2013). Toward an informational teleosemantics. In D. Ryder, J. Kingsbury, & K. Williford (Eds.), *Millikan and her critics* (pp. 21–36). Wiley-Blackwell.

Neander, K. (2017). *A mark of the mental*. MIT Press.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*(2), 135–183.

Nicholson, D. J. (2018). Reconceptualizing the organism: From complex machine to flowing stream. In D. J. Nicholson & J. Dupré (Eds.), *Everything flows: Towards a processual philosophy of biology* (Chapter 7). Oxford University Press. https://doi.org/10.1093/oso/9780198779636.003.0007

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

Olsson, E. J. (2022). *Coherentism*. Cambridge University Press.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Pickering, A. (1984). *Constructing quarks: A sociological history of particle physics*. University of Chicago Press.

Place, U. T. (1954). The concept of heed. *British Journal of Psychology, 45*, 243–255.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology, 47*, 44–50.

Poeppel, D., & Adolfi, F. (2020). Against the epistemological primacy of the hardware: The brain from inside out, turned upside down. *eNeuro, 7*(4), ENEURO.0215-20.2020. https://doi.org/10.1523/ENEURO.0215-20.2020

Poincaré, H. (2014). The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method (G. B. Halsted, Trans.). Cambridge University Press.

Putnam, H. (1960). Minds and machines. Reprinted in H. Putnam (1975), *Mind, language and reality: Philosophical papers, Volume 2* (pp. 362–385). Cambridge University Press.

Putnam, H. (1962). What theories are not. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science: Proceedings of the 1960 International Congress* (pp. 240–251). Stanford University Press.

Putnam, H. (1963). Brains and behavior. Reprinted in H. Putnam (1975), *Mind, language and reality: Philosophical papers, Volume 2* (pp. 325–341). Cambridge University Press.

Putnam, H. (1967). The nature of mental states. Reprinted in H. Putnam (1975b), *Mind, language and reality: Philosophical papers, Volume 2* (pp. 429–440). Cambridge University Press.

Putnam, H. (1973). Meaning and Reference. The Journal of Philosophy, 70(19), 699–711. https://doi.org/10.2307/2025079

Putnam, H. (Ed.). (1975a). Explanation and reference. In *Philosophical Papers* (pp. 196–214). chapter, Cambridge: Cambridge University Press.

Putnam, H. (Ed.). (1975b). The meaning of 'meaning.' In *Philosophical Papers* (pp. 215–271). chapter, Cambridge: Cambridge University Press.

Psillos, S. (2006). The, structure, the whole, structure and nothing but, the structure? *Philosophy of Science, 73*(5), 560–570. https://doi.org/10.1086/518326

Psillos, S. (2009). *Knowing the structure of nature: Essays on realism and explanation*. Palgrave Macmillan. https://doi.org/10.1057/9780230234666

Psillos, S. (2022). Realism and theory change in science. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 edition). Stanford University. https://plato.stanford.edu/archives/fall2022/entries/realism-theory-change/

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences, 3*(1), 111–132. https://doi.org/10.1017/S0140525X00002053

Pylyshyn, Z. (1984). *Computation and cognition: Toward a foundation for cognitive science*. The MIT Press.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences, 46*, e261. https://doi.org/10.1017/S0140525X22002849

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, *60*(1), 20–43.

Quine, W. V. O. (1975a). Empirically equivalent systems of the world. Erkenntnis, 9, 313–328.

Quine, W. V. O. (1975b). The nature of natural knowledge. In S. Guttenplan (Ed.), *Mind and language* (pp. 67–81). Oxford University Press.

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.
Reichenbach, H. (1938). *Experience and prediction*. University of Chicago Press.

Rescorla, M. (2024). The language of thought hypothesis. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2024 edition). Stanford University. https://plato.stanford.edu/archives/sum2024/entries/language-thought/

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... & Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Richardson, A., & Uebel, T. (Eds.). (2007). *The Cambridge Companion to Logical Empiricism*. Cambridge University Press.

Rorty, R. (1998). Charles Taylor on truth. In *Truth and progress: Philosophical papers* (pp. 84–97). Cambridge University Press.

Rupert, R. (1999). The best test theory of extension: First principle(s). *Mind and Language*, *14*(3), 321–355.

Rupert, R. (2008). Causal theories of mental content. *Philosophy Compass*, *3*(2), 353–380.

Rupert, R. (2013). Memory, natural kinds, and cognitive extension; or, martians don’t remember, and cognitive science is not about cognition. *Review of Philosophy and Psychology*, *4*(1), 25–47.

Ryle, G. (1949). The concept of mind. University of Chicago Press.

Sanches de Oliveira, G., & Baggs, E. (2023). *Psychology’s WEIRD problems*. Cambridge University Press.

Sapolsky, R. (2023). *Determined: A science of life without free will*. Penguin Press.

Schmidt, S. R. (1991). Can we have a distinctive theory of memory? *Memory & Cognition*, *19*(6), 523–542. https://doi.org/10.3758/BF03197149

Schulte, P. (2023). *Mental content*. Cambridge University Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

Searle, J. R. (1990). Cognitive science and the computer metaphor. In B. Göranzon & M. Florin (Eds.), *Artificial intelligence, culture and language: On education and work*. Springer. https://doi.org/10.1007/978-1-4471-1729-2_4

Segundo-Ortin, M., & Raja, V. (2024). *Ecological psychology*. Cambridge University Press.

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science, 1*, 253–329.

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

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese, 153*, 393–416. https://doi.org/10.1007/s11229-006-9099-8

Shapin, S., & Schaffer, S. (2017). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Sider, T. (2011). *Writing the book of the world*. Oxford University Press.

Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review, 68*, 141–156.

Song, Y., Lukasiewicz, T., Xu, Z., & Bogacz, R. (2020). Can the brain do backpropagation? Exact implementation of backpropagation in predictive coding networks. *Advances in Neural Information Processing Systems, 33*, 22566–22579.

Stampe, D. (1977). Toward a causal theory of linguistic representation. In P. A. French, T. E. Uehling, & H. K. Wettstein (Eds.), *Studies in the philosophy of language. Midwest Studies in Philosophy: Vol. 2* (pp. 81–102). University of Minnesota Press.

Stanford, P. K. (2006). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*. Oxford University Press. https://doi.org/10.1093/0195174089.001.0001

Stanford, K. (2023). Underdetermination of scientific theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2023 edition). Stanford University. https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination/

Stengers, I. (2023). *Virgin Mary and the Neutrino: Reality in Trouble*. Experimental Futures.

Stewart, J., Gapenne, O., & Di Paolo, E. A. (Eds.). (2010). *Enaction: Toward a new paradigm for cognitive science*. MIT Press.

Şahin, Y., & Deveci, I. E. (2023, May 7). Against representations that do not represent: Reply to Thomson and Piccinini [Poster presentation]. *9th International Symposium on Brain and Cognitive Science*, Özyeğin University, Istanbul, Turkey.

Tabery, J. (2014). *Beyond versus: The struggle to understand the interaction of nature and nurture*. The MIT Press.

Teller, P. (2001). Whither constructive empiricism? *Philosophical Studies, 106*(1–2), 123–150. 10.1023/a:1013170506726

Tirard, S., Morange, M., & Lazcano, A. (2010). The definition of life: A history of an elusive scientific endeavor. *Astrobiology, 10*(10), 1003–1009.

Usher, M. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language, 16*(3), 311–334.

van den Herik, J. C. (2014, July 8). *Why radical enactivism is not radical enough*. Retrieved from http://hdl.handle.net/2105/16132

VandeWall, H. (2007). Why Water Is Not H_2O, and Other Critiques of Essentialist Ontology from the Philosophy of Chemistry. *Philosophy of Science*, *74*(5), 906–919. https://doi.org/10.1086/525632

Van Fraassen, B. (1980). The scientific image. Oxford University Press.

Van Fraassen, B. (1985). Empiricism in the philosophy of science. In P. M. Churchland & C. A. Hooker (Eds.), *Images of science: Essays on realism and empiricism* (pp. 245–308). University of Chicago Press.

Van Fraassen, B. C. (1989). *Laws and symmetry*. Clarendon Press.

A. Hooker (Eds.), *Images of science: Essays on realism and empiricism* (pp. 245–308). University of Chicago Press.

Van Fraassen, B. C. (2001). Constructive empiricism now. *Philosophical Studies, 106*(1–2), 151–170. https://doi.org/10.1023/a:1013126824473

Van Fraassen, B. (2002). The empirical stance. Yale University Press.
Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy, 92*(7), 345–381.

Varela, F. J. (1979). *Principles of biological autonomy*. North-Holland.

Varela, F. J., Thompson, E., & Rosch, E. (1991). The embodied mind: Cognitive science and human experience. The MIT Press.

Villalobos, M. (2015). *The biological roots of cognition and the social origins of mind* (Doctoral dissertation, University of Edinburgh).

Villalobos, M. (2022). The concepts of mind and cognition in the autopoietic theory. *Constructivist Foundations, 18*(1), 26–27. https://constructivist.info/18/1/026

Villalobos, M., & Palacios, S. (2021). Autopoietic theory, enactivism, and their incommensurable marks of the cognitive. *Synthese, 198*(Suppl 1), 71–87. https://doi.org/10.1007/s11229-019-02376-6

Villalobos, M., & Silverman, D. (2018). Extended functionalism, radical enactivism, and the autopoietic theory of cognition: Prospects for a full revolution in cognitive science. *Phenomenology and the Cognitive Sciences, 17*(4), 719–739. https://doi.org/10.1007/s11097-017-9542-y

Villalobos, M., & Ward, D. (2015). Living systems: Autonomy, autopoiesis and enaction. *Philosophy & Technology, 28*, 225–239. https://doi.org/10.1007/s13347-014-0154-y

Vygotsky, L. S. (2012). *Thought and language* (A. Kozulin, Ed.; revised and expanded ed.). MIT Press

Walsh, D. (2015). *Organisms, agency, and evolution*. Cambridge University Press. https://doi.org/10.1017/CBO9781316402719

Ward, D., Silverman, D., & Villalobos, M. (2017). Introduction: The varieties of enactivism. *Topoi, 36*, 365–375. https://doi.org/10.1007/s11245-017-9484-6

Weber, M. (2022). *Philosophy of developmental biology*. Cambridge University Press.

Wheeler, M. (2017). The revolution will not be optimised: Radical enactivism, extended functionalism and the extensive mind. *Topoi*, *36*(3), 457–472.

Wittgenstein, L. (1953). Philosophical Investigations (G. E. M. Anscombe, Trans.). Blackwell Publishing.

Wray, K. B. (2013). The pessimistic induction and the exponential growth of science reassessed. *Synthese, 190*(18), 4321–4330. https://doi.org/10.1007/s11229-013-0276-2

Wray, K. B. (2015). Pessimistic inductions: Four varieties. *International Studies in the Philosophy of Science, 29*(1), 61–73. https://doi.org/10.1080/02698595.2015.1071551