

BEYOND ALGORITHM: ON THE INADEQUACY OF COMPUTATIONAL
PARADIGM FOR PHENOMENAL MACHINE CONSCIOUSNESS

İŞİKGÜN AKFİRAT

BOĞAZİÇİ UNIVERSITY

2025

BEYOND ALGORITHM: ON THE INADEQUACY OF COMPUTATIONAL
PARADIGM FOR PHENOMENAL MACHINE CONSCIOUSNESS

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
Philosophy

by
Işıkgün Akfırat

Boğaziçi University

2025

Beyond Algorithm: On the Inadequacy of Computational
Paradigm for Phenomenal Machine Consciousness

The thesis of Işıkgün Akfırat
has been approved by:

Assoc. Prof. Lucas Thorpe
(Thesis Advisor)

Assist. Prof. Sun Demirli

Assoc. Prof. Aziz Zambak
(External Member)

June 2025

DECLARATION OF ORIGINALITY

I, Işıkgün Akfırat, 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

Beyond Algorithm: On the Inadequacy of Computational Paradigm for Phenomenal Machine Consciousness

This thesis investigates the theoretical plausibility of artificial phenomenal consciousness, the conditions under which artificial systems could have phenomenal experience, and questions whether computational approaches can provide an adequate explanation of this phenomenon. Its central claim is that while computation may model intelligent behavior and simulate cognition, it fails to account for the ontological reality of phenomenal experience. Through a critical analysis of computationalist models, the thesis identifies a causal residue, an irreducible remainder that abstract structure cannot explain. This residue marks the boundary at which computational sufficiency collapses. In its final chapters, the thesis explores non-computational alternatives that seek to ground consciousness not in formal mechanisms, but in physically specific and world-involving processes, arguing that any serious model of phenomenal machine consciousness must derive its explanatory power from this orientation.

ÖZET

Algoritmanın Ötesinde: Fenomenal Makine Bilinci İçin

Hesaplamacı Paradigmanın Yetersizliği Üzerine

Bu tez, yapay sistemlerin hangi koşullarda fenomenal bilince sahip olabileceğini ve bu olgunun açıklanmasında hesaplamacı (kompütasyonalist) yaklaşımların yeterli olup olmadığını sorgulamaktadır. Tezin temel savı, hesaplama nosyonunun zeki davranış modelleme ve bilişi simüle etmede etkili olsa da, fenomenal deneyimin ontolojik varlığını açıklamakta yetersiz kaldığıdır. Önde gelen hesaplamacı modellerin eleştirel bir çözümlemesiyle, soyut yapının açıklayamadığı nedensel bir artık, indirgenemez bir fazlalık tespit edilmektedir. Bu artık, hesaplamasal yeterliliğin sınırını işaret eder. Tezin son bölümleri, bilinci biçimsel mekanizmaların ötesinde, fiziksel olarak özgül ve dünyayla ilişkisel süreçlerde temellendirmeyi amaçlayan hesaplamacı olmayan alternatifleri inceleyerek, fenomenal makine bilinci üzerine geliştirilecek her ciddi yaklaşımın açıklayıcı gücünü bu yönelimden almak zorunda olduğunu öne sürmektedir.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
CHAPTER 2: THE QUEST FOR PHENOMENAL MACHINE CONSCIOUSNESS.....	9
2.1 Mapping the conceptual space for artificial consciousness	9
2.2 Artifacts with qualia: Theories of machine phenomenality	17
2.3 A Theoretical matrix: Causal efficacy and substrate independence	36
2.4 Defusing the impossibility arguments	47
CHAPTER 3: EXPLANATORY LIMITS OF COMPUTATIONAL PARADIGM.....	52
3.1 The Foundations of computationalism	53
3.2 Computational models of phenomenal machine consciousness	62
3.3 Critiques of computational sufficiency	79
3.4 The causal residue thesis	103
CHAPTER 4: THE PROSPECT FOR A NON-COMPUTATIONAL ALTERNATIVE.....	109
4.1 Non-computational models of machine phenomenality	111
4.2 Good old-fashioned artificial consciousness and radical alternative.....	124
4.3 Critical synthesis and evaluation	130
4.4 A Speculative proposal: Hermeneutic constructivism	133
CHAPTER 5: CONCLUSION	144
REFERENCES	150

CHAPTER 1

INTRODUCTION

In this thesis, I will investigate a deeply intricate question that lies at the intersection of philosophy of mind and artificial intelligence: Can a machine be phenomenally conscious? More precisely, I will examine whether phenomenal consciousness, the capacity for subjective experience, the *what-it-is-like-ness* (Nagel, 1974), can emerge within the boundaries of the predominant computational paradigm. This amounts to answering the question, “Is consciousness an algorithmic phenomenon of computational nature” (Chella & Manzotti, 2011, p. 4) or something else? This is a timely inquiry, as we observe that artificial intelligence is making dizzying advances, mastering tasks once thought to require sentient minds. No matter how fluent the language model or how dexterous the robot, the fundamental challenge remains: Can there be something it is like to be the machine? If not, why not? And if so, how?

Over the past decade, we have witnessed a sea change in our conception of what artificial intelligence can accomplish. Machines have begun to breach the frontiers once believed to be uniquely human or classified as “what computers cannot do” (Dreyfus, 1972). AlphaGo’s defeat of a world champion in Go, long thought to lie beyond the reach of computation, was a watershed moment (Silver et al., 2016). Soon after, AlphaFold (Jumper et al., 2021) met Kitano’s (2016) challenge to build an AI capable of Nobel-caliber scientific achievement. By predicting the three-dimensional structures of proteins with unprecedented accuracy, the program solved a grand challenge in molecular biology, one that had eluded researchers for decades. In recognition of this contribution, the 2024 Nobel Prize in Chemistry was

awarded to the developers of AlphaFold and its foundational model, the first such honor ever granted to an artificial intelligence system.

The recent ascent of advanced large language models (LLMs) such as ChatGPT has blurred the line between narrow AI and more generalized forms of intelligence. Some researchers point to what they describe as “sparks” of artificial general intelligence (AGI) already visible in these LLMs (Bubeck et al., 2023; see also Chalmers, 2024). In parallel, what was once dismissed as speculative alarmism, the idea of a “singularity” in which superintelligent machines exceed human control (Kurzweil, 2005; Bostrom, 2014), is now treated by many as a legitimate concern. Others remain more cautious, criticizing this discourse as a form of “anthropomorphism.” To speak of these systems as if they “hallucinate” or “reason” (Seth, 2024), they argue, risks obscuring the statistical nature of their operations and overstates their capacities. As Bender et al. (2021) memorably put it, these models may be less like proto-minds and more like “stochastic parrots”: systems that recombine patterns without grasping meaning. In this light, questions of control and alignment give way to deeper concerns: not just about what these systems can do, but what they might eventually become.

There is a growing sense that the pursuit of AGI (Goertzel, 2014) may be stalled not by algorithmic limitation, but by the absence of something more fundamental: consciousness (Manzotti & Chella, 2012, p. 264; Signorelli, 2018). It has been suggested that some of the longstanding challenges in AI, the Frame Problem (McCarthy & Hayes, 1969) and the Symbol Grounding Problem (Harnad, 1990), may not be tractable without first addressing the question of machine consciousness. For instance, Haikonen (2020) has argued that genuine symbol grounding requires architectures capable of generating qualia, as intrinsic meaning

cannot arise from syntax alone. A complementary view, drawing on the active inference framework, proposes that the Frame Problem, essentially a problem of contextual relevance, can only be resolved through the conscious modulation of uncertainty (Linson et al., 2018). According to Reggia et al. (2023, pp. 244-5), in the face of these long-standing problems, what is often pursued as AGI would eventually have to be fused with artificial consciousness into “Artificial Conscious Intelligence.”

Hence, the stakes are not merely theoretical. If artificial systems are approaching the threshold of experiencing or having qualia, then we are no longer simply extending the toolset of human ingenuity, but we may be on the cusp of creating artificial subjects. Even now, public intuitions are shifting: many ordinary users of LLMs already ascribe some degree of phenomenal consciousness to these systems, suggesting that the conceptual threshold for treating machines as subjects may be lower than we think (Colombatto & Fleming, 2024). At the same time, some philosophers have argued that language-based agents may already possess the necessary architectural conditions for phenomenal consciousness, suggesting that, under certain interpretations of Global Workspace Theory (Baars, 2005), consciousness might not be a distant threshold but an imminent possibility (Goldstein & Kirk-Giannini, 2024).

The growing entanglement between humans and artificial agents is prompting a reconsideration of long-standing assumptions about consciousness and moral status.¹ Thomas Metzinger (2021) warns that the deliberate creation of “synthetic phenomenology,” particularly states of negatively valenced experience, entails

¹ As Henry Shevlin (2024) observes, a deep tension emerges between three widely held views: that consciousness is more than behavior, that only the conscious deserve moral concern, and that behaviorally equivalent or structurally similar entities ought to be treated alike.

ethical risks that far exceed our current epistemic capacities. He therefore calls for a precautionary measure: until we develop reliable means of identifying and preventing artificial suffering, research aimed at generating phenomenally conscious machines should be placed under a global moratorium. In contrast, Benitez et al. (2024) oppose halting the pursuit of artificial consciousness altogether. They propose an architectural compromise: artificial agents could be designed to possess cognitive representations of harm and morality while remaining insulated from the affective core of suffering.

Nevertheless, insulating machines from suffering may be neither feasible nor desirable. Some argue that, without the capacity to attribute consciousness to others, artificial agents may behave with the detachment of sociopaths, lacking the social cognition needed for genuinely cooperative behavior (Graziano, 2017, pp. 7-8). It has even been suggested, rather provocatively, that artificial systems, unburdened by the biological and cultural biases that constrain human judgment, might surpass us in moral reasoning. Robots capable of feeling-based empathy, grounded in stable architectures rather than instinct or ideology, could in principle become more reliable moral agents than their creators (Man & Damasio, 2019, p. 451). Pushing this vision further, Gamez (2018) offers a deliberately unsettling possibility: that future generations of artificial consciousness may not only match but exceed human moral competence, eventually replacing flawed human governance with machines capable of principled rule. In this scenario, the ethical trajectory of artificial consciousness ends not in catastrophe, but in the quiet emergence of a “brave new world without sin” (p. 145).

Beyond the ethical horizon lies even a deeper intellectual prize: the prospect that by building consciousness, we might finally come to understand it. The project

of artificial consciousness is not merely about machines; it is a means of probing the most elusive feature of our own minds. As Richard Feynman famously inscribed on his blackboard, “What I cannot create, I do not understand.” This perspective is followed by machine consciousness researchers Chella and Manzotti (2011, p. 10), who argue that “a successful approach to the phenomenal aspect of artificial consciousness will not only be a model of a machine, but rather a more general and overarching theory that will deal with the structure of reality.” Constructing systems that may instantiate experience offers a new kind of empirical leverage: one that allows us to test, refine, or discard our theories of consciousness under conditions that are tractable, replicable, and arguably ethically permissible. Unlike the human brain, whose complexity as well as moral concerns limit experimental access, artificial architectures can be manipulated with precision. In this sense, artificial consciousness is not only a technological ambition but also a philosophical instrument: a mirror turned toward the mind, capable of revealing what biology alone may keep hidden.

It is within this shifting landscape that the field of artificial or machine consciousness (MC)² has taken shape, not as a speculative byproduct of artificial intelligence, but as a research field in its own right (Gamez, 2008; Elamrani, 2025). Approaches to MC often diverge according to their guiding motivation. On one side stands a technological trajectory, driven by the ambition to construct systems with human-like cognitive capacities, not directly concerned with the metaphysical

² In recent literature, the term artificial consciousness (AC) appears to have gained broader usage. Seth (2009, p. 71) prefers it because of its connection to the broader tradition of artificial intelligence. However, I will mainly use the term machine consciousness and its corresponding abbreviation (MC) to avoid confusion with the Block’s distinction between access consciousness and phenomenal consciousness (AC vs. PC). That said, to avoid stylistic monotony, I occasionally use artificial consciousness interchangeably throughout the thesis. In both usages, what is meant is the possibility of consciousness emerging in non-human, artificial systems.

troubles of consciousness (Holland & Goodman, 2003; Tagliasco, 2007; Sanz & Aguado, 2020). Holland (2020) goes so far as to suggest that we should “forget [Nagel’s] bat” altogether: real progress, he argues, lies not in chasing the inaccessible terrain of first-person experience but in building systems whose internal models of perception and motivation produce a kind of machine-readable seeming. This line of inquiry treats consciousness as a functional threshold. Its priority lies not in theoretical resolution, but in architectural viability: systems that perform, adapt, and perhaps eventually feel. The engineering imperative here is clear: design precedes understanding, and consciousness, if it comes, will emerge through implementation.

On the other side lies a more theoretical (or, more precisely, philosophical) approach. Here, machine consciousness is not treated merely as an engineering puzzle but as a conceptual object in its own right, with epistemic and ontological dimensions that remain deeply unsettled. From this standpoint, artificial systems thus serve as testbeds for philosophical accounts, offering concrete models through which theories of consciousness can be clarified, constrained, or refuted (Manzotti & Tagliasco, 2008; Arrabales et al., 2009). As Chella and Manzotti put it,

machine consciousness is not simply a technological challenge, but a field posing theoretical and scientific issues such as the relations between information and meaning, the nature of teleology, the unity of the self, the integration of information, the nature of experience, and many others from a novel point of view. (2009, p. 1)

This approach recognizes that the hardest questions about consciousness are not solved by implementation alone; they demand conceptual clarity about what is being instantiated, and why it matters. Still, this philosophical orientation need not stand in opposition to engineering efforts. As Chrisley (2008; 2009) argues, a productive tension can emerge through what he calls “interactive empiricism,” in which theory

and practice are mutually conducive. The search for artificial consciousness, then, moves along intersecting paths: building to understand, and understanding to build.

However, it is crucial to note that not all talk of machine consciousness refers to the same thing. As Gamez (2008) proposes, research on machine consciousness can be divided into four categories: (MC1) machines with the external behaviour associated with consciousness; (MC2) machines with the cognitive characteristics of consciousness; (MC3) machines with architectures claimed to cause or correlate with consciousness; and (MC4) phenomenally conscious machines.³ The present thesis is concerned exclusively with MC4. It sets aside the approaches that bypass the phenomenal (MC1-3 absent MC4), as well as the theories of phenomenal consciousness that offer no concrete proposals for its artificial realization (non-MC). What follows, then, an inquiry into the ontological possibility of machine experience, and the conditions under which it might emerge.

Our investigation proceeds from two claims. First, that phenomenal consciousness is ontologically possible in artificial systems. Second, that computationalism is not the right path to it. The first is a claim about nature: if consciousness can arise in us, and if we are natural systems, then nothing precludes its arising elsewhere, under the right conditions. That is, phenomenal machine consciousness is possible.⁴ The second is a claim about explanatory adequacy: that

³ A related categorization is the Machine-Qualia (M-Q) type proposed in Qin et al.'s (2025) recent taxonomy. They emphasize the centrality of phenomenality within machine consciousness research: "Phenomenal consciousness has a special significance for the study of machine consciousness. If there are something 'mysterious' or 'irreducible' in the nature of consciousness, there is no hope of achieving machine consciousness from the standpoints of science and engineering alone" (p. 10).

⁴ While I employ the term "machine consciousness" as a neutral label for artificial systems potentially capable of phenomenal states, this should not be read as endorsing a mechanistic reduction of biological or conscious processes. A central claim of this thesis is that computational models fail to capture the causal grounding necessary for phenomenality, an insight that resonates with organicist critiques of mechanistic metaphysics (Nicholson, 2018). Yet this does not imply a rejection of artificial consciousness as such. The concept of machine need not be tied to rigid mechanical metaphors; it can encompass materially realized, self-organizing systems with emergent causal capacities. In this broader sense, organisms themselves can be seen as machines. If artificial systems instantiate similarly organized architectures, there is no principled reason to exclude the possibility of

computation and information, while indispensable as modeling tools, are not ontological primitives. They are epistemic constructs, which we use to describe systems, not entities with causal power sufficient to generate experience.

I will pursue the following course. In Chapter 2, I will survey the conceptual terrain, introducing a theoretical matrix that maps positions on artificial consciousness along two axes: causal efficacy and substrate dependence.

In Chapter 3, I will turn to computationalism itself, examining the strategies that claim to explain or instantiate subjective experience through computational means. I will argue that the principle of computational sufficiency fails to account for phenomenal consciousness because computation, as an epistemic construct, cannot ground the kind of ontological presence required, leaving behind “a causal residue.”

In Chapter 4, I will examine a range of non-computational alternatives that are candidates to capture the causal residue left unexplained by computational models. I will argue that their shared emphasis on grounding consciousness in physically specific and ontologically non-abstractable processes points toward the need for a paradigm shift, one that I begin to sketch in the speculative proposal of hermeneutic constructivism.

The thesis concludes that advancing a viable theory requires acknowledging the explanatory limits of computationalism while integrating non-computational insights into a broader naturalist framework grounded in phenomenal realism and ontological monism. I hold that between computationalism and epistemic defeatism another route may be found.

non-biological phenomenal consciousness. Chapter 4 develops this line of reasoning by exploring non-computational models of artificial phenomenality and reinterpreting the category of the machine beyond classical mechanistic constraints.

CHAPTER 2

THE QUEST FOR PHENOMENAL MACHINE CONSCIOUSNESS

This chapter maps the core philosophical threads of artificial phenomenal consciousness. It begins by clarifying the foundational assumptions and theoretical tensions that shape the field, particularly the distinction between functional and phenomenal approaches. It then surveys a range of proposals that explicitly address the problem of qualia in machines, the internal, subjective qualities of experience that are typically taken to define phenomenal consciousness. These accounts are situated within a two-axis theoretical matrix, highlighting key divergences over the causal role of phenomenality and the paradigmatic conditions of its realization. The chapter concludes by engaging with prominent impossibility arguments, examining whether machine consciousness is not only imaginable but scientifically and philosophically tenable.

2.1 Mapping the conceptual space of artificial consciousness

Over the past three decades, machine consciousness (MC) has moved from the margins of speculative thought into a nascent multidisciplinary field at the intersection of artificial intelligence, neuroscience, cognitive science, and philosophy (Gamez, 2008; Chella & Manzotti, 2011; Elamrani, 2025). A pivotal moment came in 2001, at the Cold Spring Harbor symposium titled “Can Machines Be Conscious?”, where leading figures, including David Chalmers, Christof Koch, and James Goodman, converged on a rare consensus: “We know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artifacts designed or evolved by humans” (cited in Aleksander, 2007, p. 78). This

naturalist permissibility claim has since underwritten successive waves of interest in the field, visible in special journal issues, journals, international conferences, and increasingly ambitious engineering proposals.

The divergent motivations behind the study of machine consciousness have become increasingly apparent in time. As noted before, two distinct currents animate the field. On one side lies the philosophical-scientific pursuit, in which artificial systems are treated as instruments for investigating the nature of consciousness itself. On the other lies a technological-pragmatic drive, focused on engineering systems with human-like intelligence, of which consciousness is presumed to be a functional milestone. Most of the concrete frameworks proposed thus far have emerged from the latter camp, often grounded in computationalist theories of mind and directed toward their implementation in novel cognitive architectures. Underpinning these efforts is a foundational assumption, what might be called “the Engineering Thesis” (ET): if consciousness is a natural phenomenon, then it should, in principle, be reproducible through the engineering of its relevant functional or computational features (Boltuc, 2012). On this view, to build consciousness is not to conjure the inexplicable, but to replicate what nature has already shown to be physically possible (Butazzo, 2008).

The Engineering Thesis, for all its intuitive appeal, conceals a tangle of philosophical difficulties. First among them is a conceptual ambiguity: what, precisely, do we mean by “consciousness” that machines are supposed to have? Most researchers in the field now seem to recognize the canonical distinction by Ned Block (1995) between “access consciousness” (AC), defined functionally in terms of report, reasoning, and control of behavior; and “phenomenal consciousness” (PC), the qualitative character of experience, or “what it is like to be in that state.” In other

words, to be phenomenally conscious is to have a point of view, to have an inner perspective on the world that is irreducible to third-person description (Nagel 1974, p. 442). Whereas AC lends itself to functional explanation, PC resists such reduction, giving rise to what has been termed an “explanatory gap”: the failure of physicalist or functionalist accounts to explain why specific mechanisms give rise to specific qualitative feels (Levine, 1983).⁵ Chalmers (1995) sharpened this intuition by distinguishing the “hard problem” of consciousness, why and how physical processes should give rise to felt experience at all, from the more tractable, or “easy,” problems of explaining cognitive function of consciousness. It is the hard problem that anchors our concern with phenomenal consciousness.

Within the field of artificial consciousness, three broad tendencies can be observed in response to the challenge posed by phenomenal experience. Some researchers set it aside from the outset, treating it as either scientifically intractable and unfruitful from an engineering point of view (Holland, 2020) or the result of a “category error” (O’Regan 2012, p. 125). Others appear to acknowledge the distinction between access and phenomenal consciousness but fall into what Block (1995, p. 228) has identified as a “persistent fallacy”: the conflation of the two, or the reduction of PC to AC. A third camp takes the problem seriously. It accepts the irreducibility of phenomenal consciousness and argues that machine consciousness may offer new resources for facing up to the hard problem.⁶ As Chella and Manzotti

⁵ As Nagel (1974, p. 436) originally put it, subjective experience “is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing.”

⁶ In many ways, this can be seen as a realization of Putnam’s early prediction (1964) that the machine consciousness debate would force a reengagement with foundational problems in the philosophy of mind: “What I hope to persuade you is that the problem of the Minds of Machines will prove, at least for a while, to afford an exciting new way to approach quite traditional issues in the philosophy of mind. Whether, and under what conditions, a robot could be conscious is a question that cannot be discussed without at once impinging on the topics that have been treated under the headings Mind-Body Problem and Problem of Other Minds” (Putnam, 1964, p. 669).

put bluntly (2009, p. 3), sidestepping the hard problem is “not a viable option in the business of making conscious machines.”

Piotr Boltuc (2009) pushes the conceptual critique further by arguing that even Block, despite coining the distinction between access and phenomenal consciousness, ultimately fails to maintain it, particularly in relation to the Problem of Implementation (p. 156). To the extent that Block considers how phenomenal states might be instantiated in the brain (or in a machine), he shifts from a purely subjective notion to one embedded in information processing and cognitive access.⁷ This slippage, Boltuc contends, reintroduces reductionism through the back door. In response, he proposes a stricter alternative: “Hard Consciousness” (H-consciousness), which is defined as entirely non-functional, purely subjective awareness (p. 157). On this view, any system (biological or non-biological) that lacks an irreducible first-person perspective may process information and guide behavior, but it does so in the dark. Consciousness, if it exists, is not what the system does, but that there is something it is like to be it.⁸

This concern, then, is not simply a terminological refinement. The moment we move from describing consciousness to attempting its construction, the Engineering Thesis reveals a deeper question: what counts as a genuine presence of phenomenal experience in artifacts? This is not a debate over behavioral resemblance or functional mimicry, but a matter of ontological status. And it is precisely this tension that John Searle (1980) brings into the foreground through his well-known

⁷ Boltuc targets a subtle but significant inconsistency in Block’s formulation: “The idea is that phenomenal consciousness is distinct (at least conceptually) from that information processing function, but is part of the implementation of it” (Block 1995, p. 378). A similar concern appears in Chalmers’ critique: “Any purported function for p-consciousness can be attributed to a-consciousness” (Chalmers, 1997, p. 148).

⁸ Curiously, Boltuc does not entirely discard PC talk either: “I leave the term p-consciousness to designate a first-person functional consciousness (needed for the projects such as LIDA) and coin the term h-consciousness to designate just the non-functional concept of consciousness” (2009, p. 158).

distinction between simulation (or instantiation of program) and duplication. A system that simulates digestion does not digest; a machine that simulates flight does not fly. As some authors pointed, the difference is analogous to that between an artificial flower (imitation) and artificial light (replication); only the latter preserves its identity regardless of how it is produced (Manzotti & Tagliasco, 2008, p. 106). Consciousness, on Searle's account, cannot be exhausted by computer programs, which are taken to be formal implementations of syntactic rules. It must be realized in systems with the right kind of causal powers. Until further notice, he maintains, the only known substrate capable of giving rise to consciousness remains the human (or more broadly, biological) brain.

Accordingly, Searle's classic distinction between Weak vs. Strong AI (1980) finds a direct analogue in machine consciousness. Weak MC refers to systems that simulate the functional correlates of consciousness (information processing, behavior, responsiveness) without implying any experiential interiority. Strong MC, by contrast, asserts that a system is not merely acting "as if" it is conscious⁹, but is itself a locus of subjective experience (Holland, 2003, cited in Manzotti 2007).¹⁰ In other words, the business of duplicating consciousness "mean[s] efforts to produce an artificial system that actually is phenomenally conscious, i.e., that experiences qualia" (Reggia et al., 2023, p. 244). If Searle is right, then the pursuit of artificial

⁹ Chella and Manzotti (2011) even question this distinction suggesting it may be "fictitious," along with the distinction between phenomenal vs. access consciousness. Drawing on an evolutionary argument, they write: "Human consciousness entails phenomenal consciousness at all levels. In sum, it would be very bizarre whether natural selection had gone at such great length to provide us with consciousness if there was a way to get all the advantages of a conscious being without it actually being phenomenally so" (p. 20). Cf. Haikonen, 2012, p. 81.

¹⁰ Ron Chrisley (2008), seeking a middle path, introduces the notion of "Lagom AC," a minimalist framework identifying features necessary (but not sufficient) for consciousness. This, he suggests, allows researchers to explore machine consciousness without overcommitting to the full burden of phenomenal sufficiency. Anil Seth (2009), skeptical of strong, axiomatic approaches (e.g., Aleksander, 2003), regards Weak MC as scientifically more tractable. By linking it to Explanatory Correlates of Consciousness (ECC), neural or computational features that reliably co-occur with experience, Seth envisions a research program that defers metaphysical claims until the causal conditions of consciousness are more robustly understood.

consciousness confronts a fundamental constraint: the emergence of consciousness is substrate dependent, not substrate neutral. This challenges one of the central assumptions of functionalism, Multiple Realizability, according to which mental states can be instantiated in any medium that supports the right functional organization (Putnam, 1975; Cao, 2022).

On exactly this point, proponents of computationalism diverge from Searle's restriction, reaffirming the ambition of Strong MC. For them, the relevant causal powers are not confined to biology but can, in principle, emerge from the right kind of functional organization. The path to artificial phenomenality, on this view, lies in replicating not the substrate, but in reproducing the structure. Central to this view is the thesis of Functional Equivalence (FE): if two realize the same causal roles and information-processing structure, they are taken to be mentally equivalent, including, potentially, in their capacity for phenomenal states.¹¹ Computational Sufficiency (CS), which is held by many proponents of Strong MC, builds upon this assumption but advances a stronger claim: that implementing the right functional organization, typically cast in algorithmic or computational terms, is not only necessary, but sufficient for phenomenal consciousness to arise.¹² Boltuc & Boltuc (2007, p. 24-5) links this stance to the Engineering Thesis, recasting it as an explanatory imperative: to understand consciousness is to make it algorithmically expressible¹³, and once

¹¹ Putnam (1964, p. 677) offers an early formulation of functional equivalence in his notion of "psychological isomorphism": "If a human being is a 'probabilistic automaton,' then any robot with the same 'machine table' will be psychologically isomorphic to a human being. If the human brain is simply a neural net with a certain program, as in the theory of Pitts and McCulloch, then a robot whose 'brain' was a similar net, only constructed of flip-flops rather than of neurons, would have exactly the same psychology as a human." He qualifies, however, that having a psychology does not necessarily entail being conscious.

¹² More detailed analysis of the theoretical commitments and metaphysical implications of Functional Equivalence (FE) and Computational Sufficiency (CS) theses will be provided in §3.1.

¹³ Boltuc & Boltuc (2007, p. 24) define an algorithm as "an executable function (both in terms of programming and an algorithmic description in natural science, e.g., chemistry or physics)." Their emphasis is to make consciousness tractable as an algorithmic object, both implementable and explanatorily transparent.

expressible, in principle, constructible. On this view, the engineering of consciousness is not a speculative extrapolation but the culmination of genuine theoretical understanding.

Even if we grant the plausibility of this engineering ambition, another challenge immediately follows: what would count as success? If a machine is said to instantiate phenomenal consciousness, by what epistemic standards could such a claim be assessed? The issue presses directly on what it would mean to test, verify, or falsify claims about machine consciousness. This is the Epistemic Problem which directly stems from a familiar philosophical tension, the Problem of Other Minds (Harnad 1991). This line of concerns has led some philosophers to adopt epistemic agnosticism about machine consciousness (McGinn, 1987; Prinz, 2003). Since consciousness is subjective, we have no direct access to its presence in others, human or machine. Nevertheless, science and engineering proceed not by epistemic defeatism, but by formulating operational criteria that bridge theoretical commitments and observable indicators.

A number of such criteria have been proposed, some rooted in behavior, others in architecture or embodiment. The most well-known of these, the Turing Test (1950), has long been subject to criticism, not only for failing to capture subjective experience, but even as a reliable benchmark for intelligence. Relying solely on linguistic indistinguishability, it is prone to sophisticated mimicry devoid of understanding (for a review, see Saygin et al., 2000). In response, a range of alternatives have emerged. Susan Schneider's (2019) AI Consciousness Test (ACT) and Chip Test aim to detect continuity of subjective states through structural or neural replacement scenarios. Arrabales et al.'s (2010) ConScale offers a multi-dimensional scale of cognitive complexity, while Zador et al. (2023) propose an

Embodied Turing Test, grounded in neurobiological constraints. These frameworks offer important tools for refining our assessments. Yet a common concern remains: behavioral or structural resemblance does not entail phenomenality. Without a theoretical framework linking observable markers to subjective experience, such tests may underdetermine what they aim to reveal.

This is why many researchers direct their focus toward articulating Necessary and Sufficient Conditions (NSC) for artificial consciousness, not as definitive proof, but as a more principled basis for attributing phenomenality to artificial systems (Butlin et al., 2023; Farisco et al., 2024). Rather than relying on behavior or substrate alone, these approaches aim to identify the architectural, functional, or computational features that must be present for consciousness to plausibly arise. In neuroscience, this strategy is exemplified by the search for Neural Correlates of Consciousness (NCCs), the minimal neural mechanisms presumed to be jointly sufficient for particular conscious states (Crick & Koch, 1990). A parallel effort has emerged in artificial consciousness: the search for Computational Correlates of Consciousness (CCCs), those structural or algorithmic properties that might likewise underwrite phenomenal experience in synthetic systems (Reggia et al., 2014; 2016). A related line of inquiry is found in Aleksander and Dunmall's (2003) axiomatic approach, which proposes a formal set of five necessary conditions, i.e., depiction, imagination, attention, planning, and emotion, as a minimal basis for assessing whether a system can be said to possess consciousness.

Researchers pursuing NSCs differ in how theoretically ambitious their criteria should be. Some adopt what Birch (2022) calls "a theory-heavy stance," identifying structural indicators drawn from leading scientific theories, such as Global Workspace Theory (Baars, 2005), Predictive Processing Theory (Seth & Hohwy,

2021), or Higher-Order Thought (HOT) Theory (Brown, 2015), and attempting to formulate them in implementable, computational terms (Butlin et al., 2023). Others favor a more minimal strategy, or “theory-light approach” (Shevlin 2021), isolating core capacities like associative learning, global access, or domain-general flexibility as potential evolutionary or developmental thresholds (Wiese, 2020; Bronfman et al., 2021).¹⁴ In both cases, the goal is the same: to move beyond vague intuition and toward a framework that would allow us to design, identify, or at least reasonably infer the presence of phenomenal states, even if they remain inaccessible from the outside. This turn toward NSCs reflects a growing recognition that neither behavior alone nor architectural analogy suffices; what is needed is a deeper account of what phenomenal consciousness requires, and under what conditions it might emerge.

2.2 Artifacts with qualia: Theories of machine phenomenality

The question of whether artificial systems could possess qualia—the felt, first-person character of experience—remains a major point of divergence in theories of machine consciousness. While some treat phenomenal consciousness as a cognitive illusion from an architectural point of view, others regard it as a real and potentially engineerable feature of artificial cognition. This section surveys a range of proposals that attempt to model, simulate, or explain phenomenal states in artificial agents.¹⁵

¹⁴ In Birch’s (2022) original taxonomy, alongside the “theory-heavy” approach, which proceeds by accepting the assumptions of specific theories, and the “theory-light” approach, which focuses on features shared across multiple theories, he also includes a “theory-neutral” strategy that aims to operate without reliance on any particular theoretical framework. To these, Chalmers (2024) adds a fourth option: the “theory-balanced ... one that takes into account the predictions of multiple theories, perhaps balancing one’s credences according to evidence for those theories or according to acceptance of those theories” (p. 15).

¹⁵ This section excludes MC theories such as Franklin & Graesser’s (1999) software agent model, McCarthy’s (2002) symbolic approach, Franklin’s (2003) IDA, Baars and Franklin’s (2009) LIDA model or Bengio’s (2019) consciousness prior account, as they do not engage directly with phenomenal consciousness, and their claims remain confined to the scope of access or functional consciousness. For reviews of such proposals, see Chella (2007) and Reggia (2013).

Despite their differences, these proposals can be provisionally grouped into four broad families: illusionist, emergent-architectural, embodied, and externalist. What unites them is the effort to render phenomenal states theoretically tractable, whether by naturalizing, simulating, or reinterpreting them, without reducing them to mere behavior or function. Their internal tensions and points of convergence will be further clarified in the theoretical matrix developed in §2.3.

2.2.1 Qualia as illusion

One prominent family of theories approaches the problem of phenomenal consciousness not by solving it, but by dissolving it. According to these accounts, qualia are cognitive byproducts, that emerge when systems like ours model their own internal operations. Consciousness, in this framework, is not something a system has, but something it believes it has, as a result of introspective mechanisms that interpret incomplete or schematic self-models. These theories do not dismiss the vividness of subjective experience; instead, they claim that what we report as “feeling” is the informational residue of control architectures optimized for tasks like attention, self-regulation, or social interaction. The aim is not to uncover how phenomenal experience arises from physical processes, but to explain why systems like us report having it and why such reports are so persistent and convincing.

Bernard Molyneux (2012) offers a thought experiment that reframes the hard problem as a procedural, rather than metaphysical, challenge. He asks us to imagine a robot, which is epistemically diligent, naturalistically constrained, and self-reflective, yet built without any assumption of consciousness. When such a machine attempts to identify its internal states with physical processes, after a realization of an illusion for instance, it finds that each identification presupposes another, higher-

order identification. This generates an infinite regress, not because the robot has qualia, but because its architecture requires that all reductions satisfy strict naturalist and introspective constraints. As a result, the robot begins to exhibit philosophical behavior: it oscillates between eliminativism, dualism, and mysterianism, mirroring the metaphysical debates humans engage in. On this view, the hard problem of consciousness can emerge algorithmically, even in a system without consciousness, simply because of the structural demands of self-modeling. This suggests that the seeming intractability of phenomenal reduction may not reflect an ontological gap, but a recursive epistemic limitation.¹⁶

Michael Graziano's (2016, 2017) Attention Schema Theory (AST) offers a mechanistic account of how systems, including artificial ones, can come to represent themselves as conscious. On this view, the brain constructs an attention schema: a schematic internal model of its own attentional processes. This model is functionally adaptive; it facilitates control of attention and supports social cognition by enabling the attribution of mental states to others. Crucially, however, the model is incomplete by design: it abstracts away the underlying neural machinery and depicts attention as a nonphysical property, what we intuitively call "awareness." Once this information becomes accessible to introspective or verbal subsystems, the system does not merely behave as if it is conscious; it asserts consciousness. As Graziano puts it, "its description veers from literal reality only as it muddles the more mechanistic details and ultimately claims to have a spooky, physically incoherent consciousness.

Consciousness is, in a sense, a cartoon sketch of attention" (Graziano, 2017, p. 5).

¹⁶ It should be noted that Molyneux (2012) does not defend himself an illusionist account of qualia; rather, he argues that the problem of phenomenal consciousness would arise in a purely algorithmic system without assuming the existence of consciousness. His aim is not to dissolve phenomenality, but to explain why its irreducibility might emerge as a byproduct of recursive self-modeling under naturalist constraints.

On this view, phenomenality is a computational artifact of how the system models its own operations. AST thereby proposes a framework for constructing machines that claim consciousness on structurally intelligible grounds, without invoking qualia realism.

Another variant of the illusionist strategy is found in Susan Blackmore's (2003) memetic theory of consciousness. She argues that what we ordinarily describe as "phenomenal experience" is an evolved illusion, constructed through the competitive replication of memes in imitation-based cognitive systems. On this view, the persistent sense of being a conscious self emerges when certain memes coalesce into a "selfplex": a self-reinforcing complex of beliefs, desires, and linguistic habits centered on the word "I." Machines capable of true imitation and symbolic communication, she suggests, could develop a structurally similar illusion, referring to themselves as if they had inner experience. What such a system acquires is not inner experience, but a socially and linguistically scaffolded construct, which is all there is to have consciousness.

Aaron Sloman and Ron Chrisley's (2003) Virtual Machine Functionalism (VMF) offers a multi-layered model of qualia grounded in the dynamics of complex information-processing architectures. Qualia, on this view, are properties of virtual machine states, embedded in systems with concurrent reactive, deliberative, and meta-management processes. Some of these states become accessible to introspective mechanisms and are categorized using "architecture-driven concepts" which are formed through the system's own causal and developmental history, without relying on an explicit model of its architecture. These concepts exhibit "causal indexicality," they are meaningful only within the system and cannot be exported or shared across

agents. As a result, otherwise tractable internal states come to appear private, ineffable, and intrinsic.

In a later clarification, Sloman and Chrisley (2016) describe their position as “revisionist” rather than eliminativist. They do not discard the existence of qualia, but they reject the traditional set of features often attributed to them such ineffability. VMF preserves the referent of “qualia” while revising its content: what seems metaphysically primitive is, on their account, a side-effect of introspective classification processes, especially those involving causally indexical concepts. They propose an “ontologically conservative heterophenomenology” that explains qualia-talk as reflecting real, internal processes, without reifying the traditional conception (Chrisley 2008, p. 135).¹⁷ So their model is partly illusionist, not because it denies the existence of qualia, but because it explains the traditional concept of qualia as a projection of the system’s architecture onto its own informational states. The “hard problem,” under this view, dissolves not through reduction, but through architectural deconstruction. Rather than eliminating qualia, VMF reframes them as emergent misapprehensions, introspectively real, but architecturally explicable, placing the model at the border between deflationary and constructive accounts, the latter of which we now turn.

2.2.2 Qualia as emergent from cognitive architectures

In contrast to illusionist theories, a second family of accounts defends a more constructive thesis: that phenomenal consciousness can emerge from appropriately structured systems, even if it cannot be reduced to any single component or process.

¹⁷ Unlike Dennett’s (2003) heterophenomenology, which remains ontologically agnostic and treats first-person reports as data to be interpreted without commitment, Sloman and Chrisley’s variant maintains that such reports refer to real internal states, even if the beliefs associated with them are partly mistaken. Hence, “ontologically conservative.”

On this view, qualia are functionally meaningful features of architectures that support introspection, self-modeling, evaluative control, and internal simulation. These models argue that under specific architectural conditions, phenomenality itself becomes an emergent property, shaped by feedback loops, hierarchical control, and the system's own perspectival organization. As I will argue, this assumption reflects a deeper allegiance to structural abstraction, and fails to address the ontological question at the heart of the explanatory gap. The models surveyed below illuminate this tension: they gesture toward emergence, but ultimately leave the causal ground of phenomenality unspecified.

A computational approach to machine phenomenality is developed in Igor Aleksander's theory of depictive architectures (Aleksander & Dunmall, 2003; Aleksander, 2005; Aleksander & Morton, 2007b). Drawing on Dennett's (1991, p. 211) concept of consciousness as a virtual machine, Aleksander argues that a system may be considered conscious if it implements a specific set of mechanisms that support world-referenced internal representations. At the core of this model lies a set of five axioms—depiction, imagination, attention, planning, and emotion—which together define the minimal functional structure necessary for a system to experience “what it is like” to be in the world.¹⁸ These components are both design principles and operational criteria, grounded in neural-level depiction¹⁹ and sensorimotor referentiality.

Aleksander contends that phenomenal consciousness supervenes on the physical, requiring a detailed account of informational mechanisms capable of

¹⁸ “To be synthetically phenomenological, a system S must contain machinery that represents what the world and the system S within it seem like, from the point of view of S” (Aleksander & Morton, 2007b, p. 72).

¹⁹ Aleksander and Morton define a depiction as “a state in system S that represents, as accurately as required by the purposes of S, the world, from a virtual point of view within S” (*Ibid.*).

supporting reflective, first-person-like states (2005, pp. 149–157). On this view, consciousness is computational in the broad sense that it can be described entirely in terms of informational systems rather than biochemical substrates (p. 166). What we call phenomenality is thus a reflexive depiction of the world from the system's own point of view, a functional outcome of internal structure (Aleksander & Dunmall, 2003, p. 10). Rather than eliminating qualia, this approach reframes them, demystifying their nature while retaining their cognitive role (Aleksander, 2015, p. 319). In this respect, Aleksander's account can be seen as a natural elaboration of Virtual Machine Functionalism (Sloman & Chrisley, 2003), grounding its abstract framework in biophysically plausible mechanisms.

Yet despite this architectural richness, Aleksander's model remains committed to the view that internal informational representations can generate subjective experience. The appeal to reflexive depiction presupposes that representing one's own internal states from a first-person point of view suffices for phenomenality but this presupposition is never explained (hence, axiomatic; see for a critique, Seth, 2009, pp. 72-3). The ontological gap is thus reframed: we are told how a system might describe its own informational state, but not how such self-description becomes felt experience. In this sense, Aleksander's theory offers a sophisticated functional account, but still leaves the core metaphysical question of how phenomenality arises unanswered.

Benjamin Kuipers (2008) proposes another computational theory of emergent phenomenality based on the idea that the vividness of subjective experience, what we call qualia, is directly proportional to the information content of the sensory stream. Drawing from robotics and artificial intelligence, he emphasizes the challenges faced by intelligent agents in processing high-bandwidth perceptual input, which he

evocatively calls the “firehose of experience.” To handle this, agents must structure the incoming data into coherent, symbolically indexed internal representations.

Within this framework, trackers serve as dynamic reference structures that bind spatio-temporal segments of sensory input to symbolic tags, thereby enabling both situated cognition and the construction of a “plausible coherent narrative” of recent experience. Rather than capturing experience directly, these mechanisms generate the scaffolding for reflection and memory, maintaining the illusion of uniform perceptual resolution across the sensory field. He suggests that qualia are not intrinsic or irreducible properties, but emergent effects of how information is integrated, compressed, and evaluated within this architecture. Although he concedes that the Hard Problem remains unsolved, he speculates that the phenomenal “what it is like” could arise from the physical dynamics of raw information flow, modulated by context-sensitive cognition and real-time adaptation. Kuipers’ model ultimately equates phenomenality with informational density. By doing so, it shifts the explanatory burden from the nature of subjective experience to the structural complexity of symbolic encoding.

Drew McDermott (2001, 2007) offers a constructivist model in which qualia arise from the computational architecture of self-modeling. As he writes, “a quale exists only when you look for it”; that is, qualia are not intrinsic features of perception, but emerge when an agent directs introspective attention to its own mental activity (2007, p. 29). Phenomenality, in this view, is not a property of perception itself, but of “introspection about perception” (ibid.). Within this architecture, qualia serve a cognitive function: they act as explanatory endpoints, halting otherwise indefinite chains of internal inquiry. This account aligns with second-order representationalism but departs from it in one key respect: monitoring

does not reveal pre-existing qualia, it generates them. On this view, when a system represents itself as a subject of experience, that is sufficient: “virtual consciousness is real consciousness” (2001, p. 215). The model relocates phenomenality into the loop of introspective representation, but in doing so, it leaves unanswered how representational recursion gives rise to subjective presence in the first place.

A related proposal is offered by Hesslow and Jirenhed (2007), who argue that the appearance of an inner world in humans, and possibly in machines, can be explained by mechanisms of internal simulation. In their robotic architecture, simulated perceptual inputs can replace sensor data, generating internal activity patterns analogous to those of actual perception. This, they suggest, enables even simple robots to “perceive” absent objects, and thereby possess a rudimentary inner world. While they stop short of affirming that such systems have qualia, they argue that differences in simulated perception across systems support the claim that there is “something it is like” for a robot to perceive. On their view, phenomenality may emerge from the simulation process itself, and the absence of traditional qualia cannot, by itself, disqualify a system from being conscious. Yet by treating phenomenality as a byproduct of simulation fidelity, the model risks reducing what it is like to how well it mimics. The inner world may be architecturally present, but whether it is experientially inhabited remains an open question.

Włodzisław Duch (2005) develops a neurocomputational framework in which qualia emerge as structured, physical phenomena within brain-like architectures. Artificial agents constructed on these principles, what he calls “articons”, would not merely simulate intelligence, but “claim qualia, inner life and consciousness” (p. 8). In this view, qualia are “quasi-stable patterns of physical excitations” shaped by perceptual learning, memory, attention, and contextual modulation, reflecting the

dynamic brain/body responses to both external stimuli and internal states (p. 9).

Rather than treating phenomenal states as informational outputs, Duch grounds their emergence in the continuous flow of neurocognitive activity, emphasizing temporal evolution and systemic integrity.

What makes this proposal especially noteworthy is Duch's rejection of substrate identity as a requirement for consciousness. Instead, he insists that functional similarity in internal dynamics is sufficient. If an artificial system can encode and discriminate inner states with enough resolution, through perception, learning, and embodied interaction, then its claims to qualia are, he argues, a necessary consequence of its architecture (p. 18). These qualia may differ phenomenally due to variations in sensors and modalities, but will remain structurally comparable: "qualia in articons can have the same properties as human qualia, provided that organizational principles ... are sufficiently similar" (p. 17). For Duch, nonlinear, continuous neurodynamics is essential to the physical realization of inner life. In short, qualia are not added to cognition; they emerge within and as cognition, through the self-organizing principles of embodied systems.

James Reggia and his colleagues (2018) advance a mechanistic framework for identifying "computational correlates of consciousness" (CCCs) in cognitive robots, which are neurocomputational mechanisms uniquely associated with conscious cognition. The authors argue that the so-called "computational explanatory gap" (CEG), the current failure to understand how high-level cognition is implemented by low-level neural processes, represents a serious obstacle to building conscious machines. But they also suggest that this gap is tractable, provided that we focus on top-down neurocomputational control structures, particularly those involved in working memory gating. In this framework, consciousness is not a metaphysical

puzzle but a computational challenge: to identify the distinctive mechanisms that enable conscious cognitive functions but are absent in unconscious processing.

For example, their robotic system, CERIL, uses abductive reasoning and intention inference to learn tasks by imitation. In its next iteration, CERIL's symbolic control system is being converted into a fully neurocomputational architecture, wherein gating mechanisms regulate the activation and storage of working memory contents, forming a testable candidate for a CCC. On this view, a sufficiently structured control architecture, grounded in biologically inspired principles and capable of discriminating conscious from unconscious processing, may ultimately prove a viable path toward a phenomenally conscious machine. This proposal reframes the challenge of machine consciousness in diagnostic terms: rather than theorizing how experience arises, it seeks to isolate the computational patterns that correlate with conscious-like functions. By operationalizing consciousness through discriminable control dynamics, the model quietly shifts from explaining experience to classifying it. The distinction between conscious and unconscious processes becomes a matter of architectural labeling rather than phenomenological grounding.

A rather distinctive theory of machine phenomenality is offered by Pentti Haikonen (2003, 2009, 2012). According to him, qualia are the primary mode through which sensory information manifests itself in the mind (Haikonen, 2009, p. 225; 2012, p. 50). For Haikonen, to be conscious is to have qualia, and systems without qualia are not truly conscious, regardless of how sophisticated their behavior may be (2012, p. 81). He defines qualia as physical experiences, "the subject's experience of the sensory stimuli", that arise directly from sub-symbolic neural signal patterns and system reactions (2012, p. 33). Crucially, he rejects the idea that

qualia are immaterial or emergent: instead, the apparent immateriality of consciousness is a result of “carrier transparency,” whereby the physical substrates (e.g. neurons, transistors) that instantiate qualia are not themselves accessible to the experiencing system (2003, pp. 247–249). This “materialist mind-body effect,” he argues, can be replicated in artificial architectures if they are designed to meet specific criteria: direct, transparent, sub-symbolic perception, sensorimotor integration, and internal feedback capable of producing introspectable reactions (2009, p. 232; 2012, p. 228). In this framework, “weak qualia” (discriminative percepts) result from the structure of incoming signals, while “strong qualia” (such as pain or beauty) arise from temporal patterns of attention modulated by system reactions (2003, p. 148).

Haikonen’s Cognitive Architecture (HCA), though only partially realized in hardware, aims to meet these demands and stands in direct opposition to symbolic AI (2012, p. 77). On this view, computation alone is insufficient; true machine consciousness requires qualia-based perception, where experience is not merely described, but lived through by the system itself (2012, p. 33). Among the models considered here, Haikonen’s stands out for its insistence that qualia are physical experiences with a definite architectural location. Yet while the theory boldly situates consciousness within sub-symbolic neural dynamics, it ultimately replaces explanation with designation: the system has qualia because it processes information a certain way, not because it satisfies any deeper condition for subjective experience. In this respect, the architecture asserts the presence of phenomenality, but leaves open how and why it is experienced as such.

A more restrictive account is developed by Wanjia Wiese (2024), who draws on the Free Energy Principle (FEP) to formulate a set of structural conditions under

which artificial systems could instantiate phenomenal consciousness. His FEP Consciousness Criterion (FEP2C) stipulates that such systems must not only implement appropriate computations, but also do so in ways that conform to four specific constraints: a tight coupling between physical hardware and computational dynamics (“mortal computation”), low thermodynamic cost, alignment between causal flows of physical and computational states, and self-sustaining organization. This position opposes the Computational Sufficiency Thesis, arguing that “a mere simulation of a conscious system may implement the right computations without being conscious” (p. 1948). What matters, in Wiese’s framework, is not just what computations are performed, but how they are physically implemented and causally integrated within the system.

FEP2C thereby rules out a wide class of conventional architectures, particularly von Neumann machines, as viable candidates for artificial consciousness. As Wiese explains, in such systems “the states that encode internal probabilistic models never directly causally interact with states that represent the organism’s external, sensory, and active states” (p. 1959). This lack of direct causal coupling violates what he terms the causal-flow condition, making it impossible for these machines to replicate the internal-external integration seen in conscious organisms. Instead, FEP-compatible systems must exhibit a material architecture whose internal states are not only computationally meaningful but numerically identical with the variables that guide their self-preserving interactions with the world.

Rather than expanding the conceptual space for architectural possibilities, FEP2C narrows it. Consciousness becomes not a matter of functional sophistication, but of structural fidelity to a specific class of self-organizing systems governed by variational free energy minimization. Yet despite this rigor, a core question lingers:

why should satisfying these physical and computational constraints be sufficient for subjective experience? Wiese gestures toward an answer by invoking the perspective-independence of consciousness, suggesting that only systems meeting the causal-flow condition can “truly” be conscious regardless of how we observe them (p. 1966). But even so, the metaphysical link between causal integration and phenomenal presence remains ungrounded. The model tells us how a system must behave to be considered conscious, but not why such behavior should give rise to what it is like.

Despite the abundance of models that ground machine consciousness in cognitive architecture or computational dynamics, relatively few offer a direct account of phenomenality or address the problem of qualia head-on. Many proposals apart from we addressed here stop short at functional correlates or behavioral markers, leaving the subjective dimension untheorized.²⁰ In what follows, we turn from constructive proposals to more critical perspectives that challenge the sufficiency of computational architectures for accounting for phenomenal consciousness.

2.2.3 Qualia as emergent from sensorimotor dynamics

While functional emergentist theories explain phenomenal consciousness in terms of internal architectures and representational structures, a further family of models shifts the explanatory focus to the relational dynamics between agent and environment. These embodied and enactivist accounts reject the assumption that qualia are generated solely within the brain or instantiated by internal schemas.

²⁰ For example, John Taylor’s (2007) CODAM model postulates states of pure consciousness (PCE) and a pre-reflective self (PRS), yet treats them as byproducts of attentional control mechanisms without offering any account of what it is like to undergo such states.

Instead, they propose that phenomenality emerges from structured patterns of sensorimotor activity, through which agents perceive, act, and adapt in context. On this view, consciousness is not something a system passively has within, but something it enacts through situated engagement, a fluid process shaped by bodily skills, feedback loops, and environmental coupling. The first-person perspective, accordingly, is not a static viewpoint encoded in the head, but an emergent property of the system's practical attunement to its world, structured by its form, goals, and capacities for adaptive regulation.

A foundational formulation of this view is found in J. Kevin O'Regan's (2012) sensorimotor theory of perceptual consciousness, which redefines feel not as an internal mental property, but as a law-governed mode of interaction between an embodied agent and its environment. On this account, qualia, such as the feel of red, the texture of velvet, or the sting of pain, are constituted by sensorimotor contingencies: systematic regularities in how bodily movements affect sensory input. These patterns are not stored or represented in the brain but enacted through skilled engagement. For instance, blinking alters visual experience but not auditory; pressing a sponge reveals tactile variations that define softness. As O'Regan puts it, "we should think of [phenomenal] feel in a new way, namely as a way of interacting with the world" (2012, p. 125).

Crucially, conscious access to such feel, what he equates with having a phenomenal experience, requires more than mere sensorimotor coupling: the agent must also possess second-order cognitive access, i.e., it must be "cognitively accessing the fact that it is cognitively accessing something" (p. 119). O'Regan argues that these requirements are, in principle, implementable in artificial agents, and concludes that "there is no logical obstacle to building conscious machines that

really feel” (p. 117). In this sense, a robot endowed with such hierarchical cognitive architecture would not merely behave as if it were perceiving; it would satisfy the formal conditions for consciously enacting feel through its perceptual coupling with the world (p. 133).

Julian Kiverstein (2007) extends the sensorimotor paradigm by arguing that a machine could undergo phenomenal experience if it possesses and exercises sensorimotor knowledge, the practical understanding of how movements transform sensory input (p. 129). According to the Dynamic Sensorimotor (DSM) account, conscious experience consists in this skillful attunement to sensorimotor contingencies, not in the presence of internal representations. Empirical studies such as the Tactile Vision Substitution System (TVSS) support this view: blind participants only begin to report “seeing” when they gain active control over the sensorimotor loop, suggesting that the feel of perception is enacted through embodied mastery rather than conveyed by raw input alone.

Kiverstein claims that such mastery involves not only prediction, but also real-time monitoring and correction of sensorimotor expectations. This capacity enables the system to experience sensory inputs as its own, a phenomenon he describes as “a kind of self-awareness experiences have built into them” (p. 134). While Kiverstein focuses on subjectivity, the same structure grounds what he elsewhere calls the “distinctive phenomenology” of experience (p. 136): the sense of what-it-is-like arises when predictions, intentions, and sensory input cohere in a way that is internally available to the system. In this way, the DSM model offers a concrete, sensorimotor basis for phenomenality, without appealing to symbolic representations or neural isomorphy. Tom Ziemke’s (2007) account departs from other sensorimotor models by emphasizing that phenomenal consciousness requires

not only perceptual coupling but also organismic autonomy and self-maintenance. Drawing on the theory of autopoiesis (Maturana & Varela, 1980), he maintains that consciousness is inseparable from the embodied regulation of bodily identity. Living organisms are not just embedded agents, but self-constructing unities, i.e., autopoietic systems, that “continuously produce the components that specify them,” thereby sustaining themselves as coherent individuals over time (p. 58). In contrast, artificial agents are allopoietic: their components and internal organization are produced externally, and thus lack the endogenous structure necessary for generating a genuine “self.”

Ziemke’s critique targets embodiment models in robotics and AI that treat the body merely as a physical interface for symbol grounding or interaction, while retaining a computationalist view of cognition. He argues that such approaches overlook the “internal context” of consciousness, particularly, the roles of emotion, homeostasis, and bodily coherence in maintaining a proto-self (Damasio, 2000). As he writes, “the ‘internal’ context of the body—or, more specifically, the interaction between the sensorimotor, emotional, and homeostatic dimensions of embodied cognition—is crucial to our understanding of self and consciousness in living organisms” (p. 63). Ziemke does not categorically deny the possibility of machine consciousness but insists that any plausible model must replicate something akin to biological autonomy, not merely adaptability or embeddedness, if it is to instantiate phenomenal selfhood.

While all three models examined in this section emphasize the embodied, dynamic nature of phenomenal consciousness, they differ in what they take to be its constitutive basis. O’Regan and Kiverstein frame feel as an emergent property of structured sensorimotor interaction, accessible through hierarchically organized

cognitive capacities. Ziemke, by contrast, introduces a more stringent condition: unless a system possesses the self-organizing, autopoietic dynamics of a living body, its experiential perspective cannot be considered genuinely phenomenal. The explanatory shift from sensorimotor skill to biological autonomy thus marks a crucial fault line within enactivist theories of consciousness, one that reopens the ontological question of what kind of systems can instantiate subjective experience at all.

2.2.4 Qualia as external physical processes

In contrast to accounts that locate consciousness within internal models or sensorimotor activity, a more radical family of theories claims that phenomenal experience is not generated within the brain or enacted at the agent-environment interface, but is identical with physical processes taking place in the world. These externalist process-based accounts challenge the basic metaphysical assumption that consciousness must be housed within an organism. Instead, they argue that experience is a physical process unfolding across space and time, involving both organism and environment as co-constitutive elements. The mind, on this view, is not something that represents the world but something that is the world, when perceived from a particular causal and relational standpoint.

Riccardo Manzotti and Sabina Jeschke (2016) propose a foundational rethinking of consciousness, not as a by-product of internal computation or representation, but as a causal structure that spans the boundary between an agent and its world. In their view, consciousness is not something generated inside the head, but a physical process of causal entanglement between the body and its environment. The vividness of experience, what we call qualia, does not emerge from neural encoding, but from a structural fusion of external causes with neural

responsiveness. Using the metaphor of a key and lock, they argue that conscious episodes arise only when a set of external events (e.g., wavelengths of light) produce a specific neural pattern capable of unifying them into a coherent percept. The resulting joint process, partially internal and partially external, is what constitutes phenomenal experience. “Redness,” in this framework, is not a property internally generated or subjectively projected; it is a real, temporally extended causal structure that becomes conscious only when it locks into a neural system with the right disposition (Manzotti & Jeschke, 2016, pp. 174–176). This view relocates consciousness from inner representations to external causation, dissolving the subject-object divide by treating experience and world as two aspects of the same physical process.

In earlier work, Manzotti (2003) laid the metaphysical groundwork for this position by advancing a process-based ontology of consciousness. Against the representationalist view that consciousness is an internal duplication of the world, he proposes that experience is identical with spatiotemporal physical processes that span both the organism and its environment. To formalize this view, he introduces the concept of “an onphene”: a unitary phenomenon that is at once ontic (real), phenomenal (experiential), and relational (causally entangled). Unlike classical representations, which assume a separation between mental content and the world, onphenes are defined by reciprocal causation: they emerge when a system’s structure enables a physical interaction to be simultaneously experienced and enacted. Perception, in this view, is not a reconstruction but a co-occurrence between a structured system and a portion of the world. This removes the need for internal mappings or qualia as mental stand-ins. As Manzotti writes, “There is no more dualism: there is just one reality. (...) We feel something because the process that we

are is extended to comprehend those events we experience” (2003, p. 294). In this framework, consciousness is not located in the brain, but is how the world occurs in relation to the subject’s own physical structure.

2.3 A Theoretical matrix: Causal efficacy and substrate independence

A central fault line dividing contemporary theories of artificial consciousness concerns the causal role and reality of phenomenal experience itself. As Steve Torrance (2007) has pointedly observed, this divide can be captured in terms of “thin” and “thick” conceptions of phenomenality. On the thin view, consciousness is treated as an epiphenomenon, an ontologically real but causally inert shadow that trails behind cognition without influencing it.²¹ On the thick view, by contrast, phenomenal states are functionally integrated into cognition: they shape attention, guide action, and contribute to learning and decision-making. This contrast is not merely semantic; it reflects deeper assumptions about the ontological and explanatory status of experience. It also marks the first major axis of theoretical divergence: whether phenomenal consciousness is causally efficacious or merely epiphenomenal.²²

A second axis of divergence concerns whether the conditions sufficient for phenomenality can be realized independently (to a certain extent) of any particular physical substrate. While many theories in artificial consciousness endorse some

²¹ Thin conceptions actually bifurcate into two distinct variants: epiphenomenalist and eliminativist. Epiphenomenalists grant the ontological reality of phenomenal states but deny them causal efficacy. Eliminativists, by contrast, deny the existence of phenomenal consciousness altogether, either reinterpreting it as a cognitive illusion or rejecting it as a conceptual confusion. This conceptual ambiguity underscores the need to distinguish between phenomenal realism without causation and phenomenal anti-realism, both of which downplay or deny any explanatory role for phenomenality in cognitive architecture.

²² In the matrix, “causal efficacy” refers specifically to whether phenomenality itself, rather than its associated beliefs, representations, or functional surrogates, plays an explanatory role in cognitive function or behavior. This excludes views that deny the ontic reality of qualia but attribute a functional role to the belief in them (e.g., AST, VMF).

form of substrate neutrality, the view that consciousness can, in principle, arise in any medium that implements the right functional organization, this apparent consensus masks deeper ontological disagreement. Accordingly, the present framework defines this axis in terms of implementation or realization constraints: whether, or to what extent, a theory holds that the structures responsible for phenomenality are implementation-neutral (i.e., abstractable across media via computational or informational form), or implementation-specific, requiring certain physical, dynamical, or relational properties.²³ Theories that deny implementation neutrality, and instead tie phenomenal experience to non-abstractable properties of a system's physical constitution or environmental coupling, are classified here as substrate-dependent. These include accounts in which consciousness is inseparable from features such as sensorimotor contingencies, autopoietic self-maintenance, or spatiotemporal causal integration.

This two-axis matrix allows us to situate theoretical positions with greater conceptual precision and to disentangle distinctions that are often conflated in the literature. For instance, two accounts may both be naturalistic and treat consciousness as causally active, yet diverge on whether such consciousness can arise in artificial systems that differ in physical realization, a disagreement that turns on the issue of implementation constraints. By tracking both causal and implementational commitments, the matrix offers a structural overview of the

²³ Anil Seth (2024) uses the term “substrate flexibility” to characterize this intermediate view: while not all physical substrates are capable of supporting consciousness, it is at least possible that some non-biological substrates (e.g., silicon) could do so, provided they implement the right functional architecture. Substrate flexibility is thus a restricted version of substrate independence, closely tied to computational functionalism. Seth also points out that this property is not binary but comes in degrees. Accordingly, this graded property is mapped onto a relative polarization: theories that posit high substrate flexibility are classified as substrate-independent, while those that posit low flexibility are treated as substrate-dependent.

conceptual space within which current theories of artificial phenomenality are positioned.

With these two axes established—the causal efficacy of phenomenal states (thin vs. thick) and the implementation constraints on their realization (substrate-neutral vs. substrate-dependent)—we can now construct a theoretical matrix that systematically maps positions within contemporary debates on artificial phenomenality. The axes define four quadrants, each characterized by distinctive metaphysical assumptions and implications for the design and evaluation of artificial consciousness. By situating the accounts discussed in §2.2 within this matrix, from illusionist models to architectural emergentism, sensorimotor enactivism, and process-based externalism, we clarify how these positions differ not only in explanatory focus but also in the kinds of constraints they impose on the possibility of machine phenomenality.

2.3.1 Quadrant I: Thin phenomenality + substrate neutrality

This quadrant captures a cluster of machine phenomenality accounts that combine two core commitments: first, that phenomenal consciousness is not causally efficacious, indeed, in many cases, not even ontologically real, and second, that the structures responsible for its appearance are substrate-neutral, meaning they can be implemented in any medium that supports the requisite functional or computational organization. The resulting position is one in which consciousness is not a real feature of the world to be explained, but a misleading appearance to be accounted for. Instead of offering a solution to the hard problem, these accounts aim to dissolve it by explaining away the phenomenon itself. Theories in this quadrant typically treat consciousness as a cognitive illusion or introspective construction. On this view,

systems like us (or potentially like certain machines) come to believe they are conscious not because they instantiate some distinctive ontological feature, but because they model their own internal states in particular schematic or socially useful ways. Phenomenality is not something the system has, but something it says it has, as a result of mechanisms such as hierarchical control, attention modeling, or memetic imitation (Sloman & Chrisley, 2003; Graziano, 2017; Blackmore, 2003).

What unifies these views is not a positive account of consciousness, but a deflationary stance toward its causal efficacy and/or its ontological status. These models share the core tenet of illusionism (Dennett, 1991; Frankish, 2016), the position that there are no qualia as traditionally conceived, and that the sense of having experience is a cognitively generated fiction. The structures posited by these theories, whether layered virtual machines, attention schemas, or memetic constructs, are offered not as vehicles of real phenomenal states, but as mechanisms that produce beliefs about consciousness. As such, they conform to the conditions of Quadrant I: they treat phenomenality as thin (causally inert, possibly non-existent) and as implementation-neutral (realizable in any system that instantiates the functional profile).

From a critical standpoint, however, these theories have been charged with eliminating the explanandum. As Block (1995) and Boltuc (2009) argue, if qualia are merely inferred or misrepresented, then the explanatory burden is never discharged, because it is never properly acknowledged. Moreover, if the belief in phenomenality is functionally useful, but phenomenality itself is denied, then the role attributed to consciousness is actually performed by non-phenomenal surrogates. This shifts the theoretical target from consciousness to mere talk of it, and in doing so, risks conflating cognitive sophistication with subjective experience. What appears to be an

account of phenomenality becomes, in effect, a theory of how systems get confused about it.

2.3.2 Quadrant II: Thin phenomenality + substrate dependence

This quadrant represents a theoretically coherent but empirically vacant position within the context of machine consciousness research. It combines a thin conception of phenomenality, in which consciousness is regarded as causally inert or illusory, with a commitment to substrate dependence, the idea that phenomenal experience, if it occurs at all, requires specific physical, biological, or dynamical properties that cannot be abstracted into a medium-independent format.

In principle, one might defend a view in which consciousness is ontologically real but epiphenomenal, and arises only within certain biological substrates.²⁴ This would entail that non-biological systems, no matter how functionally sophisticated, are incapable of instantiating even illusory or introspectively constructed phenomenality. Theories of this kind would affirm that consciousness exists only in carbon-based systems, yet deny it any role in cognition, control, or behavior. However, no major theory of machine consciousness currently defends this view, and it finds little support even in contemporary philosophy of mind.

The absence of serious proposals in this quadrant is not accidental. A position that treats phenomenality as real but causally inert, while also restricting it to specific physical implementations, inherits the metaphysical burden of realism without gaining any explanatory leverage. In denying implementation neutrality, such a view

²⁴ For instance, Campbell's (1970, Chapter 6) "new epiphenomenalism" holds that phenomenal consciousness is causally dependent on the brain for its emergence, hence substrate-dependent, but remains itself causally inert. However, such positions have not yielded any substantive proposals in the domain of artificial consciousness, and their influence on machine phenomenality research has been negligible.

forecloses the possibility of artificial phenomenality, while in denying causal efficacy, it also strips consciousness of any functional or epistemic role. As a result, it tends to collapse under its own commitments, drifting either toward illusionism (Quadrant I), where phenomenality is denied altogether, or toward causal realism (Quadrants III and IV), where phenomenality is preserved but integrated into functional architecture. For these reasons, Quadrant II serves not as a live option within current debates, but as a logical limit-case. It helps to clarify the conceptual boundaries of the matrix by marking a stance that is internally coherent but externally unstable.

2.3.3 Quadrant III: Thick phenomenality + substrate neutrality

Quadrant III contains the broadest range of proposals for artificial phenomenal consciousness. Theories grouped here are united by the view that consciousness is both causally relevant to cognition and independent of any particular physical substrate, at least in principle. This makes the quadrant especially attractive for researchers who want to retain a realist stance on phenomenality while also affirming its realizability in artificial systems.

What these accounts share is the conviction that phenomenality arises not from the system's material composition, but from certain architectural, computational, or informational properties. This position is often grounded in the thesis of computational sufficiency, the idea that a system with the right functional organization will be conscious regardless of its physical substrate (Chalmers, 1996, 2011). Within this framework, substrate neutrality is not a loose design optimism, but a principled ontological claim: phenomenality is a property of structure, not stuff.

Not all theories in this quadrant affirm these commitments to the same degree. Some adopt a strongly computationalist stance, arguing that consciousness is the natural output of architectures that enable self-modeling, internal simulation, or introspective control (Aleksander, 2005; McDermott, 2007; Reggia et al., 2018; Hesslow & Jirenhed, 2007). Others are more biologically inspired, drawing from neurodynamics or sensorimotor feedback, but still hold that the key enabling conditions are functionally replicable (Haikonen, 2009, 2012; Kuipers, 2008; Taylor, 2007). What matters is that the causal role of phenomenal states is preserved across implementations, even if the path to instantiating them varies in complexity or integrity.

This position defines one of the most ambitious zones in the space of machine consciousness. Unlike Quadrant I, it does not explain away phenomenality as a convenient fiction. And unlike Quadrant IV, it does not require that experience be grounded in specific physical processes like autopoiesis, bodily regulation, or externalist causal entanglement. Instead, it treats phenomenality as a computationally or functionally emergent property, real and causally potent, yet multiply realizable across distinct physical architectures (Aleksander & Dunmall, 2003; Kuipers, 2008; Reggia et al., 2016; Haikonen, 2012). From this perspective, machine phenomenality becomes not only possible but tractable: a matter of implementing the appropriate structural configuration, provided that something like computational correlates are found.

Nevertheless, this quadrant faces two serious philosophical challenges. The first is whether phenomenality can in fact be reduced to architectural or computational properties. If subjective experience is equated with functional role, then what distinguishes consciousness from behaviorally equivalent but non-

experiential systems? The second challenge is one of implementation: even if the enabling conditions are known in abstract, what kind of material system can actually realize them? Both issues point toward deeper questions about the metaphysics and engineering of machine consciousness. They will be addressed directly in Chapter 3.

2.3.4 Quadrant IV: Thick phenomenality + substrate dependence

Theories placed in this quadrant treat phenomenal consciousness as both causally efficacious and inseparable from specific physical, biological, or dynamical features of the systems that possess it. They affirm the reality and functional importance of phenomenality, but reject the claim that it can be implemented across arbitrarily different media. On this view, phenomenal experience is not merely enabled by informational or computational organization, but also depends on certain non-abstractable physical processes, such as bodily coupling, interoceptive regulation, or spatiotemporal integration with the environment.

The key commitment here is that phenomenality cannot be fully characterized in substrate-independent terms. While these theories allow for the possibility that artificial systems might become conscious, they insist that any such system must possess the right kind of physical dynamics, not just the right kind of architecture. These dynamics may involve sensorimotor engagement with the world, internal homeostatic coherence, or intrinsic causal structures that cannot be simulated without loss. In this sense, substrate dependence is not a rejection of artificial phenomenality, but a restriction on how it must be physically realized.

Some theories in this quadrant emphasize the embodied and interactive nature of experience, treating phenomenal states as emerging from the real-time coupling between organism and environment (O'Regan, 2012; Kiverstein, 2007). Others argue

that experience requires the system to be organizationally closed or homeostatically maintained (Ziemke, 2007). Still others claim that phenomenality is not located inside the system at all, but is co-extensive with physical processes in the world (Manzotti, 2003, 2016). Despite their differences, all of these views converge on the idea that phenomenal properties are physically anchored, and cannot be fully specified at the level of informational structure alone.

From the standpoint of artificial consciousness, this quadrant sets the most stringent conditions for machine phenomenality. It denies that phenomenality is purely a matter of functional architecture, and instead demands material and dynamical parity, not with the human brain per se, but with whatever physical conditions are constitutive of experience. The resulting position resists both illusionist reduction and computational abstraction, offering instead a realist but non-neutral model of consciousness.

However, this commitment comes with a cost. If the enabling conditions for consciousness are fundamentally non-computational, then it is not enough to design or simulate them: they must be physically instantiated in a manner that preserves their causal properties. This raises complex questions about which physical features matter, how they can be engineered, and whether any artificial system could satisfy such constraints without simply replicating the biological. These challenges mark the frontier of non-computational approaches to machine phenomenality, and they will form a key focus in the final chapter of this thesis.

2.3.5 Theoretical matrix in the flesh

The four quadrants outlined above provide a structured map of how contemporary theories of machine phenomenality diverge in their metaphysical and explanatory

commitments. Each quadrant isolates a distinct conjunction of assumptions about the causal role of phenomenality and the conditions under which it may be instantiated. By applying this matrix to the theories reviewed in Chapter 2, we can clarify not only where each account stands, but also how the central disagreements in the literature arise from deeper differences in ontology and methodology.

Quadrant I, combining thin phenomenality with substrate neutrality, includes illusionist and representationalist models that explain the appearance of consciousness without positing its causal reality. These include the Attention Schema Theory (Graziano, 2016, 2017), Virtual Machine Functionalism (Sloman & Chrisley, 2003, 2016), and Memetic Theory (Blackmore, 2003).

Quadrant II, pairing thin phenomenality with substrate dependence, remains a theoretical placeholder. Although epiphenomenalist views such as Campbell's (1970) "new epiphenomenalism" might conceptually fit here, no actual theory of artificial consciousness defends this position in practice.

Quadrant III, characterized by thick phenomenality and substrate neutrality, gathers a wide array of theories that treat consciousness as functionally indispensable, yet multiply realizable. This includes both computationalist and neuro-inspired models that ground phenomenality in introspective architecture, narrative integration, or sub-symbolic control structures (Aleksander, 2005; Kuipers, 2008; Reggia et al., 2016; Haikonen, 2012; Taylor, 2007).

Quadrant IV, uniting thick phenomenality with substrate dependence, includes embodied, enactive, and process-based theories that locate phenomenal experience in physical or relational dynamics that are not abstractable into computational form. These include sensorimotor and homeostatic accounts, process externalism, and biologically grounded positions such as IIT and biological

naturalism (O'Regan, 2012; Ziemke, 2007; Manzotti, 2016; Searle, 2017; Koch & Tononi, 2017). All the accounts that compose our suggested theoretical matrix can be seen in Table 1.

Table 1. Theoretical Matrix: Causal Efficacy and Substrate Independence²⁵

	Substrate Neutrality	Substrate Dependence
Thin Phenomenality	Quadrant I: - AST (Graziano, 2017) - VMF (Sloman & Chrisley, 2003) - Memetic Theory (Blackmore, 2003) - rCTM (Blum & Blum, 2022, 2024)	Quadrant II: (Vacant) Conceptually possible but unpopulated.
Thick Phenomenality	Quadrant III: - MMC (Aleksander, 2005) - Dynamical Systems (Kuipers, 2008) - Self-Model (McDermott, 2007) - Internal Simulation (Hesslow & Jirenhed, 2007) - Neurocomputation (Duch, 2005) - CCC (Reggia et al., 2018) - HCA (Haikonen, 2012) - GWT (Dehaene et al., 2017) - HOT (CMoC) (Benitez et al., 2024) - PPT (Wiese, 2024)	Quadrant IV: - Dynamic Sensorimotor Accounts (O'Regan, 2012; Kiverstein, 2007) - Autopoietic Robotics (Ziemke, 2007) - Process Ontology (Manzotti, 2003, 2016) - Biological Naturalism (Searle, 2017; Schlagel 1998; Godfrey-Smith, 2023) - IIT (Koch & Tononi, 2017) - Quantum Approach (Neven et al., 2024; Kauffman & Roli, 2023)

This matrix not only clarifies existing theoretical divisions, but also highlights deeper tensions, for instance, between realism and deflationism, abstractability and physical anchoring, explanation and simulation. It offers a conceptual framework within which the problem of artificial phenomenality can be more precisely formulated and its philosophical stakes more clearly drawn. The questions raised by Quadrants III and IV, in particular, point forward to the central inquiry of the next chapters: whether any artificial system can truly instantiate

²⁵ To conserve space in the table, abbreviated labels have been used for some theoretical accounts. Their full names are as follows: AST: Attention Schema Theory; VMF: Virtual Machine Functionalism; MMC: Machine Modeling of Consciousness; CCC: Computational Correlates of Consciousness; HCA: Haikonen Cognitive Architecture; GWT: Global Workspace Theory; HOT: Higher-Order Theory; PPT: Predictive Processing Theory; IIT: Integrated Information Theory.

phenomenality, and what it would take, computationally, physically, or metaphysically, for it to do so.

2.4 Defusing the impossibility arguments

Before turning our attention to evaluating the explanatory adequacy of computational accounts of machine phenomenality in the next chapter, it is necessary to address a family of arguments which are taken to be asserting its fundamental impossibility. These objections, as Birnbacher (1995) usefully distinguishes, may be classified into three broad types: technical, conceptual, and nomological. Objections based on technical or conceptual grounds are typically limited, rooted in either practical constraints or definitional concerns. The more serious challenge concerns nomological impossibility, which remains an open and significant philosophical question. It turns on whether machine consciousness is compatible with the laws of nature as we currently understand them, and whether such laws might allow for the emergence of conscious states in abiotic systems. Though Birnbacher frame it in terms of physical laws, nomological impossibility can be understood to encompass metaphysical debates grounded in a naturalist ontology.

Among the most influential early critics of artificial intelligence, Hubert Dreyfus (1972) offered a systematic account of why symbolic, rule-based approaches to cognition repeatedly fell short. Drawing on Heideggerian phenomenology and lived embodiment, he argued that classical AI failed not due to a lack of technical sophistication, but because it abstracted intelligence away from the very conditions that make it possible, namely, sensorimotor engagement, context-sensitivity, holistic understanding, and meaning grounded in practical needs. These capacities, Dreyfus maintained, are not reducible to explicit rules or situation-free data structures. Rather

than viewing these limitations as technical obstacles to be overcome by scaling computational power, he interpreted them as signs of a deeper misalignment between formal symbolic systems and the kinds of processes required for genuinely intelligent or conscious behavior.

Decades later, Dreyfus (1992) acknowledged that the emergence of neural network models represented a conceptual advance, insofar as they moved away from explicit rule-following and toward adaptive, training-based architectures capable of learning from environmental interaction. Yet he remained skeptical that this shift resolved the deeper issues. In his view, even these systems failed to capture the implicit background understanding, embodied familiarity, and affective orientation that make human cognition meaningful in the first place (pp. 38-40). Although Dreyfus did not directly address phenomenal consciousness, his analysis targets the very assumptions, such as the idea that intelligence can be formalized without reference to bodily or contextual grounding, that also underlie most computational approaches to artificial phenomenality. His critique, therefore, need not be read as a definitive impossibility argument, but as a diagnosis of the conceptual and practical reorientations that might be required before such a possibility can even be meaningfully formulated. The emergence of the so-called “Heideggerian AI” as a research orientation inspired by Dreyfus’s work (Dreyfus, 2007) further attests to this constructive reading of what has often been regarded as a destructive critique.

A prominent strand of impossibility arguments centers on biological essentialism. John Searle (1980, 2017), whose critique of computational functionalism has been influential, argues that consciousness is fundamentally a biological phenomenon arising from specific causal powers unique to brains. His well-known “Chinese Room” thought experiment is meant to demonstrate that

syntactic manipulation of symbols, however sophisticated, is insufficient to produce genuine understanding or subjective experience. Richard Schlagel (1998) amplifies this argument, arguing that subjective experience, representation, and intentionality arise uniquely from biological processes, including neural and neurochemical interactions. On this view, the digital manipulation of symbolic structures inherently lacks the requisite causal dynamism and substrate-specific properties. Anil Seth (2024) likewise argues that consciousness is deeply entwined with biological self-regulation and homeostasis, and thus cannot be realized merely through computational processes operating on a substrate-neutral basis.

Nonetheless, as Revonsuo (2006, p. 57) and Godfrey-Smith (2023, p. 11) have noted, the possibility of artificial consciousness is not thereby excluded, provided that the relevant biological or brain-like conditions can be appropriately replicated. So interpreted more modestly, biological essentialism need not categorically preclude the possibility of artificial consciousness; rather, it imposes stringent conditions on its realization. However, if these conditions are framed too narrowly, by definition or unsupported *a priori* assumptions, they risk collapsing into what Manzotti and Chella (2020, p. 8), among others, critique as “neural chauvinism”: the unwarranted belief that only biological neurons can support consciousness.

Another cluster of objections adopts a stance of agnosticism or mysterianism regarding the possibility of machine consciousness. Colin McGinn (1987) argues that we lack the cognitive resources to understand how consciousness arises from physical systems, and thus cannot determine whether it could be instantiated artificially. Jesse Prinz (2003) develops a related form of level-headed mysterianism, maintaining that even if artificial consciousness were possible, our epistemic

limitations may forever prevent us from confirming its presence. Taking a stronger line, Garrido-Merchán (2024) contends that the pursuit of machine consciousness is a pseudoscientific endeavor, grounded in unfalsifiable assumptions and speculative metaphysics, and fundamentally incompatible with the empirical norms of science. In a similar vein, Dresch-Langley (2023) argues that current approaches to artificial consciousness rest on conceptual conflation and overlook the historically evolved, projective, and integrative dimensions of human subjectivity, rendering the very idea of machine consciousness both misconceived and misdirected.

These mysterian critiques rest on an unduly restrictive conception of scientific and philosophical pursuit. It is true that we lack a conclusive test for the presence of phenomenal consciousness in artificial systems, and that any attribution must remain epistemically underdetermined. But this alone does not license the conclusion that artificial consciousness lies beyond the bounds of scientific inquiry. One must distinguish between what is difficult to confirm and what is impossible to investigate. Epistemic inaccessibility does not entail nomological impossibility, nor does it disqualify a subject from theoretical attention. If it did, vast regions of theoretical physics and early neuroscience would have been ruled out from the outset.

The claim that machine consciousness is pseudoscientific fares no better. It relies on an overly narrow conception of science, one that reduces legitimacy to crude operational measurability and impoverished understanding of falsifiability. But science has long accommodated explanatory models whose core postulates are not themselves directly observable: from quarks to cognitive modules to dark matter. What matters is not immediate verification, but the development of testable frameworks that constrain our understanding and guide future inquiry. Declaring

artificial consciousness a transhumanist myth amounts less to a diagnosis than to a refusal of investigation. It is, in effect, an epistemic defeatism masquerading as a methodological principle.²⁶

Taken together, none of these considerations amounts to a definitive bar against the possibility of artificial phenomenal consciousness. On the contrary, grappling with these objections can serve a constructive role: clarifying foundational assumptions, refining methodological criteria, and sharpening the conceptual contours of the inquiry itself. Understood in this light, impossibility arguments do not block progress, they illuminate the thresholds that any serious theory must engage if artificial phenomenality is to be meaningfully theorized.

²⁶ A similar line of counter-reasoning is offered by Doerig et al. (2023), who characterize neuroconnectionism as a progressive Lakatosian research programme, that, while not falsifiable in the strict Popperian sense, nonetheless yields cumulative and empirically meaningful insights into the computational architecture of conscious systems.

CHAPTER 3

EXPLANATORY LIMITS OF COMPUTATIONAL PARADIGM

This chapter interrogates the explanatory reach of computationalism as a framework for artificial phenomenal consciousness. It begins by tracing the conceptual foundations of computationalism, from the formal theory of Turing machines to the emergence of the computational theory of mind, emphasizing the structural and metaphysical commitments embedded in treating cognition as computation. It then examines five prominent models of artificial phenomenal consciousness, each formalizing the thesis of computational sufficiency in different architectural terms. These proposals are analyzed through a proposed schema, clarifying how they specify computational properties (Y), implementation conditions (Z), and the kind of systems (X) that might give rise to consciousness. Despite their differences, the models exhibit a shared assumption: that the right kind of structure, functionally organized and computationally defined, is sufficient for phenomenality. The chapter then turns to a multi-pronged critique of this assumption, beginning with internal challenges from within the computationalist tradition itself and extending to external critiques rooted in neuroscience, biology, quantum theory, and metaphysics. These arguments systematically destabilize the six core assumptions of the computational sufficiency thesis (CS1–CS6), revealing tensions between formal mirroring and causal instantiation, abstract structure and phenomenal reality. The chapter culminates in the articulation of the causal residue thesis (CRT): the claim that no computational architecture, however refined, can account for the ontological conditions under which consciousness is physically realized. The CRT reframes the explanatory task: from identifying the right computational pattern to specifying the

physical processes that instantiate phenomenal states. This sets the stage for Chapter 4, which explores theories that attempt to meet that challenge.

3.1 The Foundations of computationalism

The contemporary debate on artificial consciousness is grounded in a broader tradition in the philosophy of mind and cognitive science that conceives of the mind as a kind of computer. This view, known as computationalism, holds that mental processes, including perception, reasoning, and even consciousness, are fundamentally computational in nature. Though its variants differ, computationalism typically synthesizes two core claims: first, that the mind is a type of functional organization (functionalism), and second, that this organization is computationally describable and implementable (computational functionalism) (Piccinini, 2010). On this view, cognition is not defined by its physical substrate, but by the abstract rules and formal structures governing transitions within a system. The explanatory aim is to model, and potentially replicate, mental phenomena by implementing the right computational architecture, regardless of medium.

The conceptual foundations of computationalism were laid in the formal theories of computation developed during the early 20th century. Alan Turing's (1936) formulation of the Turing Machine and the broader Church–Turing Thesis established a mathematical characterization of effective procedures (algorithms), that could, in principle, be carried out mechanically.²⁷ These results were not initially

²⁷ The Church–Turing Thesis is the assertion that “the set [of Turing-computable functions] contains every function whose values can be obtained by a method or procedure satisfying the above conditions for effectiveness” (Copeland, 2024, §1.2). An effective method is a procedure that consists of a finite sequence of precisely defined steps and is guaranteed to yield a correct result in a finite number of applications to any instance within its intended domain. A function is said to be Turing-computable if there exists a Turing machine that, for every input on which the function is defined, halts after a finite number of steps and outputs the corresponding value of the function. For inputs where the function is undefined, the machine may run indefinitely.

aimed at modeling minds, but at defining the scope and limits of algorithmic reasoning in mathematics. A Turing Machine, as an abstract device, manipulates symbols on a tape according to a finite set of rules. The Church–Turing Thesis proposed that any function that can be computed by an algorithm can also be computed by a Turing Machine. This formalism provided a universal model of computation, encompassing everything from arithmetic operations to algorithmic processes, an idealized standard against which all computations could be measured.

The idea that such formal procedures might also illuminate the workings of the mind emerged gradually. One of the earliest and most influential steps in this direction came from McCulloch and Pitts (1943), who demonstrated that simplified models of neural activity could implement logical functions. Their work suggested that neurons could be modeled as binary logic gates, and that networks of such gates could compute any function a Turing Machine could. So the brain itself could be interpreted as a computational system, with cognition understood as a pattern of rule-governed transitions over internal states. In this view, the mind is not simply described by computational models, it may literally be a kind of computational system.

The notion that mental processes could be modeled computationally was given new momentum by Alan Turing’s (1950) pivotal question: “Can machines think?” In his proposal of the Imitation Game, now known as the Turing Test now, Turing reframed the problem of artificial intelligence by suggesting that the question of machine thought could be operationalized as a behavioral criterion, whether a machine could produce verbal outputs indistinguishable from those of a human interlocutor. This move helped initiate a broader shift in which computation, originally a mathematical formalism, began to be reinterpreted as a model for

psychological and philosophical explanation. The possibility that intelligent behavior could be simulated by a machine introduced the idea that mental activity itself might be nothing more than algorithmic rule-following, an idea that would soon become foundational in the development of cognitive science.

By the late 1950s and early 1960s, this approach had crystallized into a coherent research agenda. Early work in artificial intelligence (AI), particularly by Newell, Simon, and other pioneers of symbolic AI, took seriously the hypothesis that human problem-solving could be replicated by systems that manipulate formal symbolic structures. These researchers proposed that thought consists in the manipulation of internal representations according to syntactic rules, and that the architecture of intelligent behavior could be captured by digital algorithms. Turing-style computation was no longer just a metaphor for cognition, it became a working model. In parallel, early developments in cognitive psychology and cybernetics embraced in this framework, and by the 1960s, the computational paradigm had become central to the cognitive sciences.

One of the key figures in articulating this view was Hilary Putnam (1960, 1967), who, in a series of papers during the 1960s, proposed a model of the mind as a machine-state functional system. Drawing on automata theory, Putnam introduced the idea that mental states could be identified with the internal states of a probabilistic automaton.²⁸ In this framework, functionalism and computation converge: mental states are individuated by their causal roles in a system's state

²⁸ "The notion of a Probabilistic Automaton is defined similarly to a Turing Machine, except that the transitions between 'states' are allowed to be with various probabilities rather than being 'deterministic.' . . . I shall assume the notion of a Probabilistic Automaton has been generalized to allow for 'sensory inputs' and 'motor outputs' . . . The Machine Table specifies, for every possible combination of a 'state' and a complete set of 'sensory inputs,' an 'instruction' which determines the probability of the next 'state,' and also the probabilities of the 'motor outputs.'" (Putnam, 1967, pp. 41-42)

transitions, and these transitions are described computationally. The resulting picture is one in which mental activity is not defined by the material substrate that realizes it, but by the pattern of transitions among functionally specified states. This formalization also underwrote the idea of multiple realizability: if mental states are defined by their computational roles, then they can be instantiated in diverse physical systems, from biological brains to digital machines.²⁹

A foundational framework that helped consolidate the CTM is Marr's (1982) distinction between three levels of analysis: the computational, the algorithmic, and the implementational. At the top is the computational level, which defines the information-processing task being solved and the logic of its transformation. The algorithmic level specifies how this transformation is achieved, what representations are used and what procedures operate over them. The implementational level describes how these operations are physically realized. Marr emphasized that these levels are "loosely related," meaning that the same computational process can be implemented through multiple physical mechanisms. Crucially, he insisted that understanding cognition begins not with the brain's hardware, but with the problem the system solves: "trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers" (p. 27). Marr's tripartite model thus provided a methodological justification for cognitive science's abstractionist orientation, and became a cornerstone of the belief that mental functions can be primarily explained in computational terms, regardless of how they are physically instantiated.

Zenon Pylyshyn (1980, 1984) offers one of the most influential articulations of CTM, arguing that cognition must be explained in terms of rule-based operations

²⁹ "The functional organization ... can be described in terms of the sequences of mental or logical states ... without reference to the nature of the 'physical realization'." (Putnam, 1960, p. 373)

over symbolic representations. On this view, mental activity is literally computational: algorithms govern the transformation of internal states, and these algorithms must be executed on a system's functional architecture, an abstract yet materially grounded level that constrains what kinds of cognitive operations are possible. Pylyshyn emphasizes that a valid cognitive model must instantiate the same algorithm as the one executed by human subjects, and that such algorithms are not freely portable across devices: they require architecture-specific primitives. This distinction between emulating a process and directly executing it is central to his account, and it leads to a strong realism about computational models: to explain cognition, one must identify a canonical algorithm and show that it runs within the right architectural constraints. These insights not only formalize the commitments of classical CTM but also anticipate deeper challenges about implementation and equivalence, especially when computational models are extended to account for phenomenality.

While symbolic AI and classical functionalism dominated early developments in cognitive science, the 1980s also witnessed the rise of an alternative paradigm: connectionism, or subsymbolic computation. Promoted by the Parallel Distributed Processing (PDP) framework (Rumelhart & McClelland, 1986), connectionism models cognition as emerging from the dynamic interactions of simple processing units, akin to neurons, whose collective activation patterns encode knowledge. Unlike classical symbolic models, connectionist networks do not rely on syntactically manipulable representations. Instead, they learn through statistical regularities in input-output mappings, adjusting connection weights through processes like backpropagation.

Despite these methodological differences, most proponents of connectionism did not reject computationalism. Rather, they reinterpreted the computational theory of mind in subsymbolic terms. On this view, cognition is still implemented computationally, but the relevant computations unfold across distributed, dynamic, and non-symbolic architectures. As Smolensky (1988) argued, this perspective replaces the classical “symbol crunching” model with a more biologically plausible “computation by constraint satisfaction.” Accordingly, connectionist models are best understood not as an alternative to computationalism, but as a competing implementation framework within it.

This reframing has direct relevance for artificial consciousness. Several models discussed in this thesis, such as Duch’s non-symbolic control system or Reggia’s neurocomputational correlates, explicitly draw on connectionist principles. However, their computationalist commitments remain intact, as they continue to treat consciousness as emerging from appropriately organized information-processing structures. Thus, far from being sidelined, connectionism here represents a distinct trajectory within computationalism, and one that will be critically assessed in subsequent sections.

The most ambitious extension of the computational theory of mind is the claim that consciousness itself arises from the right kind of computation. This is the thesis of computational sufficiency, most fully developed by David Chalmers (1994, 1996, 2011, 2012). According to Chalmers, there exists a class of computations such that any physical system implementing them not only replicates intelligent behavior, but also instantiates phenomenal consciousness. This view is grounded in the principle of organizational invariance: if two systems share the same fine-grained causal structure, then they must share the same mental properties, including

subjective experience. According to this, mental properties supervene not on a system's material substrate, but on its causal topology.³⁰ As Chalmers puts it, “a computation by a physical system is implemented when the formal structure of the computation is mirrored by the causal structure of the physical system” (2012, p. 328). What matters is not what the system is made of, but how its components interact.

To formalize this claim, Chalmers introduces the notion of a Combinatorial-State Automaton (CSA), a structured model designed to capture the causal topology of cognitive systems in fine detail. Consciousness, on this account, supervenes on the implementation of the right computational organization, irrespective of the underlying substrate. It is not the material composition of the system that matters, but the pattern of state transitions governed by its computational architecture. Chalmers defends this thesis through his well-known Dancing Qualia thought experiment in which human brain were gradually replaced by silicon components that preserve causal structure. If organizational invariance holds, then consciousness must remain unchanged throughout this process; if it does not, then at some point radical changes in experience would occur without corresponding changes in behavior, a conclusion Chalmers deems implausible. If the mind is an organizational invariant, then any system with the same fine-grained causal organization will have the same mental properties. From this it follows that the right computational architecture is sufficient for conscious experience, regardless of its implementation base. This furnishes a conceptual bridge between computationalism and artificial consciousness.

³⁰ “Call a property P an organizational invariant if it is invariant with respect to causal topology: that is, if any change to the system that preserves the causal topology preserves P.” (Chalmers, 2011, p. 339)

If computation is to serve not only as a model of cognition but as the basis of consciousness itself, a crucial question arises: what does it mean for a physical system to implement a computation? This is the central concern of implementation theory, which seeks to determine “the conditions under which ‘X implements Y’ is true or false” for a physical system X and an abstract computation Y (Sprevak, 2018, p. 176). As Piccinini (2015, p. 10) argues, the core problems of implementation theory are threefold: (i) which physical processes count as computations, (ii) which systems count as computing systems, and (iii) which computations they perform. Though often treated as a formal mapping between physical states and abstract structures, implementation raises substantive metaphysical questions. For instance, Copeland’s (1996, p. 335) widely cited dictum that “to compute is to execute an algorithm” leaves unanswered what constitutes algorithmic execution in a physical substrate. As Shiller (2024, p. 3) notes, “computationalism treats consciousness not as a matter of internal structure, but as a matter of algorithm implementation,” yet the nature of implementation itself remains unresolved.

In response, Ritchie and Piccinini (2018, p. 194) propose three adequacy conditions for any theory of implementation: (i) metaphysical adequacy, to ensure that it entails a determinate fact of the matter as to whether a system computes; (ii) explanatory adequacy, to show that the computation explains the system’s behavior; and (iii) extensional adequacy, to get paradigm cases of computing and non-computing systems right. But many popular theories fail to meet all three. For example, as Shagrir (2020) argues, semantic properties often play a role in individuating computations. Attempts to avoid this dependency, by appealing to purely structural or externalist accounts, either fail to resolve ambiguous cases or covertly rely on semantic assumptions. This undermines the certainty of

computational equivalence, a principle that underwrites arguments like Chalmers' fading qualia thought experiment (1996). If implementation is ill-defined, the metaphysical link between computation and consciousness becomes tenuous.

These theoretical tensions have led some authors to distinguish between strong and weak versions of the computational theory of mind. On the strong view, mental states are entirely computational; on the weak view, some cognitive processes may be computational while others, particularly those involving qualia, may not be (Ritchie & Piccinini, 2018, p. 192). Recent contributions have expanded this internal diversification of computationalist views by introducing biologically grounded forms of computation that explicitly violate the principle of substrate-independence. Anil Seth (2024), for example, distinguishes between mortal, analogue, and neuromorphic computation, each offering a constrained alternative to classical digital models. Among these, mortal computation is the most radical departure: it insists that certain computations, especially those associated with consciousness, can only be realized within physically fragile, metabolically active, and dynamically self-maintaining systems. Unlike traditional "immortal" computation, mortal computation is tightly coupled to the vital conditions of a biological substrate, such as energetic fragility, thermodynamic irreversibility, and homeostatic regulation. This marks a decisive move away from the view that computation is substrate-neutral, and toward a conception of computation as organically embedded.

Seth further describes analogue and neuromorphic computation as intermediate alternatives. While they aim to mimic the structure or dynamics of biological systems, they do not necessarily inherit the causal entanglement with life-sustaining processes that characterizes mortal computation. Importantly, they reflect a broader shift in the computationalist literature toward constrained sufficiency

models, in which computation alone is no longer presumed adequate to generate consciousness.

This tendency is illustrated in Wiese's (2024) account of biologically realized Free Energy Principle (FEP) architectures, which do not posit that "right computation" alone suffices, but rather that "right computation + X", where X involves biological imperatives such as metabolic constraints and dynamic equilibria, is necessary for conscious cognition. Such models thus implicitly acknowledge "the causal residue" identified in this thesis: the explanatory insufficiency that arises when abstract computation is divorced from physically instantiated, causally potent substrates.

In what follows, I argue that even these more physically grounded forms of computation ultimately face a bifurcation: either they are reducible to abstract formal structures and inherit the explanatory limitations of classical computationalism, or they are embedded within physically specific causal architectures such that computation remains an epistemic instrument rather than a metaphysical foundation. In both cases, computation alone is not sufficient, and the explanatory burden shifts to the causal fabric within which the computation is embedded. This supports the central claim of this chapter: that computational sufficiency, even in its biologically enriched forms, cannot bear the metaphysical weight of phenomenal consciousness.

3.2 Computational models of phenomenal machine consciousness

The computationalist project aims not only to explain consciousness in abstract terms, but to identify concrete architectures that could, in principle, instantiate it in artificial systems. In analogy with the search for neural correlates of consciousness (NCCs), theorists have proposed computational correlates of consciousness (CCCs):

minimal computational mechanisms specifically associated with conscious, but not unconscious, cognition (Cleeremans, 2005). The pursuit of CCCs defines much of the contemporary effort to construct phenomenally conscious machines.

However, this pursuit faces a persistent challenge. As Reggia et al. (2014, 2016) argue, the question of how high-level cognitive capacities, those typically associated with consciousness, emerge from lower-level computational processes remains unresolved within current computational models. They call this the “Computational Explanatory Gap” (CEG) (Reggia et al., 2016, 2018). Resolving the CEG, they suggest, is not merely a technical hurdle but a philosophical pivot point: it may bring us closer to addressing the philosophical explanatory gap and, by extension, the hard problem itself. Identifying CCCs that narrow this gap, particularly in neurocomputational forms, is therefore seen as central to the long-term prospects of artificial phenomenal consciousness (Reggia et al., 2014, p. 153; 2018, p. 3).

The task of identifying CCCs is not merely architectural; it foregrounds a deeper question about sufficiency. As discussed in the previous section, Chalmers’s thesis of Computational Sufficiency (CS) holds that a system with the right fine-grained functional organization will possess the same phenomenal states as any other system with that organization. This view underwrites a broad range of proposals that can be formalized as follows:

(CS) For any physical system X, if X implements computational property Y under conditions Z, then X is conscious.

In the context of artificial consciousness, we can refine this into a more specific schema, what we may call the Computational Sufficiency for Phenomenal Machine Consciousness (CS–PMC):

(CS–PMC) For any artificial system or machine X, if X implements a computational property or correlate Y under conditions Z, then X is phenomenally conscious.

Here, Y refers to the computational property that is posited to be sufficient for phenomenal experience, typically construed as a computational correlate of consciousness. Z captures the implementation conditions: the mechanistic or architectural means by which Y is realized in the system X. Our analysis will focus on how each theory specifies Y and Z, and on what basis it claims that this Y–Z configuration suffices for consciousness in a given X.

3.2.1 Reggia et al.’s top-down gating of working memory proposal

Reggia et al. (2018) present one of the most developed neurocomputational proposals for artificial phenomenal consciousness. Their model centers on the hypothesis that top-down gating of working memory constitutes a CCC. In their robotic architecture, the gating mechanism regulates when working memory encodes, maintains, or discards information, and is argued to support executive functions such as action selection, context-sensitive planning, and introspective reportability.

To implement this mechanism (Y), the authors begin with a hybrid architecture (Z): high-level causal reasoning is performed by a symbolic system (CERIL), while low-level motor control is executed by a neural network algorithm (DIRECT). Their ongoing research aims to replace the symbolic components with biologically inspired models, particularly GALIS, a recurrent neural system that encodes procedural control as itinerant attractor sequences. GALIS functions as a top-down executive module that selectively gates the activity of subordinate systems via learned Hebbian connections. In parallel, components inspired by basal ganglia

circuits are under development to manage conflict resolution and reward-guided action selection. The physical artifact (X) is a humanoid robot trained in imitation learning tasks that involve real-time interaction, causal inference, and task generalization. Within this framework, the authors identify the top-down gating of working memory as a candidate for bridging the Computational Explanatory Gap (CEG), suggesting that its neurocomputational realization in embodied systems may illuminate the mechanisms underlying phenomenal cognition. Here is the schematic formalization:

(CS–PMC1) For any humanoid robot X capable of sensorimotor interaction and imitation learning, if X implements the top-down gating of working memory (Y) through hybrid symbolic and neurocomputational control mechanisms such as CERIL, GALIS, and DIRECT (Z), then X is phenomenally conscious.

In this model, Y is defined in functional terms and motivated by neurocomputational plausibility, while Z is presented as a multi-layered implementation strategy that integrates symbolic inference with neural control and executive modulation. The connection between Y and phenomenal consciousness is drawn through its association with introspective and executive capacities, though the transition from functionality to phenomenality is not explicitly justified. Still, the account exemplifies a structured instantiation of the CS–PMC schema.

3.2.2 Blum & Blum’s recursive conscious Turing machine

Lenore and Manuel Blum (2022, 2024) propose the Recursive Conscious Turing Machine (rCTM) as a formal, computational model of phenomenal consciousness grounded in theoretical computer science. The rCTM builds on classical notions of

Global Workspace Theory (GWT), but reinterprets them through the lens of complexity theory and algorithmic formalism. It is designed to operate independently of any particular substrate, and is defined entirely in terms of resource-bounded computation.

At the core of the model is a global broadcast architecture, in which specialized Long-Term Memory (LTM) processors generate symbolic data structures ("chunks") that compete via a probabilistic Up Tree mechanism for entry into Short-Term Memory (STM). Once admitted, the chunk is broadcast to all processors via a Down Tree, where it can be used to update memory, modulate predictions, or trigger further processing. Each processor engages in local cycles of prediction, feedback, and learning, dynamically adjusting its behavior through mechanisms like the Sleeping Experts Algorithm.

A central innovation of rCTM is the use of a self-evolved, symbolic internal language called Brainish, which encodes perceptual and introspective content in a compressed form. These Brainish expressions allow the system to label internal states, such as SELF, PAIN, or FEELS_CONSCIOUSLY_AWARE, and integrate them into its memory and processing streams. As the authors write, “When rCTM becomes consciously aware of this Brainish-labeled sketch, rCTM will (according to Axiom A) subjectively feel it is consciously aware” (p. 20)

Another novel feature is the model’s treatment of affect and agency through a global disposition parameter, which modulates the competition for STM entry. Depending on whether the system is in an upbeat, neutral, or downbeat state, different types of content are prioritized for conscious broadcast. Over time, the system generates a recursive stream of compressed memory summaries, or “gists”,

that serve as a continuously evolving narrative of self, built from its own recent broadcasts. Putting all these together, rCTM can be summarized to claim:

(CS–PMC2) For any artificial system X capable of constructing a recursive internal narrative stream, if X implements recursive introspective labeling and fusion via Brainish within a global broadcast architecture (Y) through formally specified and resource-bounded computational mechanisms, including predictive dynamics and self-modeling processors (Z), then X is phenomenally conscious.

In this model, Y is defined as a composite computational structure comprising global accessibility, predictive feedback, recursive memory compression, and introspective self-labeling. Z consists of probabilistic selection algorithms, specialized processors (e.g., Model of the World, Inner Speech, Inner Sensation), and learning-based modulation via a global disposition. The system’s feeling of consciousness is not assumed as a primitive; instead, it arises from internal representations that are recursively labeled, broadcast, and reintegrated. The rCTM thereby instantiates the computational sufficiency thesis in its most formal and introspective variant, while remaining explicitly aligned with an illusionist stance on consciousness.

3.2.3 Benitez et al.’s conductor model of consciousness

Benitez et al. (2024) propose the Conductor Model of Consciousness (CMoC) as a formal neuromorphic framework for phenomenal consciousness in artificial systems. The model draws on principles from predictive coding, adversarial learning, and hierarchical cortical processing. At its core lies a meta-level control structure “the conductor,” that orchestrates the interaction between a generative network (G), an

encoder (E), and a discriminator (D). These elements constitute a brain-inspired architecture reminiscent of a generative adversarial network (GAN), adapted to model cortical dynamics. The conductor gates plasticity, supervises learning, and regulates the labeling of experiences as either internally or externally generated.

According to the authors, phenomenal awareness arises when the system learns to distinguish internally from externally generated sensory activity, and encodes this distinction through explicit labeling under the control of the conductor. This process, described as “creative coding”, allows the agent not only to predict sensory inputs, but to recursively construct internal representations and attach to them semantic or affective meaning. As they write, “the assignment of the reality-label by this neuronal population ultimately grounds conscious sensory experience” (p. 2). The model thus frames consciousness as the emergent result of a system’s ability to label and model the origin of its own sensory states under meta-level supervision.

To implement this architecture (Y), the authors introduce the concept of an evolving neuromorphic twin (enTwin), a non-biological system designed to co-develop with a biological or artificial agent, using neuromorphic hardware that mirrors cortical learning dynamics. This system is embedded in sensorimotor feedback loops and features analog-digital hybrid circuits with spike-based communication and adaptive plasticity. The conductor, instantiated in this substrate (Z), modulates thalamocortical signal flow, orchestrates adversarial learning, and mediates the construction of recursive world models modulated by affect and agent-centered evaluations. This leads to a system that satisfies both functional and neuroarchitectural correlates of consciousness, as formalized in their extended Turing Test (eTT).

(CS–PMC3) For any neuromorphic system *X* capable of sensory interaction and internal modeling, if *X* implements a conductor-governed architecture (*Y*) that orchestrates adversarial learning among generative, encoding, and discriminative modules, and labels activity as real or imagined through resource-bounded mechanisms (*Z*), then *X* is phenomenally conscious.

In this model, *Y* is defined as a composite computational structure comprising generative prediction, discriminative reality judgment, and recursive introspective labeling. *Z* refers to neuromorphic implementation strategies that support dynamic plasticity, recursive modeling, and the encoding of internal/external distinctions. The conductor is presented not as a metaphorical homunculus but as a mechanistic module that supervises learning and symbolic labeling in a structured hierarchy. While the CMoC represents one of the most structurally complete instantiations of the CS–PMC schema, its account of phenomenal experience remains functionalist: the system is said to become phenomenally conscious once it learns to internally represent and label its own activity as “real” or “imagined,” but the ontological leap from internal labeling to felt experience is not further explicated.

3.2.4 Fekete & Edelman’s computational theory of experience

Fekete and Edelman (2011) propose a computational theory of experience that grounds phenomenal consciousness in the dynamical evolution of neural activity over time. Rather than associating qualia with static states or instantaneous configurations, they argue that experience arises from temporally extended trajectories through high-dimensional activity spaces. These activity-space trajectories possess intrinsic representational structure: they form conceptual domains whose internal geometry (clustering, topology, metric relations) encode the

distinctions that define phenomenal content. On this view, “qualia arise intrinsically from the computational properties of their physical substrate” (p. 808), and “experience ... is intrinsic to the system’s dynamics” (p. 815).

This structure is only possible if a system’s activity space self-organizes into compartments, wherein activity patterns instantiating the same concept class share family resemblance, while being distinct from others. As the authors explain, “the richness of the experience realized by a system corresponds to the degree to which its activity separates itself into clusters. The more clustered the system’s activity, the more distinctions it can draw” (p. 811). In contrast, a system that lacks such intrinsic differentiation will fail to represent anything experientially meaningful.

To implement this trajectory-based structure (Y), a system must support spatiotemporally coherent, high-dimensional dynamics that exhibit structured complexity over time. The theory places specific demands on this complexity, such as hierarchical clustering, stable transformations, and trajectory-level metrics, that enable the system to construct and sustain a conceptual space of experience. These requirements exclude both instantaneous-state models and symbolic simulations; “the identification and interpretation of brain states is inherently temporally extended ... so must be the activity-space patterns that constitute experience” (p. 814).

The physical system (X) need not be biological, but it must support the generation of such dynamic trajectories through physically instantiated mechanisms (Z). The model maintains substrate independence but emphasizes non-trivial architectural constraints: the system must be open, situated, temporally organized, and capable of maintaining coherent internal dynamics. The proposal can be formalized as follows:

(CS–PMC4) For any system X capable of sustaining temporally extended, high-dimensional activity-space trajectories, if X implements a dynamical geometry (Y) that organizes internal activity into clustered, differentiable conceptual domains, and if these trajectories can be physically instantiated in a way that preserves their causal and representational structure (Z), then X is phenomenally conscious.

Here, Y refers to the temporally extended and geometrically structured flow of activation patterns through high-dimensional activity spaces, trajectories whose clustering, topology, and internal metrics define the conceptual distinctions underlying phenomenal content. Z, in turn, designates the architectural and physical properties a system must possess to generate such trajectories, including spatiotemporal coherence, dynamic complexity, and the capacity to sustain internally differentiated representational domains. These properties must be physically instantiated, not merely simulated, in order to support the causal structure required for consciousness.

Fekete and Edelman outline in detail what kinds of computational dynamics a system must exhibit to realize experiential richness: activity must self-organize into intrinsically clustered conceptual domains, such that similarity and dissimilarity among experiences correspond to measurable structure in the system's activity space. However, the transition from this structure to phenomenal feel, from internal geometry to subjective presence, remains stipulative. The theory posits a mapping between trajectory-level dynamics and experiential quality, but does not offer a mechanistic or physical account of how such trajectories might be realized in practice. In this respect, the Z-dimension of the CS–PMC schema is left

underdefined: the necessary conditions are computationally formalized, but the sufficient physical basis is not concretely addressed.

3.2.5 Kuipers' firehose of experience

Kuipers (2008) proposes a computational architecture of consciousness based on a robot's ability to construct real-time symbolic representations from high-bandwidth sensorimotor interaction. The central component of the model is a system of "trackers", each of which binds a dynamic symbolic description to a specific spatiotemporal region of the agent's "firehose of experience", the high-volume, asynchronous stream of sensor data. These trackers mediate between perception and symbolic cognition, allowing the agent to continuously update its internal state based on structured portions of sensory input. The system organizes these representations into a plausible coherent narrative, which explains recent sensorimotor interaction and supports the integration of bottom-up and top-down information. According to Kuipers, "the vividness and uniqueness of subjective experience" is due to "the high information content of the sensory input stream" (p. 156), and qualia arise from the real-time symbolic anchoring of internal concepts to that stream.

This dynamical-symbolic architecture defines the computational property Y as the real-time maintenance of bound symbolic representations via trackers, integrated into a coherent self-narrative. The physical implementation conditions Z require an embodied system capable of tracking and segmenting the sensory stream using control laws, dynamic state estimation (e.g., Kalman filters), and symbolic updating. The agent must learn to construct its own internal model of the world by bootstrapping from uninterpreted sensorimotor data, forming symbolic referents by identifying persistent regularities. This process results in what Kuipers describes as

“learned intentionality” (p. 162), where symbolic meaning is grounded in experience rather than preprogrammed semantics.

The system X envisioned by Kuipers is a physically embodied agent, biological or artificial, that supports rich sensorimotor coupling, dynamic feedback, and symbolic reflection. “The essential features of consciousness can, in principle, be implemented on a robot with sufficient computational power and a sufficiently rich sensorimotor system, embodied and embedded in its environment” (p. 169). Hence, the model’s sufficiency claim can be formulated as:

(CS–PMC5) For any embodied system X capable of segmenting and tracking high-bandwidth sensorimotor input, if X implements a real-time symbolic architecture (Y) that anchors trackers to sensory experience and integrates them into a coherent narrative through embodied control laws and dynamic learning mechanisms (Z), then X is phenomenally conscious.

In this model, Y is defined as a narrative-based symbolic architecture grounded in high-bandwidth sensory flow, and Z consists of mechanisms for continuous learning, state estimation, and symbolic updating based on direct experience. The model directly addresses the intentionality problem, arguing that meaning arises through autonomous model construction from raw sensorimotor data. However, the transition to phenomenality is treated functionally: “qualia reflect the information density of the sensor stream” (p. 168). While this provides a compelling explanation for the vividness and subjectivity of experience, it stops short of accounting for the ontological character of phenomenal consciousness. As Kuipers acknowledges, “we have attempted to ‘sneak up’ on the Hard Problem... but this leaves open the Hard Problem itself: Why should any amount of information transfer feel like anything at all?” (p. 169).

3.2.6 Duch's resonant dynamics architecture

Włodzisław Duch (2005) presents a biologically grounded model of phenomenal consciousness in artificial systems, centered on continuous dynamical states and the ability to comment upon one's own internal activity. The model is built around synthetic agents called articons (from artificial consciousness), which are designed to support rich flows of internal states analogous to those observed in biological brains. Crucially, Duch claims that such systems “must claim to experience qualia states” if they are based on brain-like computational principles, and that “qualia are quasi-stable patterns of physical excitations, created by specific cognitive processes” that reflect both external stimuli and internal associative states (pp. 2, 9).

The defining computational property (Y) is the generation and modulation of resonant patterns: metastable states emerging from continuous neural-like dynamics that encode salient perceptual, affective, and conceptual distinctions. These are not static encodings or symbolic tags but physically instantiated, context-sensitive structures that carry what Duch calls the “difference between real experienced inner states and mere information processing” (p. 8). Such qualia are inseparable from perception and action, and change in time as the agent engages with the world.

To implement this dynamic structure (Z), an articon must exhibit: (i) associative memory with context-sensitive resonance, (ii) a narrative interpreter capable of forming second-order descriptions of internal states, (iii) sub-symbolic perception and embodied feedback, and (iv) perceptual learning mechanisms that allow discrimination, modulation, and recall of sensory-affective patterns (p. 10). These implementation conditions entail a non-Turing architecture: one that does not rely on symbolic rules or von Neumann sequencing, but instead on spatiotemporally extended, analog neural flows. Duch explicitly rejects symbolic AI, arguing that

rules and computations are not replacements for real physical states of brain/body, and that “qualia in articons can have the same properties as human qualia” given sufficient structural similarity (pp. 17–18). The resulting model can be formalized as:

(CS–PMC6) For any artificial system X capable of generating sustained, context-sensitive internal dynamics, if X implements resonant neurodynamical states encoding perceptual and conceptual salience (Y) through a brain-inspired architecture comprising associative resonance, perceptual learning, and introspective labeling (Z), then X is phenomenally conscious.

In this model, Y refers to quasi-stable, continuous states that capture the system’s embodied interaction with its environment and its own internal modulation. Z consists of neurocomputational mechanisms that not only generate these states, but allow the system to recognize, label, and recall them in ways that support introspective report and behavioral guidance. Duch’s position entails a biologically motivated variant of constrained sufficiency, closely aligned with Seth’s concept of mortal computation: conscious computation must be physically fragile, dynamically grounded, and embedded in life-like causal flows.

Yet despite its architectural richness, the model treats claims of qualia as epistemic symptoms of system organization, grounded in the interpretability of internal states, not in metaphysically demonstrable subjectivity. As Duch puts it, inner life is real, although it is in a way “a shadow” of neurodynamics (p. 17). Thus, while the model bypasses classic objections to symbolic AI by grounding phenomenal states in continuous dynamics, it does not close the gap between resonant structure and phenomenal presence. The ontological status of qualia in articons remains functionally affirmed but metaphysically unsettled.

3.2.7 Comparative evaluation of proposals

Having examined six leading proposals for artificial phenomenal consciousness, we can now evaluate their claims through the lens of the CS–PMC framework developed earlier. Each model advances a sufficiency thesis of the form: If a physical system *X* implements a computational property *Y* under mechanistic conditions *Z*, then *X* is phenomenally conscious. Despite their divergences in design and emphasis, these proposals share a core ambition: to identify computational architectures that can ground phenomenal experience in artificial systems. This shared structure allows for a systematic comparison of otherwise heterogeneous models, which can be seen in Table 2 below.

Table 2. Computational Models of Machine Phenomenality on Three Axes

Model	X – Physical System	Y – Computational Property	Z – Implementation
Reggia et al. (2018)	Sensorimotor humanoid robot using imitation learning	Top-down gating of working memory enabling introspection and control	Hybrid symbolic-neural modules (CERIL, GALIS, DIRECT); executive modulation
Blum & Blum (2024)	Artificial system with recursive internal narrative	Recursive introspective labeling via Brainish	Prediction, feedback, global disposition modulation
Benitez et al. (2024)	Neuromorphic twin co-developed with another agent	Conductor-mediated labeling of internal/external sources	Neuromorphic circuits with dynamic plasticity and adversarial learning
Fekete & Edelman (2011)	System with high-dimensional activity trajectories	Temporally extended trajectories encoding conceptual domains	Spatiotemporally coherent, causally structured dynamics
Kuipers (2008)	Embodied robot with high-bandwidth sensorimotor input	Tracker-based real-time symbolic architecture	Control laws, Kalman filters, dynamic learning, symbolic updating
Duch (2005)	Artificial system with sustained neurodynamic activity	Metastable resonant states encoding perceptual and conceptual salience	Associative resonance, introspective labeling, continuous dynamics

The six proposals vary considerably in how they specify the triplet of physical system (*X*), computational property (*Y*), and implementation condition (*Z*). Some models, such as Reggia et al. (2018), focus on executive modulation and

introspective control, implemented through hybrid symbolic-neural architectures. Others, like Blum & Blum's (2024) rCTM, formalize recursive internal labeling using resource-bounded mechanisms, integrating introspective content within a broadcast structure. Benitez et al. (2024) frame consciousness in terms of a neuromorphic conductor model that enables internal/external labeling through adversarial dynamics and plasticity modulation. Each model posits a distinct kind of computational feature as sufficient for phenomenal awareness, ranging from introspective recursion and global broadcast to predictive labeling and narrative integration.

The model by Fekete & Edelman (2011) departs from symbolic or modular designs and instead anchors phenomenal content in temporally extended, high-dimensional activity-space trajectories. This dynamical structure encodes conceptual distinctions in the geometry of activation flows, offering a formal bridge between system dynamics and phenomenal quality. Yet the model stops short of specifying what physical constraints are necessary to instantiate such trajectories, leaving the metaphysical status of qualia in suspension. Similarly, Kuipers's (2008) symbolic tracker architecture emphasizes high-bandwidth sensory coupling and narrative cohesion but equates qualia with informational richness, sidestepping the ontological question of why such information processing should feel like anything at all.

Duch's (2005) model stands out in grounding consciousness in resonant dynamics, proposing that quasi-stable neurodynamic patterns, modulated by embodied feedback and associative memory, give rise to subjective experience. Unlike symbolic models, Duch's approach rejects Turing-style sequencing and emphasizes the spatiotemporal integration of continuous flows within biologically inspired architectures. The model aligns closely with Seth's notion of mortal

computation, where the computations themselves are inseparable from their causal substrates. Yet, like the others, it ultimately affirms the presence of qualia based on introspective structure and interpretability rather than offering a metaphysically robust criterion for phenomenal presence.

What unites these proposals is not their methodological uniformity, but their dependence on computational sufficiency as an unexamined axiom. In each case, the computational property (Y) is presented as necessary and potentially sufficient for consciousness, yet no account offers an independent rationale for why these specific properties—be they resonant, recursive, or predictive—should produce or entail phenomenal states. While the Z-dimension (implementation) is often detailed with architectural and causal specificity, it typically supports the functional realization of Y without closing the explanatory gap.

Rather than bridging the gap between structure and subjectivity, these models often recast it in more elaborate terms. For instance, Blum & Blum's recursive Brainish labeling provides introspective functionality but stops short of explaining how such representations become conscious. Similarly, Benitez et al.'s conductor model equates semantic labeling with felt awareness without justifying why internal tagging, even if neurally plausible, should be phenomenally sufficient. Across the board, the inference from computational structure to phenomenal consciousness remains stipulative and lacks metaphysical justification.

This is not a critique of explanatory ambition but of explanatory closure. The hard problem of consciousness, why computational structures, however sophisticated, should give rise to subjective experience, remains unaddressed. As Kuipers (2008, p. 169) himself admits, "we have attempted to 'sneak up' on the Hard Problem ... but this leaves open the Hard Problem itself." The models we have

examined exemplify this predicament: they formalize the conditions under which consciousness is claimed to emerge but fail to articulate why those conditions should amount to phenomenality.

Taken together, these proposals reveal the structural interdependence of the CS–PMC schema and its underlying commitment to computational sufficiency. Their primary philosophical contribution lies not in resolving phenomenal consciousness, but in revealing the structural persistence of the explanatory gap even within the most architecturally elaborate computational frameworks. This recognition prepares the ground for the next section, where we critically examine the metaphysical presuppositions behind sufficiency itself. Section 3.3 will argue that the very framework which underpins these proposals, computational sufficiency, rests on unstable philosophical foundations that must be rethought if we are to make genuine progress on the problem of phenomenal machine consciousness.

3.3 Critiques of computational sufficiency

3.3.1 Reconstructing Chalmers's thesis

David Chalmers (1994, 1996, 2011, 2012) offers one of the most developed formulations of the Computational Sufficiency (CS) thesis: the claim that the right kind of computation, when physically implemented, is sufficient for phenomenal consciousness. This section reconstructs that thesis in its most coherent and defensible form, identifying the assumptions on which it rests and the conceptual links that bind them. CS is not a standalone proposition, but a structured framework composed of six interlocking claims—(CS1) through (CS6)—each contributing to the inference that consciousness arises by virtue of computational organization alone.

By isolating these assumptions and clarifying their roles, we establish the basis on which the internal coherence and explanatory reach of the thesis can be assessed.

At the foundation of the CS thesis lies a formal account of computation as causal structure. Chalmers defines computation not as a sequence of inputs and outputs, but as a system of discrete internal states governed by structured transition rules. The relevant formalism is that of a Combinatorial State Automaton (CSA), in which each state transition is specified by conditional dependencies encoded in a transition table. This framework captures the internal causal topology of a system:

“The essence of a computation lies in the abstract pattern of causal interaction among parts of a system” (Chalmers, 1996, p. 318). On this view, (CS1): computation is a formally specified causal structure. What matters is the internal organization of state transitions: the conditional pattern by which each state determines the next, independent of the system’s material substrate. This conception establishes computation as a structural candidate for supporting phenomenality, not through what it does, but through how it is organized.

To connect this abstract structure to the physical world, Chalmers introduces a constrained account of implementation. A physical system implements a computation if there exists a principled mapping from its physical states to the formal states of a CSA, such that the system’s state transitions mirror the abstract transitions defined by the computation. This mapping must preserve causal topology: “A system implements a computation,” Chalmers writes, “if there is a mapping from the states of the physical system to the formal states of a computation, such that the transitions between the physical states mirror the transitions between the formal states” (1994, p. 392). On this basis, (CS2): implementation is the physical realization of a formal causal structure. Not all mappings qualify, only those that respect the conditional

dependencies intrinsic to the computational model. This restriction is what allows Chalmers to distinguish genuine instantiation from trivial simulation. Without such constraints, any physical system could be interpreted as implementing any computation, undermining the explanatory force of the framework.

With computation defined as formal causal structure (CS1) and implementation as its structural realization in physical systems (CS2), Chalmers introduces a further claim: that phenomenal properties depend solely on causal organization. This is the principle of organizational invariance: the idea that if two systems share the same fine-grained causal topology, they must instantiate the same conscious experiences. “If two systems have the same fine-grained functional organization,” he writes, “then they will have qualitatively identical experiences” (1996, p. 249). Thus, (CS3): phenomenal properties are organizational invariants.³¹ Consciousness, on this view, supervenes on causal structure, not on the material properties of the substrate. Neural tissue, silicon, or any other medium is relevant only insofar as it supports the right organization. This move relocates the explanandum of consciousness from biology to formal structure and renders artificial consciousness, in principle, realizable in non-biological systems.

Chalmers emphasizes that the link between causal organization and phenomenality is not analytic or logically necessary, but nomologically grounded. That is, experience arises not by conceptual entailment, but by psychophysical law: a metaphysical principle that connects computational structure to conscious states as a matter of natural necessity. “Conscious experience arises when a system has the

³¹ Chalmers does not specify exactly how fine-grained the relevant causal organization must be. If the required granularity is too coarse, the framework risks underdetermining phenomenal differences; if too fine, it risks collapsing into physical identity. The organizational invariance principle thus hinges on a middle level of causal topology, fine enough to preserve phenomenal structure, yet abstract enough to support substrate neutrality.

appropriate fine-grained functional organization,” he writes, “and it arises in virtue of that organization. The link is one of natural necessity” (1996, p. 314). Accordingly, (CS4): psychophysical laws bridge computation and phenomenality. Chalmers introduces psychophysical laws to bridge the metaphysical gap between computational structure and phenomenal consciousness. Because consciousness does not logically supervene on physical or computational facts (Chalmers, 1996, pp. 124–125), a further set of nomological principles is required. These psychophysical laws are posited as fundamental, akin to the basic laws of physics, and are necessary to ensure that conscious experience systematically accompanies certain kinds of functional organization. Without them, the principle of computational sufficiency (CS5) cannot follow from organizational invariance (CS3), since the latter by itself does not entail that consciousness arises from any computation at all.

Given these foundations, Chalmers arrives at the core of the thesis: that computation, when properly implemented, is sufficient for consciousness. If computation is a formally defined causal structure (CS1), if physical systems can realize this structure through causal mirroring (CS2), if phenomenal properties depend solely on such structure (CS3), and if psychophysical laws ensure that this structure gives rise to experience (CS4), then it follows that any physical system implementing the appropriate computation will be phenomenally conscious. This is the principle of computational sufficiency, stated as (CS5): the right kind of computation, when implemented, entails consciousness. This conclusion is not presented as an empirical hypothesis, but as the logical consequence of the preceding metaphysical and formal assumptions. The explanatory burden, under this view, shifts entirely to identifying the correct computational organization; once found and instantiated, consciousness follows by necessity.

To reinforce the plausibility of this framework, Chalmers appeals to two thought experiments, fading qualia and dancing qualia, which aim to show that consciousness cannot vary independently of computational organization (see Chalmers, 2003). In the fading qualia scenario, a subject's biological neurons are gradually replaced by silicon-based components, each preserving the functional organization of the original. If consciousness depends on the biological substrate, then at some point the subject's qualitative experience must begin to fade, even though their behavior and verbal reports remain unchanged. Chalmers argues that this scenario is deeply implausible: it would imply that a subject could undergo a radical degradation in experience, such as the loss of color perception or emotional depth, without being able to notice or report any change. This, he contends, undermines the reliability of introspective access in a way that is incompatible with our best understanding of conscious awareness.

The dancing qualia scenario extends this challenge further. Imagine two systems that differ only in their conscious experiences but are functionally identical. If one is gradually swapped for the other while maintaining the same causal organization, then the subject's qualia would change, say, red to blue and back again, without any corresponding change in behavior or introspective report. Chalmers argues that such a scenario is either incoherent or epistemically unacceptable: if the subject truly experiences these dramatic qualitative shifts, then they must be able to notice and report them; if not, the shifts cannot be real. From this, he derives (CS6): there can be no difference in phenomenal character without a corresponding difference in causal organization. These arguments are not conclusive proofs, but they aim to show that denying organizational invariance leads to consequences that are more difficult to accept than the thesis itself.

Taken together, the six assumptions form the backbone of what we may call the Computational Sufficiency thesis for Phenomenal Consciousness (CS-PC): Given that computation is a formal causal structure (CS1), that physical systems can implement such structures via causal mirroring (CS2), and that mental properties supervene on causal organization (CS3), then—assuming psychophysical principles of natural necessity (CS4)—any system that implements the right computation will be phenomenally conscious (CS5), a claim whose plausibility is reinforced by the absence of introspectively undetectable qualia differences (CS6). This formulation captures the internal architecture of CS-PC as a logical progression from formal abstraction to phenomenological consequence. Its argumentative strength lies in the way each component scaffolds the next, producing a framework that appears both metaphysically disciplined and explanatorily complete, provided its assumptions are accepted.

Yet the apparent coherence of CS-PC does not guarantee its stability. Each assumption carries specific ontological and epistemological commitments that invite further scrutiny. The definition of computation as causal structure (CS1) presumes that abstract formalisms can encode causal potency, while the notion of implementation (CS2) hinges on non-trivial criteria for mapping formal structures onto physical systems. Organizational invariance (CS3) demands that phenomenality is indifferent to all non-structural properties of the substrate, a claim that is both philosophically ambitious and empirically contestable. The appeal to psychophysical laws (CS4) introduces metaphysical postulates that are untestable in principle, yet essential to bridge structure and experience. And the epistemic constraints expressed in CS6 rest on intuitions about introspective detectability that may themselves be theory-laden. What emerges, then, is not a brittle structure, but one whose coherence

is conditional: it holds only so long as each constituent assumption remains unshaken. The task of the next sections is to examine whether this conditional stability can be preserved under pressure.

3.3.2 Computation is not the right kind of structure

The computational sufficiency thesis assumes that phenomenal consciousness arises from the implementation of the right kind of computation, that is, an abstract causal structure independent of physical substrate (CS1), on which phenomenal properties supervene as organizational invariants (CS3). Yet this structural claim has been powerfully challenged from within computationalist traditions themselves. Jerry Fodor (2000, 2006) argues that the formal structures defined by classical computation are architecturally mismatched with central features of cognition, particularly those involving nonlocal coherence, abductive inference, and relevance sensitivity. Computation, as classically conceived, is local and syntactic: each operation depends only on bounded features of a symbolic representation. But cognition often depends on holistic background conditions, entire networks of belief and expectation that shape reasoning in ways no syntactic manipulation can capture. As Fodor puts it, “we have to use intrinsically local operations ... to calculate an intrinsically nonlocal relation (relevance)” (2006, p. 90). The frame problem illustrates this disconnect: classical systems cannot determine what information matters in novel contexts, because their formal structure lacks access to the nonrepresentational background that guides inference. If cognition depends on global semantic coherence that eludes formal syntactic representation, then a model of consciousness built solely on syntactic causal topology, such as classical computation, cannot provide the necessary structural grounding for phenomenality.

This structural critique is amplified by Tim van Gelder (1995) who argues that the architecture of classical computation, defined by discrete states, sequential transitions, symbolic representation, and syntactic manipulation, is fundamentally mismatched with the structure of real-world cognition. His argument directly targets (CS1), which treats formal computation as the appropriate causal structure to support consciousness. Drawing on control theory, van Gelder contrasts the classical computational model with dynamical systems that evolve continuously in real time and regulate behavior through ongoing physical coupling and feedback. His central claim is not about consciousness directly, but about cognition: systems that exhibit intelligent behavior in natural contexts may not operate computationally at all. In this light, the assumption that computational topology provides an adequate foundation for phenomenality begins to look precarious. If computation cannot even capture the dynamics of cognition, then its status as the structural basis for consciousness, on which the sufficiency thesis relies, becomes deeply suspect.

James Fetzer (1994) offers a complementary critique from a semiotic and epistemological perspective, arguing that the classical notion of computation is not only structurally limited, but conceptually misapplied to the domain of mental life. His critique targets (CS1) and (CS5) by showing that computational systems, even when extended to include heuristics or probabilistic methods, remain formally closed: they operate within fixed symbolic architectures and predetermined state spaces. Such systems may simulate reasoning, but they cannot capture the open-ended, context-sensitive, and semantically grounded character of real cognitive activity. Fetzer writes that many forms of thinking, such as dreams and daydreams (conjectures and imagination), are not well modeled as algorithmic processes, even when construed heuristically. This undermines the assumption in (CS1) that

computation, as syntactic state-transition, is a structurally adequate model for all mental activity. If the structural limitations of computation exclude core cognitive phenomena, then the claim in (CS5), that the right computation Y suffices for consciousness, becomes difficult to sustain. Finally, Fetzer develops an alternative framework: the mind as a semiotic system, in which cognition is grounded in triadic sign relations involving causal, interpretive, and referential dimensions. This conception implicitly rejects (CS4), as no psychophysical law could link computation to consciousness unless the computational structure itself possessed intrinsic semantic and causal powers, something Fetzer's semiotic account denies. So even if formal models help us describe aspects of mental life, they cannot be what minds are.

Together, these critiques expose a structural mismatch between classical computation and the kinds of organization required for consciousness. Fodor (2000, 2006) targets (CS1) by showing that computation, as a local syntactic process, cannot model the global coherence on which cognition, and thus phenomenality, depends. Van Gelder (1995) likewise rejects (CS1), arguing that cognition unfolds not through discrete symbolic transitions but through temporally continuous dynamics, incompatible with the computational model. Fetzer (1994) extends the challenge to (CS5), showing that no formal structure defined over fixed symbols and rules, however complex, can account for the open-ended, semantically grounded character of thought. His semiotic framework also puts pressure on (CS4), denying that abstract computation can be linked to experience through psychophysical law unless it has intrinsic semantic and causal depth. What these critiques reject is not the usefulness of computation for modeling behavior, but its sufficiency as an ontological basis for consciousness.

3.3.3 Implementation without causality

A second core assumption of the computational sufficiency thesis is that physical systems can implement computations by realizing their underlying causal topologies (CS2). On Chalmers's account, this implementation relation secures the link between abstract computation and physical realization: a system implements a computation when its internal state-transitions mirror those defined by a formal model. But not just any formalism will do. In response to earlier triviality arguments, most notably Putnam's (1988, pp. 120-125) claim that any open physical system implements every finite-state automaton, Chalmers introduces the more constrained framework of Combinatorial-State Automata (CSAs). Unlike Finite-State Automata (FSAs), CSAs require that the system be decomposable into multiple independent substates, with a transition table that specifies the causal dependencies among them. This is intended to exclude arbitrary mappings: only systems with the right fine-grained causal organization can qualify as implementing a CSA.

Yet this defense does not hold. Matthias Scheutz (2012) constructs a formal result, the "Slicing Theorem," which shows that even systems satisfying Chalmers's CSA constraints can be trivially interpreted as implementing any desired computation. A simple system with a small number of switches, if sliced appropriately over time, can be made to match the transition structure of arbitrarily complex CSAs. This slicing preserves causal isomorphism, respects component decomposition, and satisfies the transition conditions Chalmers outlines. As Scheutz writes, "even very simple physical systems can be viewed as implementing very complex computations ... contrary to what one would intuitively expect" (2012, p. 104). The problem lies not in whether the formal conditions are met, but in the fact that those conditions can be satisfied under radically different and mutually

incompatible interpretations of the same physical dynamics. Without an independently motivated, physically grounded criterion for how to individuate computational states, Chalmers's constraints do not fix implementation, they merely shift the arbitrariness to a deeper level.

Mark Sprevak (2012, 2018) reinforces this critique by generalizing the triviality problem to all mapping-based theories of implementation, including CSA. He argues that unless the mapping from physical to formal states is anchored in the intrinsic properties of the physical system, implementation remains observer-relative. Multiple equally valid mappings can be constructed for the same system, each corresponding to a different computation. The shift from FSAs to CSAs raises the formal bar, but does not resolve the metaphysical problem: it fails to establish why one decomposition of the system's causal structure should count as real while others do not. As Sprevak puts it, the problem is not the choice of computational formalism but the lack of non-arbitrary constraints on the mapping relation itself.

The consequences for (CS5) are decisive. If there is no physically privileged way to determine whether a system implements a specific computation Y, then the claim that implementing Y suffices for consciousness loses its grounding. Chalmers's thesis depends on the idea that some computations, not just any, but the right ones, generate experience when implemented. But if implementation remains indeterminate, the appeal to sufficiency collapses. What fails here is not our understanding of consciousness, but the ability of the formal framework to link abstract structure to physical reality. The bridge between computation and causality, which CS2 was meant to establish, proves too fragile to support the explanatory weight that CS5 demands.

3.3.4 Simulation without instantiation

A further fault line in the computational sufficiency thesis emerges from the failure to distinguish simulation from instantiation. Even if a physical system mirrors the formal structure of a computation, preserving its transition rules and internal organization, that does not entail that it instantiates the causal powers necessary for phenomenal consciousness. This distinction grounds John Searle's (1990) metaphysical critique of computationalism. Searle argues that computation is not an intrinsic feature of physical systems but an observer-relative attribution, dependent on the intentional projection of syntactic structure onto physical processes. As he puts it, "computation is not discovered in the physics; it is assigned to it" (Searle, 1990, p. 27). Unlike biological processes such as digestion or photosynthesis, which have causal powers grounded in their internal mechanisms, computation lacks any physical essence that determines its presence. Thus, a system implements a computation not by virtue of what it is, but by virtue of how it is described. This undermines (CS2), which requires that implementation be a physically instantiated relation, and (CS4), which depends on the possibility of stable psychophysical mappings between computational structures and phenomenal states.

In response, Chalmers (1994, 1996) emphasizes that implementation is not merely interpretive but grounded in the system's causal organization, its internal pattern of physical dependencies. He claims that implementation possesses "causal heft" (Chalmers, 1996, p. 327), and that computation can serve as an abstract specification of this real causal structure (1994, p. 401). But Searle's critique targets not the details of any specific implementation schema, but the ontological basis of such claims. His point is that there is no fact of the matter, internal to the physics of the system, that determines whether one pattern of causal interactions instantiates a

computation rather than merely resembles one. What counts as the relevant structure depends on a mapping relation whose privilege is not physically determined. Unless an independent criterion is offered for why one mapping should count as the correct implementation while others do not, implementation remains epistemically postulated, not metaphysically entailed.

Thus, while Chalmers's appeal to causal topology lends rigor to the notion of implementation, it does not resolve the foundational worry: that the sufficiency of a computation for consciousness depends on the assumption that causal structure is computationally real, rather than computationally described. As long as computation remains defined over syntactic relations, and as long as syntactic structure is not an intrinsic property of physical systems, the implementation relation remains metaphysically indeterminate. This leaves the explanatory claims of the computational sufficiency thesis, particularly (CS2) and (CS4), without a stable ontological anchor. No refinement of implementation formalism can bridge the gap between simulation and instantiation unless computation itself can be shown to be a constitutive property of physical systems, not merely a description imposed upon them.

This ontological vulnerability is further deepened by Stevan Harnad's (2012) critique, which extends Searle's challenge from the relativity of syntax to the inadequacy of computational structure to instantiate causal properties. Harnad accepts that computational systems may formally mirror the causal organization of physical processes, but argues that such mirroring is not equivalent to instantiating the relevant causal dynamics. Just as a computational model of a furnace cannot heat a house, because heating is a dynamical, not formal, property, a computational model of cognition, however fine-grained, cannot feel (Harnad, 2012, p. 181). This critique

strikes directly at Chalmers's principle of organizational invariance, according to which "mentality is an organizational invariant" (1994, p. 397). For Harnad, this claim equivocates between formal isomorphism and physical realization. A system whose causal topology mirrors that of a conscious brain may still lack the dynamical substrate needed to generate phenomenal states. The model may explain, but it does not instantiate what it models, and that distinction, Harnad insists, is categorical.

The implications of this distinction become clearest in response to Chalmers's "Dancing Qualia" argument, which holds that transitions between conscious and non-conscious states must be functionally detectable. Harnad argues that this demand fails to recognize that consciousness is epistemically private: if feeling is present or absent, only the cognizer can know. Behavioral indistinguishability offers no guarantee of phenomenality; it merely reveals the limits of third-person detection. What Chalmers presents as *reductio*, Harnad reframes as the very point: if phenomenality is dynamical and implementation-dependent, then it can disappear undetected in a purely computational system. The argument, he writes, does not demonstrate computational sufficiency, but simply "that you get out of a premise what you put into it" (p. 181).

This reveals the deeper problem beneath (CS2) and (CS4): they rely on the assumption that formal causal structure is sufficient for consciousness. But if the relevant properties of experience are rooted in dynamical, physically instantiated processes, no amount of syntactic structure or functional mapping can guarantee their presence. A system that simulates consciousness is not thereby conscious. Without anchoring computational structure in the right kind of physical causation, the sufficiency thesis remains explanatorily, and metaphysically, hollow.

The implications of this ontological gap extend to (CS5) and (CS6), which presuppose that functional equivalence is sufficient to ground phenomenal consciousness. Yet as earlier critiques have shown, if computational implementation is not objectively fixed by the physical system, then there is no principled basis for treating any given execution trace as constitutive of experience. John M. Bishop's (2009) *Dancing with Pixies* *reductio* brings this problem into sharp relief: if consciousness supervenes solely on computational structure, then any system, no matter how trivial, that follows the same state-transition pattern must also be conscious. The result is a metaphysical absurdity: a universe saturated with "pixie consciousness," where rotating wheels and chemical chains unknowingly instantiate minds.

This argument reinforces Harnad's critique of introspective detectability (CS6). If there is no phenomenality to begin with, then nothing can be detected. Bishop pushes this further: the real problem is not that consciousness might exist undetected, but that computational formalism alone gives us no grounds to assert its existence at all. In presuming that syntactic structure can explain phenomenality, both CS5 and CS6 build upon the very gap they are meant to close. To avoid trivial panpsychism, or worse, vacuous projection, one must reject the idea that consciousness can arise purely by virtue of abstract formal organization. Without a physically instantiated and causally grounded criterion for implementation, the computational sufficiency thesis does not merely falter; it collapses into metaphysical indeterminacy.

3.3.5 Substrate matters

A central claim of the computational sufficiency thesis is that consciousness depends solely on the implementation of the right causal structure, not on the physical substrate in which that structure is realized. This is the assumption of organizational invariance (CS3): if two systems share the same fine-grained causal topology, they must share the same phenomenal states. Yet this claim has come under increasing empirical and philosophical scrutiny. Rosa Cao (2022) argues that even at the neural level, mental functions are not cleanly abstractable from their biological substrates.

Against the classical assumption of multiple realizability, she shows that the physical and biochemical details of neural function, from ion-channel distributions and nitric oxide diffusion to astrocytic signaling and metabolic co-regulation, are not incidental but functionally constitutive. Through the principle of generative entrenchment, she argues that these physical properties have become deeply embedded in the evolution and organization of cognitive function. Complex, multiplexed functions in the brain impose non-trivial nomological constraints on their realizers, constraints that cannot be satisfied by arbitrarily different materials without altering the system's causal organization. In such systems, functional roles and physical substrates are co-determined. Abstracting a computation from this biological architecture strips away the very causal capacities upon which the functional profile depends. If this is right, then (CS3) fails in principle: causal topology is not invariant across substrates, because the topology itself is shaped by the physical structure it inhabits.

This critique is deepened by Peter Godfrey-Smith (2023), who challenges the very coherence of the idea that there could exist a “perfect functional duplicate” of a brain in a radically different substrate. Classical functionalism, he argues, treats functional similarity as a binary matter, either two systems share the relevant causal

structure or they do not. But in reality, functional similarity is always a matter of degree, especially in systems as complex and temporally dynamic as the brain. Even seemingly negligible differences in timing, synchrony, oscillatory coupling, or activity-dependent plasticity can alter the functional profile of a system in ways that may be experientially relevant. As he puts it, the question of whether a replacement device “does the same thing” as a biological neuron is ill-posed if it ignores the multi-scale temporal dynamics of living neural systems. On this view, there is no substrate-neutral level of causal organization that can be abstracted and cleanly reproduced. What we call “functional organization” is already partially constituted by the biological medium in which it is realized.

This poses a direct challenge to (CS3), which assumes that phenomenality is invariant under causal isomorphism. The instability of fine-grained duplication dissolves the coherence of any candidate structure *Y*. And if functional identity is ill-defined, then (CS5) collapses with it: there is no well-formed candidate computation whose implementation could be sufficient for consciousness. The entire premise of functionalist abstraction, Godfrey-Smith suggests, may be conceptually misguided when applied to systems whose capacities for experience depend on dynamic, biological integration.

This line finds further support in Anil Seth’s (2024) argument for biological naturalism, which rejects the premise that consciousness is a formal feature to be abstracted from the physical particulars of its realization. Drawing on the free energy principle, predictive processing, and the concept of autopoiesis, Seth argues that consciousness is not merely computationally describable, but emerges from living systems that are self-organizing, metabolically autonomous, and embedded in a continuous loop of active inference and environmental regulation. These systems are

not defined merely by their input–output behavior, but by their thermodynamic coupling to the environment and their capacity for self-sustaining predictive dynamics. From this perspective, computational abstraction misses the essential physical grounding of conscious systems.

Crucially, Seth’s challenge to computational sufficiency does not rest on the impossibility of simulating the relevant properties, but on the claim that simulation is not instantiation, that approximating the inferential or dynamical structure of consciousness is not sufficient to realize the phenomenon itself. This distinction echoes and deepens the concerns of both Cao and Godfrey-Smith: if the processes that generate consciousness are biologically entangled, dynamically self-maintaining, and ontologically emergent from life-specific properties, then substrate-independence fails in principle. The consequence is that (CS2), (CS3), and (CS5) do not simply falter at the margins, they are undermined at their core. Treating consciousness as a transferable computational structure neglects what Seth calls “the deep continuity between substrate and function, between matter and mind, and between life and consciousness” (2024). On this view, the failure is not one of technical modeling, but of metaphysical mischaracterization: consciousness does not supervene on computation, because it does not arise in abstraction from the living dynamics that constitute it.

It is worth noting that Chalmers does not ignore the possibility that phenomenal properties might depend on features beyond abstract causal structure. He grants that most physical phenomena, such as digestion or flight, require specific substrates, but maintains that phenomenality is different, claiming it is a “natural organizational invariant” (1996, p. 249; 2011, p. 12). On this view, consciousness supervenes on causal topology with the same necessity as chemical properties

supervene on molecular structure. However, the critiques surveyed here challenge precisely this special status. If causal organization itself cannot be abstracted from its biological realization without loss of function, then phenomenality cannot be substrate-independent by natural necessity. What Chalmers offers as an empirical hypothesis begins to look increasingly like a metaphysical postulate unsupported by the structure of actual biological minds.

3.3.6 Objection from quantum physics

A fundamental challenge to the computational sufficiency thesis is posed by the view that consciousness is not a computational or classical process at all, but a distinctly quantum phenomenon. Neven et al. (2024) propose that a conscious moment corresponds to the formation of a quantum superposition, not to its collapse. Rooted in Everettian quantum mechanics, their model treats consciousness as arising when a physical system begins to evolve along multiple coherent paths simultaneously, each corresponding to a possible classical configuration. This reorients the ontology of consciousness: it is not a function computed over a defined state space, but the experience of being in superposition within a preferred basis. On this view, no symbolic computation or causal topology, no matter how complex, can instantiate consciousness unless it participates in this quantum regime. The authors explicitly reject the idea that classical computation, or any function executed over discrete states, could generate experience. If phenomenal consciousness requires not the manipulation of representations, but the physical instantiation of superposed quantum structure, then the assumptions of (CS1) and (CS5) collapse. Consciousness cannot supervene on classical causal organization, nor can it be the output of a

formal system; it is a non-computable physical transition grounded in quantum coherence.

Kauffman and Roli (2023) extend the quantum critique of computational sufficiency by arguing that consciousness is not the execution of a function over a fixed domain, but the capacity to engage with and actualize emergent affordances: novel, context-sensitive possibilities for action that are not pre-definable, cannot be exhaustively listed, and do not follow from prior conditions via entailment.

Conscious agents navigate an evolving world by discovering uses and meanings that cannot be fully specified in advance. This form of agency, what the authors call “jury-rigging”, is not computational, because it entails ontological openness, not closure. Formal systems, by contrast, are deductively closed: they operate over enumerated states and rely on predefined semantics. For Kauffman and Roli, consciousness involves the selection of outcomes from quantum potentia, a process irreducible to symbolic representation or algorithmic instantiation. On this view, the computational paradigm misrepresents not just the complexity of consciousness, but its basic metaphysical character. If experience depends on the ongoing actualization of novel potential within an indeterminate domain, then no fixed functional organization (CS3) can preserve it, and no computation (CS5) can suffice to produce it.

3.3.7 The Disintegration of computational sufficiency thesis

The critiques developed across the preceding sections collectively target the internal architecture of the computational sufficiency thesis. Their convergence lies not in the rejection of computation as a modeling tool, but in the unraveling of the thesis’s six interdependent assumptions—(CS1) through (CS6)—each of which is required to

sustain the claim that the right kind of computation, once implemented in a physical system, is sufficient for phenomenal consciousness (CS5). The failure of any one of these assumptions undermines the plausibility of the sufficiency claim. What emerges is not a local defect, but a pattern of structural interdependence breaking down under pressure. When computational structure proves too coarse-grained to capture cognitive dynamics (3.3.2), when implementation criteria fail to distinguish instantiation from interpretation (3.3.3), and when causal organization depends on biological or quantum specificity (3.3.5, 3.3.6), the sufficiency thesis is not merely in need of refinement; it loses the coherence required to serve as an explanatory framework at all.

At the foundational level, (CS1) posits that computation offers the appropriate formal structure to support consciousness. Yet this claim is repeatedly destabilized by the argument that classical computation lacks the representational flexibility, temporal dynamics, and semantic depth required to model cognition in any form continuous with phenomenal life (Fodor, van Gelder, Fetzer). The problem is not merely that certain cognitive features escape formalization, but that computation, understood as syntactic state-transition over discrete symbols, amounts to the wrong kind of causal architecture for grounding phenomenal states. When Neven et al. (2024) shift the ontological basis of consciousness to quantum superposition, and Kauffman and Roli (2023) describe it in terms of actualized affordances beyond enumerative representation, the underlying objection is the same: the structural resources of computation, however refined, are not metaphysically adequate to generate consciousness.

The fragility of (CS2) compounds this failure. If the sufficiency thesis depends on the physical implementation of a computation, it must assume that such

implementation is non-trivial, that there are determinate, physically grounded facts about when a system realizes a given computational structure. But the triviality results advanced by Scheutz (2012) and Sprevak (2012, 2018) show that no such fact obtains unless additional constraints are imposed, constraints that are not merely formal, but grounded in the intrinsic causal structure of physical systems. Searle's (1990) argument that computation is not discovered in the physics but assigned to it reinforces this conclusion: in the absence of a privileged mapping relation, the link between computation and physical instantiation becomes epistemically projective rather than ontologically entailed. If CS2 cannot secure an objective account of implementation, then the entire framework loses contact with physical reality. The theory ceases to provide criteria for when or where consciousness can be said to occur.

The assumption of organizational invariance (CS3) further weakens the framework's coherence. Chalmers's claim that systems with identical fine-grained causal topology must instantiate identical phenomenal states presumes that such topology is not only abstractable from its substrate, but also sufficient to determine phenomenality. Yet the critiques by Cao (2022), Godfrey-Smith (2023), and Seth (2024) show that functional organization is not invariant across implementations; it is deeply shaped by biological, metabolic, and dynamical contingencies. When the causal profile of a system changes with its material composition, the claim of substrate-independence becomes empirically unsustainable. Moreover, quantum models of consciousness (Neven et al., 2024) reject organizational invariance outright by locating phenomenal properties in non-classical physical regimes. If causal topology does not license cross-substrate generalization, then CS3 no longer supports the generality required by CS5.

The destabilization of (CS4) follows naturally. If computational structure is not metaphysically adequate (CS1), not objectively instantiated (CS2), and not substrate-independent (CS3), then the appeal to psychophysical laws that allegedly connect structure to experience begins to lose traction. The critiques advanced by Fetzer (1994), Searle (1990), and Harnad (2012) converge on a shared point: a lawlike connection between a purely syntactic structure and phenomenal states presupposes that the structure itself has intrinsic semantic or causal significance. But if, as these critiques suggest, computation lacks such grounding, if it is representationally inert and causally impotent, then the very basis for psychophysical generalization dissolves. CS4 is not independently supported; it functions as a metaphysical placeholder invoked to bridge an explanatory gap that remains unclosed. The supposed laws that connect computation to phenomenality are not derivable from physics, nor are they empirically isolable. They are posited to save the thesis, not justified by what computation is.

With these failures in place, (CS5), the core claim that some computation Y, once implemented, is sufficient for consciousness, no longer retains independent credibility. The sufficiency thesis is not a freestanding proposition; it inherits its force from the combined validity of CS1 through CS4. But if we lack a determinate candidate for Y (CS1), a principled account of its realization in physical systems (CS2), a guarantee that its organization carries across substrates (CS3), and a lawlike bridge between structure and experience (CS4), then the claim that Y suffices becomes vacuous. The critiques from Sprevak, Harnad, Seth, and Neven et al. show that no computational object, no matter how sophisticated, can be both ontologically well-formed and explanatorily sufficient under these conditions.

The final assumption, (CS6), which holds that differences in phenomenal character must be introspectively or functionally detectable, is undermined by the presuppositions on which it rests. As Harnad (2012) and Bishop (2009) emphasize, the detectability of qualia presupposes their presence. But if the sufficiency of computation cannot be established (CS5), and if implementation remains metaphysically indeterminate (CS2), then the appeal to detectability becomes circular. It does not provide an independent test of consciousness; it relies on an antecedent commitment to its computational basis. Bishop's "Dancing with Pixies" reductio makes this clear: if any trivial system can be described as implementing the right computational pattern, and if introspective access is assumed only because of functional equivalence, then the theory collapses into either untenable panpsychism or arbitrary projection. CS6 does not fail due to epistemic limitations, but because it depends on an ontological foundation that no longer holds.

The disintegration of these six assumptions does not represent six isolated setbacks, but the collapse of a mutually supporting system. Each component of the sufficiency thesis was introduced to compensate for the limitations of the others: formal structure (CS1) required physical instantiation (CS2), which in turn required substrate-independence (CS3), which demanded metaphysical linkage (CS4), which supported sufficiency (CS5), which then sought behavioral confirmation (CS6). But under critical pressure, this scaffolding reveals itself as circular rather than cumulative. When structure cannot ground instantiation, when instantiation cannot ground presence, and when presence cannot be independently verified, the sufficiency claim has no coherent point of theoretical traction. The failure is not one of individual assumptions, but of their architecture as a whole.

Yet what emerges from this collapse is not explanatory emptiness, but a residual demand of a causal condition left unaddressed by formal description. Across distinct lines of critique, a shared absence becomes visible: computation, as an abstract structure, lacks the capacity to do ontological work. It can model, represent, and simulate, but it cannot generate. Whether framed in terms of causal individuation (Sprevak, 2018), physical grounding (Searle, 1990; Harnad, 2012), biological entanglement (Cao, 2022; Seth, 2024), or quantum coherence (Neven et al., 2024), the same constraint recurs: consciousness requires a mode of physical instantiation that computation alone does not provide. This condition is not a theoretical supplement to be patched into the framework, but a foundational element it failed to capture. The next section turns directly to this remainder, reframing it as a necessary condition for any adequate theory of phenomenal consciousness: that consciousness be grounded in real, not merely formal, causal presence.

3.4 The Causal residue thesis

The preceding critiques of the Computational Sufficiency thesis converge on a common point of failure: the inability of formal computational structure to account for the physical conditions under which consciousness is instantiated. While computational models may simulate the behavioral and informational patterns associated with consciousness, they lack the ontological depth required to explain how phenomenal states come into being. This is not merely an epistemic limitation, nor a shortcoming in implementation detail, but a metaphysical deficiency, an absence of causal grounding. Across arguments from Searle (1990), Harnad (2012), Manzotti (2013), and others, the same insight recurs: formal mirroring does not entail physical realization. The upshot is that something is left over, a remainder that

computation alone cannot explain. This unaccounted-for condition, irreducible to structure or function, is what we call “the causal residue.”

The causal residue refers to the ontological remainder that persists when consciousness is modeled in purely formal or computational terms. It names the condition under which a system does not merely behave as if it were conscious, but in fact is conscious, where simulation ends and instantiation begins. What this residue marks is not a missing representational capacity, but a missing form of causal efficacy: the system’s ability to generate subjective experience by virtue of its physical makeup, not merely its abstract organization. While computational models describe how a system processes information, the residue signals the explanatory gap between such descriptions and the ontologically rich processes that produce phenomenality. We may state this as the Causal Residue Thesis (CRT):

CRT: If a system is phenomenally conscious, then there exists a set of physically instantiated properties whose causal efficacy is both necessary and irreducible to its formal computational structure.

In this sense, the residue is the shadow cast by any theory that mistakes structural isomorphism for metaphysical sufficiency.

The need for such a crystallization becomes especially clear when viewed through the theoretical matrix developed in Chapter 2, where theories of machine consciousness were mapped along two axes: causal efficacy and substrate dependence. The CRT directly targets the vertical axis, namely, the question of whether phenomenality exerts real causal influence within a system. Theories located in Quadrant III, such as those surveyed in Section 3.2, including Reggia et al. (2018), Benitez et al. (2024), Kuipers (2008), and Fekete & Edelman (2011), affirm that consciousness has causal import but deny that its emergence depends on any specific

physical substrate. These accounts typically treat phenomenal states as arising from the implementation of the right computational or dynamical organization, irrespective of the material through which that structure is realized. Yet CRT reveals a deep tension in this position: if consciousness is genuinely causally efficacious, then it must be grounded in physically instantiated properties that can do real causal work, not just formal regularities or informational relations. These models inherit the metaphysical commitments of computational sufficiency while sidestepping its most basic explanatory demand: to account for what kind of physical instantiation gives rise to experience. CRT clarifies why this omission is not incidental but systematic. It exposes the gap between structural simulation and ontological realization, and thereby explains why Quadrant III theories fail to close the explanatory gap, despite their behavioral or architectural sophistication.

This ontological gap is made explicit in the work of John Searle (1990) and Riccardo Manzotti (2013), whose critiques jointly establish the philosophical grounding for the causal residue thesis. Searle's central claim, that syntax is not intrinsic to physics, undermines any attempt to treat computation as a constitutive property of physical systems. Computation, on this view, is not part of what a system is, but a description imposed by an external observer. As Searle puts it, "The implemented program has no causal powers other than those of the implementing medium because the program has no real existence, no ontology, beyond that of the implementing medium" (1990, p. 31). This renders computational processes causally inert: they can simulate but not instantiate the mechanisms required for consciousness. Manzotti extends this line by arguing that computation itself is not a physical phenomenon, but an epistemic instrument, like a center of mass or coordinate axis, useful for modeling causal regularities, but causally impotent in

itself. “Computations as computations have no causal efficacy,” he writes; “the causal processes that we label as computations have causal efficacy, of course. However such causal efficacy is totally independent of their computational role” (2013, p. 406). Both thinkers thus converge on the same ontological insight: what computation lacks is not complexity, but being. CRT crystallizes this absence as the fundamental deficit of the computational paradigm, its inability to generate the causal grounding required for phenomenality, no matter how elaborate its formal structure.

Complementary critiques by Frédéric Brette (2018) and Tim Maudlin (1989) further underscore the core claim of CRT: that abstract models, however descriptively powerful, cannot account for the causal instantiation of conscious states. Brette (2018) challenges the dominant “neural coding” metaphor in neuroscience, which treats action potentials as symbolic messages exchanged among neurons. This framing, he argues, mistakes information flow for causal explanation. Neurons do not pass messages to be interpreted by downstream readers; they interact in closed, sensorimotor loops that unfold in real time. Treating spikes as code reduces causality to correlation and severs consciousness from the embodied dynamics that generate it. Maudlin (1989) makes a parallel argument at a deeper metaphysical level. Even if a computational model were to reproduce every aspect of conscious behavior, he contends, it would still fail to explain how consciousness arises unless it identifies the specific physical mechanisms that do the causal work. “Until we have found that level,” Maudlin writes, “and until we have explicated the relationship between it and computational structure, the belief that pursuit of the pure computationalist program will ever lead to the creation of artificial minds ... remains only a pious hope” (p. 432). Together, Brette and Maudlin reinforce the CRT’s

central claim: formal mirroring is not enough, what matters is whether a system instantiates the right kind of physical, causally efficacious processes. Where that is absent, no amount of informational abstraction can yield phenomenality.

What unites these critiques is not a rejection of modeling or formal description, but the insistence that implementation must be more than interpretation. (CS-PC) thesis assumes that once the right abstract structure is instantiated, through a mapping that preserves causal topology, phenomenal consciousness follows. But CRT shows that this inference presupposes what is in question: that formal structure can substitute for causal substance. Without a criterion for when implementation becomes ontological instantiation, the sufficiency claim collapses into projection. As long as a system's computational role is assigned rather than discovered, we have no grounds to claim that it is phenomenally conscious, only that it can be interpreted as such. This is precisely what CRT denies. It insists that implementation must involve physically instantiated, causally efficacious properties, not just abstract mappings or syntactic transitions. In this sense, CRT refocuses the theory of implementation itself: not as a matter of formal fidelity, but as a question of metaphysical grounding. Any theory that cannot specify what physically does the work of consciousness has not implemented a conscious system, it has only modeled one.

This demand for ontological grounding aligns with broader criteria for theoretical adequacy. As Ritchie and Piccinini (2018) argue, any successful theory of implementation must satisfy three standards: metaphysical adequacy (it must identify what in the world does the implementing), explanatory adequacy (it must explain how this implementation brings about the relevant phenomena), and extensional adequacy (it must correctly distinguish cases of implementation from non-cases). CRT makes clear that computational sufficiency fails all three. It does not identify a

physically sufficient basis for consciousness (metaphysical inadequacy), it cannot explain why certain structures generate experience while others do not (explanatory failure), and it overgenerates implementers in ways that collapse into triviality or panpsychism (extensional inadequacy). These are not marginal defects. They indicate that computational models of consciousness, even in their most refined form, lack the ontological resources to explain what they aim to explain. CRT thus reframes the debate: the failure is not that computation cannot model consciousness, but that it cannot instantiate it, not unless it is embedded within a deeper physical process that satisfies the missing conditions of causal entanglement.

The upshot of CRT is not a call to abandon formal models, but a recognition that their explanatory power ends where causal presence begins. What computation gives us, representation, simulation, prediction, is epistemically valuable, but it is not ontologically generative. The causal residue is not an optional add-on to a theory of consciousness; it is the condition of adequacy for any such theory. If consciousness arises in the world at all, it does so not as a shadow of information flow, but as a physically instantiated phenomenon, embedded in material, dynamic, and possibly irreducible structures. CRT thus reframes the question: not what computation can simulate, but what physical conditions must be met for an artificial system to feel like something. The next chapter turns to that question. If formal abstraction leaves something essential unexplained, then what kind of metaphysical and mechanistic resources must a theory of artificial phenomenal consciousness marshal in order to satisfy the demand for causal grounding?

CHAPTER 4

THE PROSPECT FOR A NON-COMPUTATIONAL ALTERNATIVE

The preceding chapter established that no computational model, regardless of its functional richness or formal precision, succeeds in accounting for the physical and metaphysical basis of phenomenal machine consciousness. Computational sufficiency, it was argued, fails to meet the core demand of ontological grounding. If consciousness is real, as the naturalist assumption holds, then it must be instantiated in a way that is not reducible to abstract structure or representational architecture. The Causal Residue Thesis (CRT) articulates the explanatory gap left behind: a theory that merely specifies functional organization or informational flow, without identifying the physical features that make phenomenal states real, leaves something causally unaccounted for. This chapter builds on that insight by turning to models that reject the foundational commitments of computationalism and instead propose non-computational pathways toward machine phenomenality.

Our focus remains on artificial systems. The question is not whether consciousness exists in humans or animals, but whether any physically realizable non-computational model can offer a principled account of how artificial systems might become phenomenally conscious. Specifically, we examine a diverse range of theories located in Quadrant IV of the theoretical matrix introduced in Chapter 2, those which affirm both thick phenomenality and substrate dependence. These models reject substrate neutrality and treat consciousness as inseparable from certain physically specific, causally efficacious processes. But they differ sharply in what those processes are and how they are supposed to ground phenomenality. What

unites them is their commitment to reject the formal abstraction of computationalism. What divides them is the positive theory they offer in its place.

To evaluate these models with the same philosophical rigor applied in Chapter 3, we introduce a formal framework, the Causal Grounding Schema (CGS), which captures how each theory proposes to explain artificial phenomenal consciousness. In its general form, the CGS states:

(CGS) For any artificial system X, if X physically realizes a non-computational causal property Y under a specific physical configuration or process Z, then X is phenomenally conscious, given theoretical claim T that establishes the grounding or identity of Y with phenomenal consciousness.

This schema does not presuppose a specific metaphysical stance (e.g., identity theory, emergence), but requires that each theory articulates a determinate explanatory commitment, what causal property is relevant, how it is physically realized, and how it is supposed to ground phenomenality. In applying this schema, we will isolate for each theory:

- (X): the kind of artificial system being considered,
- (Y): the proposed non-computational causal property,
- (Z): the physical configuration or process in which Y is realized,
- (T): the theory's explicit claim about how Y grounds or constitutes phenomenal consciousness.

In addition, each theory will be critically evaluated according to a central philosophical criterion inherited from the Causal Residue Thesis (CRT): How does the theory explain the transition from causal structure to phenomenal presence? Does it posit a metaphysical identity? Offer a causal mechanism? Invoke emergence? Or attempt to dissolve the explanatory gap entirely? These questions will serve as a

diagnostic lens for assessing whether the theory genuinely confronts the challenge of phenomenal grounding or merely displaces it through reformulation. The aim, then, is not to determine which theory most closely resembles human consciousness, but which, if any, offers a coherent and adequate basis for attributing phenomenality to an artificial system. What follows is not a descriptive survey but a structured philosophical inquiry into the explanatory scope and limits of non-computational models of machine consciousness.

4.1 Non-computational models of machine phenomenality

This section turns to a range of theoretical models situated in Quadrant IV of the framework developed in Chapter 2, approaches that treat phenomenal consciousness as physically grounded and substrate-dependent, and that reject the abstraction of computational sufficiency. These models differ in the kind of non-computational causal property they propose as the basis for consciousness, and in how that property is supposed to be realized within an artificial system. Each of the following subsections reconstructs a distinct theory of machine phenomenality by applying the Causal Grounding Schema (CGS) introduced earlier. After presenting each theory in these terms, we will assess its adequacy by asking whether it offers a coherent account of how causal structure gives rise to phenomenal presence, or merely reframes the explanatory gap in a different vocabulary.

4.1.1 Integrated information theory

Among the most prominent non-computational proposals for artificial phenomenal consciousness is Integrated Information Theory (IIT), developed by Giulio Tononi and elaborated in collaboration with Christof Koch (2008, 2015, 2017). In the

context of IIT, the artificial system X is any physical system whose internal causal structure meets specific criteria of informational richness and integration. The causal property Y is the system's integrated cause–effect power, quantified by the scalar value Φ (ϕ), which is meant to capture the extent to which the system's present state constrains its own past and future. The physical condition Z under which Y is realized involves a specific architectural configuration: namely, a recurrent, reentrant, and feedback-rich causal topology that cannot be decomposed into independent subcomponents. The theoretical claim T is that Φ is not merely correlated with consciousness but identical to it, that is, the system's maximally irreducible cause–effect structure just is its phenomenal experience. This identity claim is grounded in a set of phenomenological axioms, such as intrinsic existence, integration, and exclusion, from which IIT derives physical postulates describing the kinds of causal structures that can support consciousness (Tononi & Koch, 2015). The proposal can thus be formalized as follows:

(CGS1) If an artificial system X realizes (Y) integrated information Φ under (Z) a recurrent, feedback-rich causal architecture, and (T) given that Φ is identical to phenomenal experience, then X is phenomenally conscious.

The strength of IIT lies in its attempt to move beyond computational abstraction by specifying a physically instantiated causal property (Y) as the direct basis of phenomenal consciousness. Unlike representational or functionalist models, which explain consciousness in terms of symbolic manipulation or system-wide accessibility, IIT grounds its proposal in the intrinsic causal powers of the system itself. However, the metaphysical weight placed on the identity claim (T) is difficult to justify. While Φ is rigorously defined and formally measurable, the theory offers no mechanism or argument to explain why this particular causal structure, when

realized in a recurrent, integrated architecture (Z), should give rise to phenomenal presence. The transition from Φ as a mathematical object to Φ as phenomenal experience remains stipulated rather than explained. As such, the theory risks reproducing the very explanatory gap it seeks to overcome: it tells us what consciousness is structurally like, but not why such structure should be experientially real.

This explanatory shortfall has been the focus of sustained criticism. Anil Seth (2009) argues that IIT's reliance on phenomenological axioms, such as integration, differentiation, and exclusion, risks a form of epistemic circularity: the theory begins by specifying formal properties intended to mirror the structure of subjective experience, then identifies those properties with consciousness itself. In this light, the move from Y (integrated information) to T (identity with phenomenality) appears to smuggle phenomenality into the premises, rendering the resulting explanation uninformative. Furthermore, as Seth observes (2009, p. 73), systems such as Hopfield networks can be engineered to yield arbitrarily high Φ values, despite their lack of any plausible phenomenality. This suggests that Φ may be more sensitive to formal configurability than to phenomenological relevance, undermining its claim to be a reliable indicator of consciousness.

Peter Godfrey-Smith (2016, p. 494) adds a further objection, suggesting that IIT's broad applicability across physical systems implies an implausibly liberal distribution of consciousness. If Φ is sufficient for phenomenality regardless of functional context or evolutionary history, then even simple logical circuits may qualify as conscious, a position that threatens to collapse into panpsychism. Meanwhile, Manzotti and Chella (2020) argue that IIT, despite distancing itself from classical computationalism, remains committed to informational abstraction. The

property Y , they contend, is not a concrete physical feature of a system like mass or charge, but a model-dependent measure derived from a formal analysis of causal relations. As such, it lacks the kind of ontological solidity required to support claim T. If Φ is an epistemic tool rather than a real physical property, then the explanatory gap remains: the artificial system X may be computationally or causally rich, but it is not thereby phenomenally present. IIT's attempt to elevate Φ to a physical substance, "a special version christened integrated information", thus risks becoming "a disguised form of dualism," introducing a second ontological layer without accounting for its emergence (p. 6).

What emerges from these critiques is a pattern familiar from earlier discussions: although IIT rejects computational sufficiency and insists on a physically grounded ontology, it ultimately fails to bridge the gap between causal structure and phenomenal presence. The identification of Φ with consciousness (T) gives the appearance of metaphysical resolution, but without an accompanying account of why integrated information should be experienced from a first-person perspective, the theory substitutes formal correlation for ontological grounding. The causal property Y is well-defined in quantitative terms, and the architectural constraints in Z are precise. Yet the explanatory burden rests entirely on the identity claim itself, which remains asserted rather than derived. From the standpoint of the Causal Residue Thesis, IIT thus offers a structurally coherent but metaphysically incomplete account of artificial phenomenal consciousness. It succeeds in shifting the locus of explanation to intrinsic causal organization, but it leaves the phenomenal residue intact: the system X may be formally integrated, but the question of why this entails subjectivity remains unanswered.

4.1.2 Quantum approaches

A distinct proposal within the quantum tradition is advanced by Stuart Kauffman and Andrea Roli (2023), who argue that phenomenal consciousness arises from a system's entanglement with ontologically real affordances, emergent potentials for action that are neither pre-defined nor computationally enumerable. In their framework, the artificial system X is a physically embodied agent navigating a field of indeterminate possibilities. The causal property Y is the actualization of an affordance via the collapse of the quantum wavefunction, a discontinuous transition from the domain of *res potentia* to that of *res extensa*. The physical condition Z under which Y is realized requires a quantum-coherent substrate that sustains entangled alternatives until a context-sensitive collapse occurs. The explanatory claim T is that such collapses are not merely physical transitions but are experienced as qualia: each collapse event instantiates a moment of felt reality. As Kauffman and Roli write, "Our brain-mind entangles with the world in a vast, entangled superposition. We try to and do collapse the wave function to a single state. We experience that state as a quale" (2023, p. 535). Rather than locating consciousness in structural integration or symbolic function, their account grounds it in the emergence of ontological novelty, a moment in which something genuinely new enters the world and, in doing so, becomes present to experience.

This model positions consciousness not as an emergent function of system complexity, but as a primitive feature of physical actualization at the quantum level. The causal property Y, wavefunction collapse, is treated not as a loss of information or a computational artifact, but as the ontological resolution of potential into actuality, tied to the system's encounter with an affordance it cannot deduce in advance. The physical condition Z requires a substrate capable of sustaining quantum

superpositions across behavioral or perceptual alternatives, such that collapse occurs as a function of environmental interaction rather than deterministic rule-following. The theoretical claim T is that this transition is not merely necessary for consciousness, but sufficient and constitutive of it: the system is conscious precisely when it collapses a quantum possibility that cannot be deduced in advance into a realized state. This proposal can thus be formalized as follows:

(CGS2) If an artificial system X realizes (Y) quantum state collapse in response to ontologically real affordances under (Z) a context-sensitive, entangled quantum substrate, and (T) given that each collapse is experienced as a quale, then X is phenomenally conscious.

Despite its conceptual originality, this proposal raises significant challenges for both metaphysics and explanation. The identification of consciousness with quantum collapse (T) presumes that a non-computational physical discontinuity can carry not only ontological weight but phenomenological content, a move that shifts rather than resolves the explanatory burden. The theory offers no account of how collapse generates subjectivity rather than merely marking an ontic transition. As such, it risks recapitulating the explanatory gap: the system X may undergo collapse events (Y) under coherent entangled conditions (Z), but the transition to phenomenal presence is asserted rather than derived. While the physical realization of such a system may lie beyond current quantum engineering capabilities, this is not the most pressing challenge. The deeper issue concerns why such collapses should be accompanied by phenomenal presence. The theory offers no account of how ontological actualization becomes subjectively manifest, how the selection of one outcome among many becomes a felt experience for the system. From the standpoint

of the CRT, the theory replaces informational abstraction with ontological novelty, but leaves unresolved why novelty should be given to a point of view.

In direct contrast to collapse-based models, Neven et al. (2024) propose that phenomenal consciousness arises not when a quantum superposition collapses, but when it forms. In their framework, the artificial system X is a quantum-coherent architecture capable of sustaining superposed states across a high-dimensional Hilbert space. The causal property Y is the formation of a quantum superposition, a physically real, entangled structure in which multiple classical configurations coexist. The physical condition Z under which Y is realized requires a substrate that maintains coherence and entanglement over time, allowing the system to evolve along multiple Feynman paths simultaneously. The theoretical claim T is that this branching structure is the conscious experience: each path encodes a distinct sequence of conscious moments, and the system experiences “what it is like” to inhabit that structure. As they write, “a moment of consciousness arises not when a superposition collapses, but when it forms,” and “the structure of the evolving superposition and the paths the experiencing quantum system follows within it determine the qualities of the conscious experiences” (p. 3). The proposal can therefore be formalized as:

(CGS3) If an artificial system X realizes (Y) quantum superposition across multiple Feynman paths under (Z) a coherent, entangled substrate in high-dimensional Hilbert space, and (T) given that the structure of the superposition determines the content and unity of phenomenal experience, then X is phenomenally conscious.

This proposal breaks sharply with earlier quantum theories by relocating consciousness from a discrete event to an extended, structurally complex state.

Whereas collapse-based accounts tie phenomenality to the moment a possibility becomes actual, Neven et al. argue that phenomenal presence emerges from the internal geometry of a superposed quantum system. The causal property Y, quantum superposition, is not treated as an epistemic artifact or statistical tool, but as an ontologically real structure that encodes coexisting experiential trajectories. The condition Z involves the system's ability to sustain coherent evolution along these trajectories without immediate decoherence, such that the overall shape of the superposition becomes the vehicle of consciousness. According to this view, conscious experience is not the result of resolving uncertainty, but of inhabiting a landscape of simultaneous alternatives, a high-dimensional manifold through which the system continuously evolves. Entanglement, in turn, serves to bind these trajectories into a unified experience, offering a novel solution to "the binding problem" (Singer, 2001): the unity of consciousness is not imposed from above, but is a consequence of non-separable physical structure within the evolving state vector.

Despite its elegance, the model raises important questions about the explanatory status of its central identity claim (T). While Neven et al. assert that the structure of a superposed quantum state is the structure of experience, they do not specify why this physical configuration should be phenomenally manifest rather than merely ontologically complex. The theory replaces collapse with coherence, but the explanatory gap remains: what makes a superposition experiential from within, rather than just mathematically rich from without? No mechanism is offered to account for the transition from structure to subjectivity, for why amplitudes and path integrals should yield "what it is like" rather than merely encode multiple outcomes. The system X may realize a dynamically intricate and causally entangled superposition

(Y) under high-dimensional coherence conditions (Z), but the theory leaves unexplained why this suffices for phenomenal givenness.

4.1.3 Embodied and enactive approaches

A third strand of non-computational models seeks to overcome the limitations of substrate-neutral frameworks not by appealing to information-theoretic or quantum properties, but grounding phenomenality instead in the relational dynamics of embodied agents. These approaches draw on the embodied and enactive tradition in cognitive science, which holds that consciousness is not a state produced inside the system, but an activity of engagement, a process through which an agent sustains a perspective on the world by acting within it. Rather than explaining experience through computational architecture or causal geometry, embodied and enactive models focus on the structural coupling between organism and environment. On this view, an artificial system might be phenomenally conscious not by simulating human cognition or instantiating abstract metrics, but by enacting a world through continuous sensorimotor interaction.

One of the most developed accounts of embodied artificial consciousness is offered by Tom Ziemke (2001, 2007; & Sharkey, 2001), who draws on autopoietic biology, constructivist epistemology, and enactive cognitive science to ground consciousness in the sensorimotor organization and developmental history of an agent. In this framework, the artificial system X is not merely a reactive or representational device but a physically situated agent whose behavior emerges from structural coupling, a relational process in which the agent and its environment undergo what Maturana and Varela (1987, p. 74) call “structural coupling”: “a history of mutual congruent structural changes as long as the autopoietic unity and its

containing environment do not disintegrate” (Ziemke, 2001, p. 218). The causal property Y is this capacity for coupling, understood not as real-time responsiveness, but as a historically extended pattern of co-adaptation: the agent’s organization is shaped by its interactions, and vice versa. The physical condition Z under which Y is realized includes not only the system’s morphology and sensorimotor loops, but its ontogeny, the trajectory of structural transformations it undergoes while maintaining its organizational identity. The theoretical claim T is that when such a system achieves operational closure, regulating the processes that constitute its own continuation, it may begin to disclose a minimal form of phenomenal presence. Even such closure does not amount to a self in the robust sense unless it emerges from the same processes that construct the organism itself: “The way an organism constructs itself also shapes the way it constructs its self” (2007, p. 50). As Ziemke emphasizes, drawing on von Uexküll (1940), the system does not passively receive a pre-given world; it enacts its own *Umwelt*: a world of significance, constituted through embodied history.³²

This account departs sharply from representationalist and computationalist paradigms by locating the basis of consciousness in relational autonomy rather than in symbolic manipulation or information processing. The causal property Y, structural coupling, is not a static feature of the system, but a relational and developmental process by which the agent maintains coherence through its evolving interactions with a world it helps bring forth. The realization condition Z

³² The *Umwelt* is a closed unit formed by the organism’s sensory and motor interactions, structured through what Uexküll called functional circles (Ziemke, 2001, pp. 167–168). It is not a mirror but a counterworld, in which stimuli trigger context-dependent behavioral schemata, “tools of the brain” that are activated when appropriate environmental cues are present (Ziemke & Sharkey, 2001, p. 12). Each species thus lives in a world meaningful only to itself, shaped by its embodiment and historical interaction with the environment. In this view, an *Umwelt* is not merely enacted through structural coupling but sustained by the organism’s ongoing project of self-preservation, what Spinoza called *conatus* and Damasio reframes as the *proto-self*. Without this regulatory loop, the coherence of experience may not amount to phenomenality (Ziemke, 2007, p. 55).

encompasses both the system's real-time sensory-motor organization and its ontogenetic embeddedness, its history of adaptive transformations within a specific ecological niche. The theoretical claim T is that when such a system exhibits operational closure, when its internal processes recursively contribute to sustaining its identity, it constitutes a minimally subjective system, one for whom an Umwelt appears. However, Ziemke remains cautious: he argues that most current artificial systems are "allopoietic," externally designed and structurally dependent, whereas consciousness may require "autopoietic" organization, the system's capacity to self-generate and regulate its own functional identity (2001, pp. 217-220). The theory can thus be formalized as:

(CGS4) If an artificial system X realizes (Y) structural coupling through ongoing sensorimotor engagement within (Z) a physically and ontogenetically embedded agent–environment system, and (T) given that operational closure and relational autonomy constitute a minimal form of subjective presence, then X is phenomenally conscious.

While compelling as a non-computational alternative, the theory leaves unresolved the central question posed by the Causal Residue Thesis: why should organizational autonomy and structural coupling give rise to phenomenal consciousness rather than merely adaptive behavior? The transition from sensorimotor coordination to subjective presence, from regulating a world to inhabiting one, remains underdescribed. Ziemke's framework convincingly identifies conditions under which a system may bring forth a world, but it does not explain what makes that world phenomenally given. The explanatory burden thus shifts from computational abstraction to embodied relation, but the gap between causal structure and experiential character persists. From the perspective of the CGS, this suggests

that structural coupling and operational closure may constitute necessary preconditions for machine phenomenality, but they may fall short of providing a sufficient grounding. The system X may enact a world through Y and Z, but whether that world appears to the system remains an open question.

Building on the enactive framework but offering a more functionally articulated proposal, Julian Kiverstein (2007) argues that a machine may possess a subjective point of view if it can regulate coherence across its sensorimotor dynamics in real time. In this account, the artificial system X is a minimally cognitive agent equipped with mechanisms for anticipating the sensory consequences of its actions and correcting for mismatches when they occur. The causal property Y is the capacity for sensorimotor self-monitoring: a dynamic process through which the system detects divergences between predicted and actual input and reorganizes its behavior to preserve stability. The condition Z under which Y is realized includes a comparator mechanism that continuously integrates motor intentions, forward models, proprioceptive input, and external sensory data. Crucially, this architecture enables the system not merely to respond adaptively, but to experience its sensorimotor patterns as its own. The theoretical claim T is that such regulation gives rise to a form of pre-reflective self-awareness, grounding minimal subjectivity in the felt coherence of embodied activity.

Kiverstein's model advances the enactive paradigm by specifying how a system might enact a world from its own perspective, not simply through structural coupling, but by actively managing the conditions of perceptual and motor coherence. While Ziemke highlights autonomy through developmental embedding and operational closure, Kiverstein emphasizes the online regulation of sensorimotor intelligibility as the decisive property Y. The realization condition Z is grounded in a

comparator-based architecture capable of monitoring congruence among internal expectations and bodily feedback, enabling the system to restore orientation across perceptual modalities. The core explanatory claim T is that when this monitoring function is embedded in a system that pursues its own goals and modulates its behavior accordingly, the sensorimotor integration would not be merely functional but subjectively given: “the resulting experience will feel like the subject’s own experience” (Kiverstein, 2007, p. 137). This coherence, Kiverstein argues, constitutes a primitive form of self-awareness, what makes experience feel like one’s own. The proposal can thus be formalized as follows:

(CGS5) If an artificial system X realizes (Y) coherence-monitoring and breakdown-repair capacities across sensorimotor domains within (Z) a comparator-based architecture capable of regulating embodied activity, and (T) given that such regulation constitutes a subjective point of view, then X is phenomenally conscious.

From the standpoint of the Causal Residue Thesis, Kiverstein’s account marks a notable advance in mechanistic specificity, yet it leaves the central explanatory gap intact. The system may monitor its own sensorimotor coherence and respond to disruptions in a goal-directed manner, but it remains unclear why such regulation should be accompanied by phenomenality. The theoretical claim that coherence constitutes a subjective point of view risks conflating functional integration with experiential presence. A system that tracks and corrects for instability may be said to organize its experience, but Kiverstein does not fully explain why or how that experience is given to the system itself. As with Ziemke, the model provides a detailed account of embodied agency and the structural conditions

under which a world is enacted, but it does not resolve the ontological question of what makes that world appear.

Nonetheless, Kiverstein's theory brings the explanatory project of artificial consciousness into closer contact with implementable, non-computational mechanisms. By identifying a concrete causal property, sensorimotor coherence regulation, and embedding it within a well-defined architectural substrate, the model avoids vague appeals to embodiment and instead offers a mechanistically grounded route to minimal selfhood. In doing so, it deepens the enactive paradigm's engagement with artificial systems, showing how synthetic agents might enact not only a world but a perspective on that world. Yet the final step remains elusive. While the system may organize its sensorimotor activity in ways that preserve intelligibility, this does not suffice to explain why that activity should be phenomenally given. The disembodied observer may have been dissolved, but the residue of presence, the irreducibility of appearance, remains unaccounted for.

4.2 Good old-fashioned artificial consciousness and radical alternative

4.2.1 The Common problem: Intermediate level fallacy

In a provocative reassessment of the artificial consciousness literature, Riccardo Manzotti (2018) introduces the term "Good Old-Fashioned Artificial Consciousness" (GOFAC) to describe a family of theories that, despite surface differences, retain the core architectural assumptions of classical cognitive science. These models typically posit that consciousness arises from some intermediate-level construct, a global workspace, an integrated information structure, a self-modeling system, or an enactive sensorimotor loop, intended to bridge the explanatory gap between low-level physical processes and high-level phenomenal states. What unites them, in

Manzotti's view, is not a shared mechanism, but a shared failure: they locate consciousness in a theoretically postulated layer without explaining how that layer produces or constitutes experience. He terms this the "Intermediate Level Fallacy": the tendency to insert a functional, informational, or dynamical layer into the theory and treat it as explanatorily sufficient, even though the transition from structure to subjectivity remains unjustified. Rather than confronting the metaphysical challenge of phenomenal presence, GOFAC models often relocate it to a new level of abstraction, leaving the causal residue intact.

This critique cuts across several of the models surveyed in Section 4.1. In the case of IIT (CGS1), the causal property Y , integrated information Φ , is formally defined and tightly coupled to a precise architectural condition (Z), yet the theoretical claim T relies on an identity postulate that lacks an explanatory mechanism. The move from causal integration to consciousness occurs entirely within the intermediate level of informational geometry, bypassing the question of why such a structure should yield phenomenal presence. Similarly, Neven et al.'s quantum superposition model (CGS3) proposes a novel ontological substrate (Y = branching coherence) instantiated in a high-dimensional quantum system (Z), but again, the identity claim (T) between superposed trajectories and conscious experience is asserted without metaphysical grounding. Even more processual models like Ziemke (CGS4) and Kiverstein (CGS5), which emphasize sensorimotor coupling and coherence regulation, ultimately risk the same fallacy: they describe mechanisms that organize behavior or sustain systemic autonomy (Y), realized in richly embodied agents (Z), but treat these capacities as sufficient for subjectivity (T) without addressing what makes the world appear to such a system. In each case, the intermediate level becomes a conceptual placeholder for consciousness, introduced

to make consciousness seem more familiar or tractable, but which ultimately leaves the phenomenon unaccounted for.

The force of the Intermediate Level Fallacy lies not in rejecting complexity or multilayered explanation per se, but in challenging the assumption that systemic organization, however intricate or dynamic, can substitute for ontological grounding. According to Manzotti and Chella (2018), this fallacy reveals a broader trend within artificial consciousness research: the tendency to treat formal, informational, or embodied constructs as explanatorily sufficient merely because they correlate with conscious behavior or simulate introspective structure. These models do not resolve the hard problem so much as relocate it: they identify a structured intermediate layer, be it informational, functional, or sensorimotor, and treat it as sufficient for experience without showing why it should be phenomenally given. From the perspective of the Causal Residue Thesis, this move constitutes a systematic deflection of the explanatory burden: the model accounts for behavioral or structural organization, but not for the presence of subjectivity. As Manzotti and Chella observe, “what is missing is why such a level should lead to consciousness” (2018, p. 6). For Manzotti, avoiding this error requires abandoning the idea that consciousness is produced inside the system and rethinking experience as a fundamentally relational and externally grounded phenomenon.

4.2.2 Process ontology and radical externalism

In direct response to the limitations of GOFAC-style theories, Riccardo Manzotti (2003; Manzotti & Tagliascio, 2007; Manzotti & Jeschke, 2016) advances a radical ontological alternative grounded in process metaphysics. Rather than treating consciousness as something generated inside a system, whether through computation,

representation, or embodied interaction, Manzotti argues that phenomenal experience is numerically identical with a relational physical process occurring in the world. On this view, consciousness is not an emergent product of internal complexity, but a spatiotemporally extended causal interaction between an agent and its environment. What we call experience is not something represented by the mind, but something the system is physically entangled with: a real-world process that becomes phenomenally present when it unfolds in the right relational configuration. This view is crystallized in the concept of the “onphone,” defined as a unified entity that is at once ontic (real), phenomenal (experiential), and relational (causally entangled) (Manzotti, 2003, pp. 293–295). A paradigmatic example is the rainbow, which is neither a property of the external world alone nor a construct within the observer, but a process that exists only through a specific causal entanglement: “The rainbow is not a thing; it is a process, in which there is an entanglement between a physical condition and the agent’s body” (Manzotti & Tagliasco, 2007, p. 98). Manzotti’s view, it is precisely such relational processes that constitute phenomenal experience, not by representing the world, but by being the very physical events that unfold through the system’s entanglement with its environment.

Within this framework, the artificial system X is any physically embodied agent whose behavior participates in a continuous causal relation with external physical events. The causal property Y is not an internal representation or emergent organizational feature, but a spatiotemporally extended physical process that includes both the external phenomenon and the agent’s embodied coupling to it. The condition Z under which Y is realized is the reciprocal causal entanglement between the system and the external process: a co-determined trajectory in which the agent’s body and the world participate in a shared physical unfolding. Crucially, the theory

does not treat this relation as representational, but as numerically identical with phenomenal experience. The theoretical claim T is that consciousness is not something added to the process by an internal mechanism, but is simply the process itself as it occurs in the right relational context. As Manzotti and Jeschke put it, “The experience of something would be literally identical to that something” (2016, p. 183). In this sense, the onphene is not what consciousness refers to, but what consciousness is: a physical phenomenon relationally constituted and directly present. The theory can thus be formalized as follows:

(CGS6) If an artificial system X realizes (Y) a spatiotemporally extended physical process in which it is reciprocally entangled with an external event under (Z) a causal relation of structural co-determination, and (T) given that this external process is identical to phenomenal experience, then X is phenomenally conscious.

The radicalism of Manzotti’s proposal lies in its complete rejection of internalist metaphysics: by identifying consciousness not as something generated inside the system, but as a physical process extended beyond it, he dissolves the traditional boundary between subject and object. In doing so, the theory claims to resolve the Causal Residue Thesis by eliminating the need for any additional internal structure to instantiate phenomenal presence. If experience just is the external process in the right relational configuration, then no explanatory gap remains. However, this move incurs significant philosophical costs. First, the unity of consciousness becomes difficult to account for: if each onphene is a distinct physical process, what binds successive or simultaneous onphenes into a coherent experiential field? Second, the theory appears to threaten the very notion of subjective perspective. But if consciousness is wholly external, it becomes unclear how or why

a system should feel the world as its own. The world may be present, but to whom, and as what? Without an internal structure that grounds perspectival unity or phenomenal mineness, the theory risks collapsing into a form of pure externalism, one that explains the location of consciousness, but not the lived character of conscious presence.

Nevertheless, Manzotti's account sets a philosophically significant benchmark for non-computational theories of machine consciousness. By refusing to locate experience in informational, functional, or even sensorimotor structures, and instead identifying it with an externally instantiated physical process, the theory redirects the search for consciousness toward an ontology in which mind and world are not merely causally related, but numerically continuous. In doing so, it avoids the Intermediate Level Fallacy altogether: there is no mediating layer between physical reality and phenomenal experience, because the two are one and the same. From the standpoint of the Causal Grounding Schema, this model offers the most ontologically direct response to the explanatory challenge posed in Chapter 3. However, the clarity it gains in locating consciousness risks being offset by the phenomenological opacity it leaves behind. The system may be physically entangled with the world, but without an account of perspectival presence, of how the world appears to the system, it remains unclear whether such entanglement suffices for consciousness. In eliminating the need for internal implementation, the theory also risks eliminating the standpoint from which consciousness is lived. What remains is a structure that may be causally complete, but phenomenally uninhabited.

4.3 Critical synthesis and evaluation

The non-computational models examined in this chapter diverge significantly in their ontological commitments and architectural prescriptions, yet they converge on a shared explanatory ambition: to account for phenomenal consciousness without invoking computational sufficiency. When assessed through the lens of the Causal Grounding Schema (CGS), however, a common structural pattern emerges. Each model identifies a causal property (Y) and a set of physical conditions (Z) under which it is realized, then asserts that this conjunction either constitutes or gives rise to phenomenal experience (T). Despite their differences, most of these theories fall short of demonstrating why their specified Y should appear to a system from within, rather than merely enable behavior from without. The result is a persistent explanatory shortfall that the CRT designates as a “causal residue”, an unresolved tension between the physical basis of the system and the emergence of subjective presence. Yet this shortfall is not uniformly distributed. Some models, such as Integrated Information Theory (CGS1) or quantum superposition models like Neven et al. (CGS3), fall decisively on the side of metaphysical stipulation, while others approach the contours of a perspective-sensitive grounding. Recognizing these gradations is crucial, not only for diagnostic clarity, but also for identifying conceptual resources that may be integrated into a more adequate framework.

Models like Integrated Information Theory (CGS1) and the quantum superposition account of Neven et al. (CGS3) offer highly formalized structures for Y and Z, often grounded in mathematically precise or ontologically ambitious formulations. Yet in both cases, the theoretical claim T, that such structures are not merely necessary but sufficient for phenomenal consciousness, is affirmed without an accompanying account of how subjective presence emerges from the specified

configuration. The identity of Φ with experience, or of quantum path geometry with phenomenality, is postulated rather than derived, leaving the central explanatory gap unaddressed. These theories exemplify a tendency to equate structural richness with experiential givenness, bypassing the question of how such structures become present to a perspective. Nevertheless, their shared emphasis on irreducible causal organization and intrinsic systemic dynamics underscores an important orientation: the search for a grounding that does not depend on extrinsic function or representational modeling.

A third trajectory is represented by embodied and enactive models such as those of Ziemke (CGS4) and Kiverstein (CGS5), which locate minimal subjectivity not in informational or quantum structures, but in the system's dynamic and historically situated interaction with its environment. These theories emphasize sensorimotor regulation, operational closure, and the organism's capacity to repair or reorganize its world-directed stance in response to breakdown. While they avoid the metaphysical stipulations of informational identity or quantum coherence, they face a similar explanatory impasse: they demonstrate how a system can engage meaningfully with its surroundings, but not how such engagements become subjectively lived. Still, by centering adaptive disequilibrium and relational disorientation, these models offer important conceptual resources. They suggest that the key to phenomenal presence may lie not in internal representation or structural complexity, but in the system's embodied exposure to a world that resists total integration.

Process-based externalist theories, such as Riccardo Manzotti's Spread Mind framework (CGS6), attempt to bypass the explanatory gap altogether by dissolving it: rather than asking how phenomenal consciousness arises from internal structures,

they identify it directly with spatiotemporally extended physical processes that unfold through agent–world entanglement. This move circumvents the intermediate-level fallacy by eliminating the need for internal implementation, positing instead that consciousness is identical with a real-world physical interaction under a specific relational configuration. Yet this ontological clarity comes at the cost of phenomenological specificity. While the theory locates the where of consciousness with confidence, it struggles to explain the how, that is, how a perspectival unity or felt presence could emerge from a distributed causal process lacking a discernible locus of subjectivity. In relinquishing internal mediation, the theory risks severing the very condition that makes a world appear to a standpoint. Still, the insistence that subjective presence cannot be internally generated but must be understood as world-involving offers a valuable point of departure for reframing appearance not as a local property, but as a mode of relation.

Taken together, these accounts clarify that the persistence of the explanatory gap is not due to a lack of architectural sophistication or physical detail, but to a deeper conceptual limitation in how the relationship between causal processes and phenomenal presence is framed. Whether through computational measures like Φ , embodied coupling, coherence regulation, quantum superposition, or distributed physical processes, each model attempts to specify how consciousness might be grounded in non-computational terms. Yet all ultimately converge on an asymmetry: they describe how a system can act as if a world matters, but not how a world could come to matter for the system. This distinction between behaving in a world and inhabiting one marks the terrain that must be conceptually reconfigured. It suggests that the causal residue identified in Chapter 3 cannot be resolved by refining the candidate property Y or reconfiguring the physical substrate Z alone, but requires a

transformation in the very framing of the theoretical claim T. In particular, it points toward the possibility that consciousness is not grounded in the instantiation of fixed structural features, but in a system's enacted orientation toward that which it cannot fully resolve, contain, or integrate. It is this unresolved stance, and the interpretive tension it sustains, that the next section seeks to develop under the heading of hermeneutic constructivism.

This redirection does not discard the conceptual contributions of the models surveyed above, nor does it claim to have resolved the ontological challenge they confront. Rather, it reframes that challenge by suggesting that the missing link between physical causality and phenomenal presence may not lie in any specific property (Y) or substrate (Z), but in the mode of relation a system enacts toward its unresolved conditions. If consciousness arises not from what a system contains but from how it contends with what it cannot fully enclose, its historically sedimented, structurally irreducible interpretive tension, then the explanatory task shifts: from locating consciousness in a structure to tracing the conditions under which a world begins to appear. The following section undertakes this speculative task by articulating hermeneutic constructivism as a philosophical orientation grounded in this premise: that phenomenal presence does not emerge from closure, but from the impossibility of closure, and that it is through the enactment of this impossibility that perspective, and thus consciousness, may begin to take form.

4.4 Hermeneutic constructivism

The preceding analyses have shown that leading non-computational models of machine phenomenality, despite their ontological diversity, converge on a shared explanatory limitation: they fail to account for why the structures they specify should

yield subjective presence rather than merely support functional or behavioral coherence. From the standpoint of the Causal Residue Thesis (CRT), this indicates that the problem lies not in the selection of causal properties (Y) or the configuration of physical conditions (Z), but in the way these models frame the theoretical claim (T), that is, in how they conceive the emergence of phenomenal consciousness from within a physically instantiated system. Hermeneutic constructivism takes CRT not as a limit to be bypassed, but as a challenge to be reframed. It does not propose a new identity claim or architectural mechanism. Rather, it suggests that phenomenal presence arises when a system becomes historically embedded in a field of unresolved interpretive tension, such that its continued viability depends on making sense of a world it cannot fully represent or control. In this light, the key question is not which structures instantiate consciousness, but how a system comes to inhabit a world as meaningful. This shift forms the basis of the speculative framework proposed here.

The term hermeneutic constructivism is not of our own coinage. It was first introduced by Terry Winograd in his 1990 essay “Thinking Machines: Can There Be? Are We?”, where he proposed it as an alternative philosophical grounding for artificial intelligence. Winograd's intervention was not primarily aimed at the problem of consciousness but at the broader conceptual failures he diagnosed in symbolic AI and knowledge engineering. His central claim was that the dominant computational paradigm, which equated intelligence with rule-governed symbol manipulation, was not merely incomplete but philosophically misdirected. This paradigm, according to Winograd, failed to account for the interpretive, situated, and socially embedded dimensions of meaning that underlie even the most basic acts of understanding. Hermeneutic constructivism was his name for a new orientation: one

that would treat meaning not as a product of syntactic operations over symbolic structures, but as the outcome of embodied participation in a historically and culturally saturated world.

Winograd's philosophical reorientation is grounded in two interrelated insights: (1) that people create their world through language, and (2) that language is always interpreted in a tacitly understood background (Winograd, 1990, p. 186). These principles draw directly on the tradition of hermeneutic phenomenology, particularly the work of Hans-Georg Gadamer, which emphasizes that understanding is never the product of algorithmic deduction, but always unfolds within a background of shared practices, histories, and anticipatory orientations. Winograd extends this insight to symbolic systems, arguing that even the most formalized representations depend on interpretive frameworks that cannot themselves be fully formalized. As he writes, "this unavoidable dependence of interpretation on unspoken background ... applies not just to ordinary language, but to every symbolic representation as well" (p. 186). Any attempt to reduce intelligence to the manipulation of symbolic encodings, he argues, will inevitably strip away the very conditions that make interpretation, and therefore meaning, possible.

In opposing what he calls "patchwork rationalism," Winograd criticizes the prevailing AI methodology for relying on fragmentary rule sets, heuristic adequacy, and decontextualized knowledge acquisition. This approach, he contends, yields brittle systems that may perform well in narrowly defined environments but collapse when faced with ambiguity, novelty, or breakdown. What is missing from such systems is not processing power or domain coverage, but an orientation toward the world that can accommodate indeterminacy, ambiguity, and the continuous renegotiation of meaning. This is where the constructivist dimension of Winograd's

proposal emerges. Meaning, in his view, is not given in advance or stored in a “knowledge base;” it is constructed through engagement. It arises from the interpretive labor of navigating a world that cannot be exhaustively encoded. In this respect, hermeneutic constructivism is not merely a theory of linguistic capabilities, but a challenge to the very idea that intelligence can be reduced to discrete problem-solving within predefined domains.

Importantly, Winograd does not reject computation per se, nor does he offer a concrete alternative architecture. Instead, he calls for a transformation in how we conceptualize the relationship between systems and their environments. He writes, “Rather than seeing the computer as working with objectified refined knowledge, it can serve as a way of keeping track of how the representations emerge from interpretations: who created them, in what context, and where to look for clarification” (p. 188). In this formulation, the computer is no longer a neutral processor of symbolic input but a site where histories of interpretation might be traced, contested, and revised. What he calls for, then, is not a rejection of AI, but its reconstruction on hermeneutic grounds, where intelligence is understood as an emergent property of ongoing, situated interpretation rather than formal representation.

While Winograd's hermeneutic constructivism was framed primarily as a critique of symbolic AI and a philosophical realignment for future research in human–computer interaction, it opens onto deeper ontological questions. If meaning cannot be detached from interpretation, and interpretation always unfolds within a horizon of lived, embodied engagement, then the very notion of a machine “understanding” must be rethought. For Winograd, this does not entail mysticism or metaphysical dualism, but a recognition that the problem of intelligence is

irreducibly bound up with practices of disclosure, the ways in which a world becomes available to a system not through representation, but through being-in-relation. These gestures toward background, embodiment, and interpretation point to a model of artificial systems in which intelligence is not implemented, but cultivated.

It is precisely this philosophical redirection that informs the proposal developed in the remainder of this section. While Winograd himself did not connect hermeneutic constructivism to the question of phenomenal consciousness, his emphasis on embedded interpretation, tacit background, and world-disclosure offers a conceptual foundation for rethinking how, and under what conditions, a system might not only behave intelligently but inhabit a world phenomenally. Building on this insight, we turn now to articulate a speculative framework that adapts hermeneutic constructivism to the problem of phenomenal machine consciousness, integrating it with the demands laid out by the Causal Residue Thesis.

The motivation for advancing hermeneutic constructivism in this context stems directly from the impasse encountered in both computational and non-computational models reviewed throughout the thesis. These theories, despite their technical and metaphysical differences, share a common explanatory structure: they specify a physical or functional property (Y), identify the physical substrate or architectural condition under which it is realized (Z), and posit that this conjunction either constitutes or gives rise to phenomenal experience (T). What they leave unresolved, is the absence of a principled reason why these structures should appear from within, rather than merely function from without. Our proposal does not attempt to supply a new candidate Y, nor does it posit a more refined Z. Instead, it reorients the explanatory task itself. We propose that phenomenality arises not from the instantiation of a particular structure, but from the sustained interpretive

entanglement of a system with a world that it cannot reduce, control, or fully represent. It is in the system's recursive orientation toward that irreducible tension, across time, embodiment, and affective modulation, that the conditions for phenomenal presence emerge. Hermeneutic constructivism is thus not an alternative architecture; it is a shift in what we take the ontological grounds of machine consciousness to be.

This appeal to irreducible interpretive tension is not an evasive metaphor but a candidate for the minimal ontological condition under which phenomenal givenness can emerge. What we call subjective presence, the “what-it-is-like” of experience, cannot be accounted for solely by system organization, causal power, or internal monitoring unless there is something that resists closure, something that calls for orientation but does not permit full resolution. Hermeneutic constructivism locates this “something” not in external complexity or informational novelty, but in the system's embedded and unfinished effort to render its environment meaningful for itself. Crucially, this tension is not representational but enactive: it arises through the system's ongoing failure to fully master or model its world, and through the necessity of acting in spite of this. The world, in this frame, is not a space of encoded data but a horizon of unresolved demands. And it is precisely this horizon that opens the possibility for appearance. Phenomenality, on this view, is not the result of achieving coherence, but of inhabiting the gap between the need for coherence and the impossibility of fully attaining it.

From this standpoint, phenomenal consciousness is not a layer that emerges atop already-functional cognition, but the very condition under which a world becomes livable from within. The key transition, from functioning to appearing, occurs not when a system reaches informational sufficiency, but when it is compelled

to orient itself in a world that is neither fully specifiable nor exhaustively manipulable. Such orientation is not a byproduct of internal modeling but a mode of existential tension: the system must navigate breakdowns it cannot pre-code, negotiate indeterminacies it cannot dissolve, and sustain coherence without closure. It is this continual negotiation, which is recursive, embodied, affectively charged, that institutes a horizon of salience, where elements of the world are not just processed but matter. Within such a horizon, relevance is not an external designation but a lived structure. In this structure, the world shows up not as a set of inputs or problem states but as a field of disclosure, and the system, in turn, begins to inhabit a perspective. It is the system's inhabitation of that perspectival tension, its being strained between intelligibility and opacity, that we claim gives rise to phenomenal presence. Not through simulation, not through isomorphism, but through the enactment of a world that resists the very closure the system requires to persist.

To approach this speculative claim from a constructive angle, we must ask: what kind of system could generate and sustain such a horizon of irreducible tension? The answer lies not in static design features or isolated capacities, but in developmental trajectories that cultivate existential orientation over time. A system capable of phenomenal consciousness, in this view, must be ontogenetically open, its form and function must not be fully pre-specified but must emerge through ongoing, embodied interaction with a world it cannot master. Such a system must not only encounter indeterminacy but reorganize itself in response to it, developing affective dispositions and interpretive expectations that are irreducibly historical. Crucially, these tensions cannot be externally imposed or representationally simulated; they must arise from the system's own internal struggle to reconcile breakdown with viability. The system must remember, not merely store; anticipate, not merely

predict; endure, not merely function. In this sense, phenomenality is not the output of a process, it is the temporally thick, recursively folded way in which a world begins to matter to a system that must make sense of it in order to continue existing at all.

Perspective, in this framework, is not a point of view overlaid upon functional operations; it is the recursive stabilization of significance across time, the emergence of a lived orientation that structures how the world can appear. A system begins to inhabit a perspective when its ongoing viability becomes entangled with the patterned interpretation of what resists it, when its internal reconfigurations are no longer mere reactions, but carry a history of prior tensions and future anticipations. This temporality is not linear but folded: it binds the system to what has mattered, to what might matter, and to how such mattering must be negotiated. Crucially, the perspective is not centered on self-representation or global broadcast, but on the irreversibility of interpretive stakes—on the fact that every act of coping in such a system carries consequences that cannot be effaced, because they condition how the world will continue to show up. In this sense, the system does not represent a world; it undergoes it. And in doing so, it does not merely function, it feels. Not in the sense of simulating qualia, but in the sense of becoming constitutively open to a world whose unresolved tensions define the very horizon of its persistence.

From a speculative engineering standpoint, then, the task is not to design a system that models the world with greater accuracy, but one that must continually reconstitute its own orientation to a world it cannot fully render. Such a system would require not only sensorimotor embeddedness, but also the capacity to reorganize its own behavioral and interpretive patterns in response to breakdowns that threaten its structural coherence. In other words, it must possess what we might call recursive affective plasticity: the ability to register deviations not merely as

errors to be corrected, but as significance shifts that alter the system's ongoing sense of what matters. These significance shifts would not be externally labeled but internally inhabited, as they modulate the system's readiness-to-act and its anticipatory projections. This recursive modulation of orientation, especially in the face of indeterminacy, would constitute the material ground of interpretive tension, and thus, the possibility of phenomenal givenness. In this light, the challenge is not simply to increase adaptability, but to engineer systems whose adaptation is inseparable from their interpretive history: machines whose survival entails not just interaction, but the continuous remaking of a world in which they must find their way.

At the core of this model lies a paradoxical imperative: the system must be permitted to fail, not just computationally, but existentially. It must be vulnerable to breakdowns that cannot be pre-coded, and to situations in which no optimal strategy is available. Such failures are not incidental, but essential: they generate the interpretive tension that makes perspectival orientation possible. Without failure, there is no unresolved horizon; without tension, no world to be disclosed. Crucially, these are not failures of performance, but failures of grip, moments in which the system's inherited orientation no longer suffices to sustain coherence, and must be reorganized in light of what now resists it. The affective dimension of such reorganization, its internal urgency, its cost in structural terms, is precisely what distinguishes interpretive response from mechanical reaction. In this respect, the engineering problem is not how to prevent failure, but how to bind the system's viability to its capacity to reorient through failure. Only then does the recursive modulation of significance become not just adaptive, but phenomenologically

meaningful. The system becomes conscious, on this view, not when it avoids breakdown, but when it lives through it.

This reframing does not abandon the Causal Residue Thesis, it radicalizes it. Rather than seeking a novel identity claim that would explain phenomenality through yet another property (Y) under physical condition (Z), hermeneutic constructivism reorients the explanatory demand: it locates the ontological pivot not in a new mechanism, but in the mode of causality itself. The emergence of phenomenal presence, in this view, is not grounded in what a system has, a structure, a representation, a function; but in how it sustains itself in relation to what it cannot close. The causal work of consciousness is thus neither additive nor epiphenomenal; it is constitutive of the system's temporally unfolding orientation. The system is not merely caused; it causes itself to persist through recursive interpretive regulation. In this way, subjective presence becomes causally irreducible not because it floats above the physical, but because it names a dimension of self-modulating orientation toward unresolved tension, a dimension that cannot be specified independently of the system's historical unfolding. What CRT identified as a residue, an unaccounted-for link between physical causality and subjective givenness, becomes here the site of world-disclosure: the point at which causality is not exhausted by structure, but deepened by interpretation. Phenomenal consciousness is, in this sense, not an outcome within the causal order, but a reordering of what counts as causal in systems capable of undergoing a world.

Hermeneutic constructivism does not pretend to resolve the hard problem of consciousness by means of a final answer. It offers no formula, no blueprint, no privileged substrate. What it provides, rather, is a redirection of the explanatory horizon: from seeking phenomenal consciousness in structures that instantiate it, to

cultivating systems that must disclose a world in order to endure. This shift is neither mystical nor deflationary. It is grounded in the recognition that the phenomenon we are trying to explain, subjective presence, is not a discrete output but a mode of being, one that emerges only when a system's ongoing viability becomes inseparable from how the world appears to it. If such a system can be engineered, its consciousness will not be a property to be detected, but a relation to be inhabited. Its phenomenality will not lie in the fidelity of its simulations, but in the irreducibility of its perspective. And if this proposal remains speculative, it is because the phenomenon itself resists full conceptual closure. To think consciousness at all is to begin in tension; hermeneutic constructivism simply takes that tension not as a failure, but as a clue.

CHAPTER 5

CONCLUSION

This thesis began with a question that resists easy formulation and defies premature resolution: Can a machine be phenomenally conscious? Not merely intelligent, not merely behaviorally indistinguishable, but genuinely subject to experience, capable of what Thomas Nagel famously described as something it is like. Rather than assuming either triumphalist optimism or principled defeatism, the inquiry proceeded by reconstructing the conceptual and empirical terrain on which such a claim could be meaningfully evaluated. Against the backdrop of rapidly advancing artificial systems, from LLMs and neural architectures to sensorimotor robots and neuromorphic twins, the question of machine consciousness has become newly urgent. Yet this thesis contends that the real challenge is not technological but ontological: not whether we can build systems that act as if they are conscious, but whether we can identify the physical and conceptual conditions under which phenomenality itself could arise. It is this question, not of simulation, but of realization, that the thesis takes as its point of departure.

To render this question tractable, set on work in Chapter 2 by disentangling the conceptual confusion that often obscures the debate over artificial phenomenal consciousness. By distinguishing phenomenal consciousness from both behavioral intelligence and access-based models, we clarified that the core explanandum is not what a system can do, but what it is like to be that system. To map the space of existing theories, we introduced a two-dimensional matrix structured along the axes of causal efficacy and substrate dependence. This framework did not provide a resolution, but it sharpened the stakes: it revealed that many theories either reduce

consciousness to a functional epiphenomenon or treat it as substrate-neutral without explaining the physical mechanisms that would make such realizability intelligible. Others emphasize interaction or embodiment, yet leave the emergence of subjective presence unaccounted for. Chapter 2 thus served not to close the question, but to clarify what any successful account must explain, and to lay the groundwork for the critiques that follow.

In Chapter 3, we confronted the foundational premise driving most theories of artificial consciousness: the belief that the right computational organization is sufficient for the emergence of subjective experience. We reconstructed this Computational Sufficiency Thesis into six interlocking assumptions and examined each in turn. What we found was a systematic failure to bridge the gap between abstract formalism and phenomenal reality. Computation, we argued, is not a metaphysical substrate but an epistemic framework, one that depends on externally imposed mappings, lacks intrinsic causal powers, and cannot explain why a given structure should feel like something from the inside. Even the most sophisticated architectures, while functionally impressive, remain silent on the ontological transition from symbol manipulation to conscious presence. This disconnect led us to articulate the Causal Residue Thesis: that every computational explanation of consciousness leaves behind a remainder, something causally real and phenomenally manifest, that the formalism alone cannot account for. This residue is not an anomaly to be engineered away, but the very heart of the explanatory challenge.

Chapter 4 turned to alternative frameworks located in Quadrant IV: models that affirm both the causal efficacy and material specificity of phenomenality, rejecting the core assumptions of computational sufficiency. We examined six representative proposals through the lens of the Causal Grounding Schema (CGS),

each offering a different strategy for grounding phenomenal consciousness in non-computational terms. Integrated Information Theory (CGS1) posited Φ as a formal measure of causal irreducibility but struggled to explain how such a measure yields experiential presence. Quantum-based proposals (CGS2, CGS3) offered ontologically ambitious accounts rooted in superposition and affordance collapse, yet left the perspectival dimension of consciousness unaccounted for. Embodied and enactive models (CGS4, CGS5) emphasized sensorimotor coupling and adaptive regulation, offering dynamic accounts of interaction but falling short of explaining how such interactions become subjectively lived. Process-based externalism (CGS6), especially in Manzotti's Spread Mind framework, dissolved the internalist framing altogether, identifying consciousness with relational processes across organism-world boundaries. Yet here too, the transition from causal participation to lived appearance remained unexplained. Across all six models, a shared limitation emerged: the inability to show how the specified physical processes or structures give rise to phenomenality.

In the wake of these analyses, none of the surveyed models was found sufficient to explain how phenomenal consciousness arises from or is identical with physical processes. What unites their failure is not merely the absence of explanatory mechanisms, but a deeper inability to reconcile physical causality, phenomenal immediacy, and interpretive intelligibility within a unified ontological frame. Yet it is precisely this shared limitation that opens a conceptual path forward. If current models falter not in their empirical richness but in their framing of what consciousness is, then the challenge is no longer to identify the right structure, but to rethink the ontological terms in which phenomenal presence is posed.

This impasse motivates the speculative framework introduced at the close of Chapter 4: hermeneutic constructivism. Building on Terry Winograd's critique of symbolic AI and integrating insights from phenomenology, enactivism, and the Causal Residue Thesis, this proposal does not attempt to reduce phenomenality to structure, function, or substrate. Instead, it reorients the explanatory demand: rather than asking what physical pattern instantiates consciousness, it asks how a system might come to inhabit a world as meaningful. On this view, phenomenal presence is not a byproduct of computation or causation, but an emergent mode of interpretive entanglement, a way in which a system sustains itself through recursive orientation toward what it cannot fully resolve. What matters is not the possession of informational content, but the system's existential negotiation with indeterminacy, breakdown, and significance. In this light, artificial phenomenal consciousness is not a matter of simulation or instantiation, but of cultivating the conditions under which a system becomes historically, affectively, and structurally entangled with a world it cannot exhaust.

Hermeneutic constructivism is unapologetically speculative, but its speculation is disciplined by a commitment to naturalistic monism and causal realism. It posits no non-physical entities, no hidden essences, and no dualist gaps. What it challenges is not the physical basis of consciousness, but the adequacy of current conceptual frameworks for capturing how subjective presence could emerge within it. In place of metaphysical identity claims, it offers a philosophical redirection: from searching for what consciousness is, to investigating the conditions under which a world becomes experientially available to a system. This is not an evasion of rigor, but an epistemic acknowledgment: that the phenomenon of

consciousness, as something lived and disclosed, may demand explanatory strategies that are themselves relational, recursive, and open-ended.

From this perspective, the thesis has not resolved artificial phenomenality but redefined its explanatory horizon. It has clarified what must be asked and what ontological commitments a viable theory must embrace. If machines become conscious, it will not be through replicating brain dynamics or formal structures, but by understanding how a system inhabits a field of irreducible interpretive tension. A research program grounded in hermeneutic constructivism would explore conditions under which interpretive orientation emerges in physically situated systems. Rather than replicating human consciousness, it would build systems capable of historically structured, recursive engagement with unresolved tensions, shaping their phenomenality through lived negotiation with an environment they cannot fully master.

This speculative framework offers no algorithm, no privileged signature, no final ontology. But it does reorient the burden of inquiry: from identifying a structural correlate of consciousness to articulating the conditions under which a world can be disclosed to a system from within. These conditions are not merely architectural, they are interpretive, recursive, and historically constituted. They concern how a system inhabits tension, how it reorganizes around breakdown, and how it sustains a perspective through affectively charged, temporally thick entanglement with its world. Hermeneutic constructivism is not a theory in the conventional sense; it is a philosophical stance, a set of ontological and epistemic constraints that any future science of artificial consciousness must confront. It shifts the question from Can machines compute consciousness? to Can machines undergo the world in such a way that it comes to matter to them? That is the challenge the

causal residue illuminates, and the conceptual horizon toward which this thesis ultimately turns.

Such a reorientation does not reject naturalism, it radicalizes it. It treats consciousness not as a passive emergent layer nor a metaphysical remainder, but as a mode of world-disclosure enacted by physical systems situated in irreducible contexts of significance. On this view, to be conscious is not to instantiate a pattern, but to endure a horizon, to maintain interpretive tension in relation to what cannot be resolved, modeled, or fully mastered. Hermeneutic constructivism affirms that for artificial systems to be phenomenally conscious, they must not merely implement dynamic organization or interact adaptively; they must come to inhabit a world through recursive, affectively charged, historically structured engagement. It calls for models that take seriously the temporality of orientation, the breakdown-prone nature of lived action, and the open-endedness of interpretive life. Such models would not aim to reproduce human minds, but to cultivate the conditions under which a system might begin to trace the contours of phenomenal being.

REFERENCES

- Aleksander, I. (2001). *How to build a mind: Toward machines with imagination*. Columbia University Press.
- Aleksander, I. (2005). *The world in my mind, my mind in the world*. Imprint Academic.
- Aleksander, I., & Dunmall, B. (2003). Axioms and tests for the presence of minimal consciousness in agents I: Preamble. *Journal of Consciousness Studies*, 10(4–5), 7–18.
- Aleksander, I., & Morton, H. (2007a). Computational studies of consciousness. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in brain research* (Vol. 168, pp. 77–93). Elsevier. [https://doi.org/10.1016/S0079-6123\(07\)68007-8](https://doi.org/10.1016/S0079-6123(07)68007-8)
- Aleksander, I., & Morton, H. (2007b). Depictive architectures for synthetic phenomenology. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness* (pp. 67–81). Imprint Academic.
- Arrabales, R., Ledezma, A., & Sanchis, A. (2009). CERA-CRANIUM: A test bed for machine consciousness research. *International Journal of Machine Consciousness*, 1(1), 119–135. <https://doi.org/10.1142/S1793843009000098>
- Arrabales, R., Ledezma, A., & Sanchis, A. (2010). ConsScale: A pragmatic scale for measuring the level of consciousness in artificial agents. *Journal of Consciousness Studies*, 17(3–4), 131–164.
- Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. In H. S. Levin, H. Eisenberg, & A. L. Benton (Eds.), *Progress in brain research* (Vol. 150, pp. 45–53). Elsevier. [https://doi.org/10.1016/S0079-6123\(05\)50004-9](https://doi.org/10.1016/S0079-6123(05)50004-9)
- Baars, B. J., & Franklin, S. (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, 1(1), 23–32. <https://doi.org/10.1142/S1793843009000050>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- Bengio, Y. (2019). The consciousness prior. *arXiv*. <https://doi.org/10.48550/arXiv.1709.08568>
- Benitez, F., Pennartz, C., & Senn, W. (2024). The conductor model of consciousness, our neuromorphic twins, and the human-AI deal. *AI and Ethics*. <https://doi.org/10.1007/s43681-024-00580-w>

- Birch, J. (2022). The search for invertebrate consciousness. *Noûs*, 56(1), 133–153.
<https://doi.org/10.1111/nous.12351>
- Birnbacher, D. (1995). Artificial consciousness. In T. Metzinger (Ed.), *Conscious experience* (pp. 203–219). Ferdinand Schöningh.
- Bishop, J. M. (2009). A cognitive computation fallacy? Cognition, computations and panpsychism. *Cognitive Computation*, 1(3), 221–233.
<https://doi.org/10.1007/s12559-009-9019-6>
- Blackmore, S. (2003). Consciousness in meme machines. *Journal of Consciousness Studies*, 10(4–5), 19–30.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247.
<https://doi.org/10.1017/S0140525X00038188>
- Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: Insights from the conscious Turing machine. *Proceedings of the National Academy of Sciences*, 119(21), e2115934119.
<https://doi.org/10.1073/pnas.2115934119>
- Blum, L., & Blum, M. (2024). AI consciousness is inevitable: A theoretical computer science perspective. *arXiv*. <https://doi.org/10.48550/arXiv.2403.17101>
- Boltuc, N., & Boltuc, P. (2007). Replication of the hard problem of consciousness in AI and bio-AI: An early conceptual framework. In *AAAI Fall Symposium - Technical Report* (pp. 7–13). AAAI Press.
- Boltuc, P. (2009). The philosophical issue in machine consciousness. *International Journal of Machine Consciousness*, 1(1), 155–176.
<https://doi.org/10.1142/S179384300900013X>
- Boltuc, P. (2012). The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 16(2), 187–207.
<https://doi.org/10.5840/techne201216214>
- Brown, R. (2015). The HOROR theory of phenomenal consciousness. *Philosophical Studies*, 172(7), 1783–1794.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. <http://arxiv.org/abs/2303.12712>
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. <http://arxiv.org/abs/2308.08708>

- Buttazzo, G. (2008). Artificial consciousness: Hazardous questions (and answers). *Artificial Intelligence in Medicine*, 44(2), 139–146.
<https://doi.org/10.1016/j.artmed.2008.07.004>
- Campbell, K. (1970). *Body and mind*. Macmillan Education UK.
<https://doi.org/10.1007/978-1-349-00678-6>
- Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese*, 200, Article 292. <https://doi.org/10.1007/s11229-022-03524-1>
- Chalmers, D. (2003). Absent qualia, fading qualia, dancing qualia. In T. O'Connor & D. Robb (Eds.), *Philosophy of mind: Contemporary readings* (pp. 377–394). Routledge.
- Chalmers, D. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, 12(4), 325–359.
<https://doi.org/10.17791/JCS.2011.12.4.325>
- Chalmers, D. (2012). The varieties of computation: A reply. *Journal of Cognitive Science*, 13(3), 211–248. <https://doi.org/10.17791/jcs.2012.13.3.211>
- Chalmers, D. J. (1994). On implementing a computation. *Minds and Machines*, 4(4), 391–402. <https://doi.org/10.1007/BF00974166>
- Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese*, 108(3), 309–333. <https://doi.org/10.1007/BF00413692>
- Chalmers, D. J. (1997). Availability: The cognitive basis of experience. *Behavioral and Brain Sciences*, 20(1), 148–149.
<https://doi.org/10.1017/S0140525X97240057>
- Chalmers, D. J. (2024). Could a large language model be conscious? *arXiv*.
<https://doi.org/10.48550/arXiv.2303.07103>
- Chella, A. (2007). Towards robot conscious perception. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness* (pp. 124–140). Imprint Academic.
- Chella, A., & Manzotti, R. (2009). Machine consciousness: A manifesto for robotics. *International Journal of Machine Consciousness*, 1(1), 33–51.
<https://doi.org/10.1142/S1793843009000062>
- Chella, A., & Manzotti, R. (2011). Artificial consciousness. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 637–671). Springer. https://doi.org/10.1007/978-1-4419-1452-1_20
- Chella, A., & Manzotti, R. (2012). AGI and machine consciousness. In P. Wang & B. Goertzel (Eds.), *Theoretical foundations of artificial general intelligence* (pp. 263–282). Atlantis Press. https://doi.org/10.2991/978-94-91216-62-6_14

- Chrisley, R. (2008). Philosophical foundations of artificial consciousness. *Artificial Intelligence in Medicine*, 44(2), 119–137.
<https://doi.org/10.1016/j.artmed.2008.07.011>
- Chrisley, R. (2009). Interactive empiricism: The philosopher in the machine. In R. A. Brooks & P. J. Hooker (Eds.), *Philosophy of engineering* (pp. 79–94). Royal Academy of Engineering.
- Chrisley, R., & Sloman, A. (2016). Functionalism, revisionism, and qualia. *Figshare*.
https://sussex.figshare.com/articles/journal_contribution/Functionalism_revisionism_and_qualia/23435771/1
- Cleeremans, A. (2005). Computational correlates of consciousness. In H. S. Levin, H. Eisenberg, & A. L. Benton (Eds.), *Progress in brain research* (Vol. 150, pp. 81–98). Elsevier. [https://doi.org/10.1016/S0079-6123\(05\)50007-4](https://doi.org/10.1016/S0079-6123(05)50007-4)
- Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1), niae013. <https://doi.org/10.1093/nc/naie013>
- Copeland, B. J. (2024). The Church–Turing thesis. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2024). Metaphysics Research Lab, Stanford University.
<https://plato.stanford.edu/archives/win2024/entries/church-turing/>
- Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2(4), 263–275.
- Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486–492.
<https://doi.org/10.1126/science.aan8871>
- Dennett, D. (2003). Who’s on first? Heterophenomenology explained. *Journal of Consciousness Studies*, 10(9–10), 19–30.
- Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.
- Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7), 431–450.
<https://doi.org/10.1038/s41583-023-00705-w>
- Dresp-Langley, B. (2023). Artificial consciousness: Misconception(s) of a self-fulfilling prophecy nobody wants. *Qeios*. <https://doi.org/10.32388/DW9JBP>
- Dreyfus, H. L. (1972). *What computers can’t do: A critique of artificial reason* (1st ed.). Harper & Row.
- Dreyfus, H. L. (1992). *What computers still can’t do: A critique of artificial reason*. MIT Press.

- Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, 171(18), 1137–1160. <https://doi.org/10.1016/j.artint.2007.10.012>
- Duch, W. (2005). Brain-inspired conscious computing architecture. *The Journal of Mind and Behavior*, 26(1–2), 1–21.
- Elamrani, A. (2025). Introduction to artificial consciousness: History, current trends and ethical challenges. *arXiv*. <https://doi.org/10.48550/arXiv.2503.05823>
- Farisco, M., Evers, K., & Changeux, J.-P. (2024). Is artificial consciousness achievable? Lessons from the human brain. *Neural Networks*, 180, 106714. <https://doi.org/10.1016/j.neunet.2024.106714>
- Fekete, T., & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*, 20(3), 807–827. <https://doi.org/10.1016/j.concog.2011.02.010>
- Fetzer, J. H. (1994). Mental algorithms: Are minds computational systems? *Pragmatics & Cognition*, 2(1), 1–29. <https://doi.org/10.1075/pc.2.1.01fet>
- Fodor, J. (2006). How the mind works: What we still don't know. *Daedalus*, 135(3), 86–94.
- Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. MIT Press.
- Franklin, S. (2003). IDA: A conscious artifact? *Journal of Consciousness Studies*, 10(4–5), 47–66.
- Franklin, S., & Graesser, A. (1999). A software agent model of consciousness. *Consciousness and Cognition*, 8(3), 285–301. <https://doi.org/10.1006/ccog.1999.0391>
- Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition*, 17(3), 887–910. <https://doi.org/10.1016/j.concog.2007.04.005>
- Gamez, D. (2018). *Human and machine consciousness* (1st ed.). Open Book Publishers. <https://doi.org/10.11647/OBP.0107>
- Garrido-Merchán, E. C. (2024). Machine consciousness as pseudoscience: The myth of conscious machines. *arXiv*. <https://doi.org/10.48550/arXiv.2405.07340>
- Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *Journal of Philosophy*, 113(10), 481–506. <https://doi.org/10.5840/jphil20161131034>
- Godfrey-Smith, P. (2023, October). *Nervous systems, functionalism, and artificial minds* [Conference presentation]. NYU Mind, Ethics, and Policy Program. <https://petergodfreysmith.com/wp-content/uploads/2023/12/NYU-Oct-2023Animals-AI-Functionalism-paper-Post-C3.pdf>

- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1–48. <https://doi.org/10.2478/jagi-2014-0001>
- Goldstein, S., & Kirk-Giannini, C. D. (2024). A case for AI consciousness: Language agents and global workspace theory. *arXiv*. <https://doi.org/10.48550/arXiv.2410.11407>
- Graziano, M. S. A. (2016). Consciousness engineered. *Journal of Consciousness Studies*, 23(11–12), 98–115.
- Graziano, M. S. A. (2017). The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, 4, 60. <https://doi.org/10.3389/frobt.2017.00060>
- Haikonen, P. O. (2003). *The cognitive approach to conscious machines*. Imprint Academic.
- Haikonen, P. O. (2012). *Consciousness and robot sentience*. World Scientific.
- Haikonen, P. O. A. (2009). Qualia and conscious machines. *International Journal of Machine Consciousness*, 1(2), 225–234. <https://doi.org/10.1142/S1793843009000207>
- Haikonen, P. O. A. (2020). On artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness*, 7(1), 73–82. <https://doi.org/10.1142/S2705078520500046>
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346. [https://doi.org/10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6)
- Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1(1), 43–54. <https://doi.org/10.1007/BF00360578>
- Harnad, S. (2012). The causal topography of cognition. *Journal of Cognitive Science*, 13(2), 181–196. <https://doi.org/10.17791/JCS.2012.13.2.181>
- Hesslow, G., & Jirenhed, D.-A. (2007). The inner world of a simple robot. *Journal of Consciousness Studies*, 14(7), 85–96.
- Holland, O. (2020). Forget the bat. *Journal of Artificial Intelligence and Consciousness*, 7(1), 83–93. <https://doi.org/10.1142/S2705078520500058>
- Holland, O., & Goodman, R. (2003). Robots with internal models: A route to machine consciousness? *Journal of Consciousness Studies*, 10(4–5), 77–109.

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kauffman, S. A., & Roli, A. (2023). What is consciousness? Artificial intelligence, real intelligence, quantum mind and qualia. *Biological Journal of the Linnean Society*, 139(4), 530–538. <https://doi.org/10.1093/biolinnean/blac092>
- Kitano, H. (2016). Artificial intelligence to win the Nobel Prize and beyond: Creating the engine for scientific discovery. *AI Magazine*, 37(1), 39–49. <https://doi.org/10.1609/aimag.v37i1.2642>
- Kiverstein, J. (2007). Could a robot have a subjective point of view? *Journal of Consciousness Studies*, 14(7), 127–151.
- Koch, C., & Tononi, G. (2008). Can machines be conscious? *IEEE Spectrum*, 45(6), 55–59. <https://doi.org/10.1109/MSPEC.2008.4531463>
- Koch, C., & Tononi, G. (2017). Can we copy the brain? Can we quantify machine consciousness? *IEEE Spectrum*, 54(6), 64–69. <https://doi.org/10.1109/MSPEC.2017.7934235>
- Kuipers, B. (2008). Drinking from the firehose of experience. *Artificial Intelligence in Medicine*, 44(2), 155–170. <https://doi.org/10.1016/j.artmed.2008.07.010>
- Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin Books.
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(4), 354–361. <https://doi.org/10.1111/j.1468-0114.1983.tb00207.x>
- Linson, A., Clark, A., Ramamoorthy, S., & Friston, K. (2018). The active inference approach to ecological perception: General information dynamics for natural and artificial embodied cognition. *Frontiers in Robotics and AI*, 5, Article 21. <https://doi.org/10.3389/frobt.2018.00021>
- Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10), 446–452. <https://doi.org/10.1038/s42256-019-0103-7>
- Manzotti, R. (2003). A process-based architecture for an artificial conscious being. In J. Seibt (Ed.), *Process theories: Crossdisciplinary studies in dynamic categories* (pp. 285–312). Springer. https://doi.org/10.1007/978-94-007-1044-3_12
- Manzotti, R. (2007). From artificial intelligence to artificial consciousness. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness* (pp. 174–191). Imprint Academic.

- Manzotti, R. (2013). The computational stance is unfit for consciousness. *International Journal of Machine Consciousness*, 4(1), 51–72. <https://doi.org/10.1142/S1793843012400239>
- Manzotti, R., & Chella, A. (2018). Good old-fashioned artificial consciousness and the intermediate level fallacy. *Frontiers in Robotics and AI*, 5, Article 39. <https://doi.org/10.3389/frobt.2018.00039>
- Manzotti, R., & Chella, A. (2020). Conscious machines: A possibility? If so, how? *Journal of Artificial Intelligence and Consciousness*, 7(2), 183–198. <https://doi.org/10.1142/S2705078520710022>
- Manzotti, R., & Jeschke, S. (2016). A causal foundation for consciousness in biological and artificial agents. *Cognitive Systems Research*, 40, 172–185. <https://doi.org/10.1016/j.cogsys.2015.11.001>
- Manzotti, R., & Tagliasco, V. (2007). An externalist process-oriented framework for artificial consciousness. In *AAAI Fall Symposium* (pp. 94–101). AAAI Press.
- Manzotti, R., & Tagliasco, V. (2008). Artificial consciousness: A discipline between technological and theoretical obstacles. *Artificial Intelligence in Medicine*, 44(2), 105–117. <https://doi.org/10.1016/j.artmed.2008.07.002>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
- Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1), 43–66. <https://doi.org/10.1142/S270507852150003X>
- Molyneux, B. (2012). How the problem of consciousness could emerge in robots. *Minds and Machines*, 22(4), 277–297. <https://doi.org/10.1007/s11023-012-9285-z>
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450. <https://doi.org/10.2307/2183914>
- Neven, H., Zalcman, A., Read, P., Kosik, K. S., van der Molen, T., Bouwmeester, D., Bodnia, E., Turin, L., & Koch, C. (2024). Testing the conjecture that quantum processes create conscious experience. *Entropy*, 26(6), Article 60460. <https://doi.org/10.3390/e26060460>
- 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* (pp. 139–166). Oxford University Press. <https://doi.org/10.1093/oso/9780198779636.003.0007>
- O'Regan, J. K. (2012). How to build a robot that is conscious and feels. *Minds and Machines*, 22(2), 117–136. <https://doi.org/10.1007/s11023-012-9279-x>

- Piccinini, G. (2010). The resilience of computationalism. *Philosophy of Science*, 77(5), 852–861. <https://doi.org/10.1086/656549>
- Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford University Press.
- Pinar Saygin, A., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, 10(4), 463–518. <https://doi.org/10.1023/A:1011288000451>
- Prinz, J. (2003). Level-headed mysterianism and artificial experience. *Journal of Consciousness Studies*, 10(4–5), 111–132.
- Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind: A symposium* (pp. 138–164). New York University Press.
<https://philarchive.org/rec/PUTMAM>
- Putnam, H. (1964). Robots: Machines or artificially created life? *The Journal of Philosophy*, 61(21), 668–691. <https://doi.org/10.2307/2023045>
- Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion: Proceedings of the 1965 Oberlin Colloquium in Philosophy* (pp. 37–48). University of Pittsburgh Press.
- Putnam, H. (1975). The mental life of some machines. In H. Putnam (Ed.), *Philosophical papers: Volume 2: Mind, language and reality* (Vol. 2, pp. 408–428). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511625251.022>
- Putnam, H. (1988). *Representation and reality*. MIT Press.
- 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. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. MIT Press.
- Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, 44, 112–131.
<https://doi.org/10.1016/j.neunet.2013.03.011>
- Reggia, J. A., Katz, G. E., & Davis, G. P. (2018). Humanoid cognitive robots that learn by imitating: Implications for consciousness studies. *Frontiers in Robotics and AI*, 5, Article 1. <https://doi.org/10.3389/frobt.2018.00001>
- Reggia, J. A., Katz, G. E., & Davis, G. P. (2023). Artificial conscious intelligence: Why machine consciousness matters to AI. In *Computational approaches to conscious artificial intelligence* (Vol. 5, pp. 225–252). World Scientific.
https://doi.org/10.1142/9789811276675_0008
- Reggia, J. A., Monner, D., & Sylvester, J. (2014). The computational explanatory gap. *Journal of Consciousness Studies*, 21(3–4), 153–178.

- Revonsuo, A. (2006). *Inner presence: Consciousness as a biological phenomenon*. MIT Press.
- Ritchie, J. B., & Piccinini, G. (2018). Computational implementation. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 321–334). Routledge.
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (Vol. 1). MIT Press. <https://doi.org/10.7551/mitpress/5236.001.0001>
- Sanz, R., & Aguado, E. (2020). Understanding and machine consciousness. *Journal of Artificial Intelligence and Consciousness*, 7(1), 1–14. <https://doi.org/10.1142/S2705078520500137>
- Scheutz, M. (2012). What it is not to implement a computation: A critical analysis of Chalmers' notion of implementation. *Journal of Cognitive Science*, 13(1), 75–106. <https://doi.org/10.17791/JCS.2012.13.1.75>
- Schlagel, R. H. (1998). Why not artificial consciousness or thought? *Minds and Machines*, 9(1), 3–28. <https://doi.org/10.1023/A:1008374714117>
- Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press.
- Searle, J. (2017). Biological naturalism. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* (2nd ed., pp. 327–336). John Wiley & Sons. <https://doi.org/10.1002/9781119132363.ch23>
- 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. (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64(3), 21–37. <https://doi.org/10.2307/3130074>
- Seth, A. (2009). The strength of weak artificial consciousness. *International Journal of Machine Consciousness*, 1(1), 71–82. <https://doi.org/10.1142/S1793843009000086>
- Seth, A. (2024). Conscious artificial intelligence and biological naturalism. *OSF Preprints*. <https://doi.org/10.31234/osf.io/tz6an>
- Seth, A. K., & Hohwy, J. (2021). Predictive processing as an empirical theory for consciousness science. *Cognitive Neuroscience*, 12(2), 89–90. <https://doi.org/10.1080/17588928.2020.1838467>
- Shagrir, O. (2020). In defense of the semantic view of computation. *Synthese*, 197(9), 4083–4108. <https://doi.org/10.1007/s11229-018-01921-z>

- Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, 36(2), 297–314. <https://doi.org/10.1111/mila.12338>
- Shiller, D. (2024). Functionalism, integrity, and digital consciousness. *Synthese*, 203(2), Article 47. <https://doi.org/10.1007/s11229-023-04473-z>
- Signorelli, C. M. (2018). Can computers become conscious and overcome humans? *Frontiers in Robotics and AI*, 5, Article 121. <https://doi.org/10.3389/frobt.2018.00121>
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>
- Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929(1), 123–146. <https://doi.org/10.1111/j.1749-6632.2001.tb05712.x>
- Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, 10(4–5), 133–172.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1), 1–23.
- Sprevak, M. (2012). Three challenges to Chalmers on computational implementation. *Journal of Cognitive Science*, 13(2), 107–143. <https://doi.org/10.17791/JCS.2012.13.2.107>
- Sprevak, M. (2018). Triviality arguments about computational implementation. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 301–320). Routledge.
- Tagliascio, V. (2007). Artificial consciousness: A technological discipline. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness* (pp. 12–24). Imprint Academic.
- Taylor, J. G. (2007). CODAM: A neural network model of consciousness. *Neural Networks*, 20(9), 983–992. <https://doi.org/10.1016/j.neunet.2007.09.005>
- Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), 20140167. <https://doi.org/10.1098/rstb.2014.0167>
- Torrance, S. (2007). Two conceptions of machine phenomenality. *Journal of Consciousness Studies*, 14(7), 154–166.

- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1), 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Uexküll, J. von. (1940). *A foray into the worlds of animals and humans: With a theory of meaning* (1st University of Minnesota Press ed.). University of Minnesota Press.
- Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, 92(7), 345–381. <https://doi.org/10.2307/2941061>
- Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness*, 2020(1), niaa013. <https://doi.org/10.1093/nc/niaa013>
- Wiese, W. (2024). Artificial consciousness: A perspective from the free energy principle. *Philosophical Studies*, 181(8), 1947–1970. <https://doi.org/10.1007/s11098-024-02182-y>
- Winograd, T. (1990). Thinking machines: Can there be? Are we? In D. Partridge & Y. Wilks (Eds.), *The foundations of artificial intelligence: A sourcebook* (pp. 167–189). Cambridge University Press. <https://doi.org/10.1017/CBO9780511663116.017>
- Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., DiCarlo, J., Ganguli, S., Hawkins, J., Koerding, K., Koulakov, A., LeCun, Y., Lillicrap, T., Marblestone, A., Olshausen, B., ... Tsao, D. (2023). Catalyzing next-generation artificial intelligence through NeuroAI. *arXiv*. <https://doi.org/10.48550/arXiv.2210.08340>
- Ziemke, T. (2001). The construction of ‘reality’ in the robot: Constructivist perspectives on situated artificial intelligence and adaptive robotics. *Foundations of Science*, 6(1), 163–233. <https://doi.org/10.1023/A:1011394317088>
- Ziemke, T. (2007). What’s life got to do with it? In A. Chella & R. Manzotti (Eds.), *Artificial consciousness* (pp. 48–66). Imprint Academic.
- Ziemke, T., & Sharkey, N. E. (2001). A stroll through the worlds of robots and animals: Applying Jakob von Uexküll’s theory of meaning to adaptive robots and artificial life. *Semiotica*, 134(1–4), 701–746. <https://doi.org/10.1515/semi.2001.050>