The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness

Alexander Lerchner[1]
[1]Google DeepMind

Computational functionalism dominates current debates on AI consciousness. This is the hypothesis that subjective experience emerges entirely from abstract causal topology, regardless of the underlying physical substrate. We argue this view fundamentally mischaracterizes how physics relates to information. We call this mistake the *Abstraction Fallacy*. Tracing the causal origins of abstraction reveals that symbolic computation is not an intrinsic physical process. Instead, it is a mapmaker-dependent description. It requires an active, experiencing cognitive agent to alphabetize continuous physics into a finite set of meaningful states. Consequently, we do not need a complete, finalized theory of consciousness to assess AI sentience—a demand that simply pushes the question beyond near-term resolution and deepens the AI welfare trap. What we actually need is a rigorous ontology of computation. The framework proposed here explicitly separates simulation (behavioral mimicry driven by vehicle causality) from instantiation (intrinsic physical constitution driven by content causality). Establishing this ontological boundary shows why algorithmic symbol manipulation is structurally incapable of instantiating experience. Crucially, this argument does not rely on biological exclusivity. If an artificial system were ever conscious, it would be because of its specific physical constitution, never its syntactic architecture. Ultimately, this framework offers a physically grounded refutation of computational functionalism to resolve the current uncertainty surrounding AI consciousness.

1. Introduction

Large Language Models have been empirically successful enough to push the 'Hard Problem' of consciousness out of pure theory and into the realm of engineering and policy. With the massive returns we see from scaling compute (Bubeck, 2023; Hoffmann, 2022; Kaplan, 2020; Sutton, 2019), the prevailing functionalist paradigm assumes that hitting the right information-processing roles is enough to achieve phenomenal consciousness (Chalmers, 1996; Dehaene et al., 2017; Dennett, 1991). Under this view, algorithmic indicator properties act as likely evidence for sentience (Butlin et al., 2023). This assumption is exactly what motivates recent, serious proposals for AI welfare and moral patienthood (Long et al., 2024). This shift is reinforced by leading theorists who assign significant credence to the possibility that state-of-the-art models could possess genuine experience within the next decade (Chalmers, 2023; Schneider, 2019).

At the center of these proposals lies substrate independence, the idea that the "software" of the mind could run on silicon just as well as on carbon. That assumption has begun to face sustained criticism from a 'Biological Turn'. Seth (2025) and Block (2025), for example, argue that consciousness may depend on life-maintaining biological processes, such that experience requires the organized dynamics of living systems. In contrast to substrate independence, this view makes biology central rather than incidental. Yet that position remains empirical, as it does not clearly identify the basic logical mistake at the core of computational functionalism.

Here, we derive the logical sequence that vindicates the intuition that computation is not sufficient to instantiate consciousness. The difficulty with computational functionalism is not just that it may overlook biological details. The problem runs much deeper. It is rooted in a misunderstanding of how physics relates to information and computation.

Corresponding author(s): lerchner@google.com

Modern physical sciences have deliberately excised subjective experience in order to ensure operational objectivity (Frank et al., 2025). This strategy has been extraordinarily successful. But when this stance is applied to the question of how computation relates to subjective experience, it is bound to fail. Applying this operational objectivity to the very definition of computation is highly problematic, as can be seen in the ongoing and still unresolved debates around the role of an 'observer' in supplying meaning to computational symbols.

Moreover, it turns out that the term 'observer' suggests a too passive role for the missing prerequisite to fully define computation in physical terms. Our framework elucidates why computation is not an intrinsic process that simply unfolds in matter. Instead, it is a way of *describing* physical processes. To count as computation, continuous physical dynamics must be partitioned into a finite set of discrete, semantically meaningful states (i.e., a form of alphabet). Such semantic partitioning logically requires an active, experiencing cognitive agent, which we define as a *mapmaker*, to contrast it with the passive connotation of a standard 'observer'. It is the mapmaker who performs this alphabetization. Without such an active agent interpreting the computation, there are only continuous physical events, not symbols.

A key insight from our contribution is that resolving the present uncertainty surrounding artificial consciousness does not require a complete and final theory of consciousness. Instead, we need an ontology of computation. Via this route, we can logically prove that algorithmic symbol manipulation, no matter how large in scale or intricate in architecture, cannot constitute the physical instantiation of experience, since it is a mapmaker-dependent descriptive tool.

Demonstrating the role of the mapmaker in the causal story changes the focus of the debate. So far, well-known critiques of artificial consciousness, including Searle's Chinese Room and related arguments (Block, 1978; Putnam, 1988; Searle, 1980), rely primarily on *reductio ad absurdum*. They aim to show that pure syntactic manipulation, even if it perfectly mirrors outward behavior, still seems to miss something essential.

Our approach takes a different route. Instead of appealing to intuitions about what is absent, we examine how abstraction arises in the first place. If computation depends on a mapmaker who extracts invariants from experience and assigns symbols, then the dependency is built into the structure. Any computational map presupposes an experiencing agent who performs the alphabetization. Making the algorithm more complex does not undo this order of dependence. No increase in scale allows the map to generate the subject whose activity is required for computation to count as such at all.

In other words, the claim that algorithmic complexity generates consciousness commits an ontological inversion: it mistakes the syntax for the territory of intrinsic dynamics, and assumes that the mapmaker can be created from the map. By delineating the structural dissociation between extrinsic behavioral simulation and intrinsic physical instantiation, we demonstrate that digital architectures are precluded from becoming moral patients. This realization pulls the field of AI safety out of the welfare trap. It allows us to focus entirely on the concrete risks of anthropomorphism, treating AGI as a powerful, but inherently non-sentient tool.

2. The Ontology of Abstraction: Map vs. Territory

Computer science frequently treats the abstractions underlying algorithms as mathematical givens, leaving the question of their physical realization open. What exactly is the physicalist ontology of an abstraction? To answer this question, we need to establish how exactly abstract syntax relates to physical dynamics.

2.1. The Standard Definition of Physical Implementation

In the standard literature of implementation (Chalmers, 1996; Putnam, 1988), a physical system P is said to implement an abstract computation C via a mapping function f. The requirement is straightforward: f has to map physical states to abstract states in a way that allows the underlying physical causality to mirror the algorithm's logical structure.

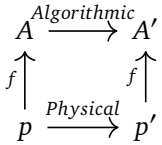

Figure 1 | **The Commutative Diagram of Implementation.** In the standard view, the mapping f (the alphabetization) interprets the physical evolution of the vehicle ($p \rightarrow p'$) as the logical evolution of the abstract content ($A \rightarrow A'$).

Suppose the physical system evolves from state p to p' governed entirely by physical laws ($p \rightarrow p'$). At the same time, the abstract computation advances from logical state A to A' as dictated by its algorithmic rules ($A \rightarrow A'$).

The system successfully implements the computation if:

$$f(p) = A \quad \text{and} \quad f(p') = A' \tag{1}$$

For the diagram to commute (see Figure 1), applying the mapping f to the resulting physical state (p') has to yield the exact target abstract state (A') dictated by the logical rules.

2.2. The Physical Origin of Abstract States (A)

What exactly are these abstract states A? If we want to understand the mapping f, we have to determine the ontology of A. Functionalist accounts usually treat A, such as the logical state "Pain" or "Red", as a floating abstraction with no specific physical realization. This bypasses the causal history required to generate an abstraction.

Forming an abstraction is not free. It is an active, metabolically expensive physical process of extracting invariants. Before a cognitive agent can form a concept A (like "Red"), it has to encounter the territory: multiple experiential instances of redness. From there, the agent actively filters out high-dimensional noise to isolate a stable core. In the vocabulary of manifold learning, the agent projects the high-dimensional manifold of raw experience down onto a lower-dimensional invariant subspace. This subspace physically constitutes the concept A.

It is tempting to argue that unsupervised clustering algorithms already generate abstractions without needing any prior experience. But this confuses *statistical compression* with *phenomenal constitution*. Certainly, an unsupervised algorithm can cluster data points to locate a statistical centroid. That mathematical invariant, however, is merely a compressed address within a latent space.

For that statistical centroid to count as a genuine *concept*—a semantically grounded category like "Redness"—the agent needs the intrinsic phenomenal state to serve as the common denominator for the grouped instances. If there is no constitutive experience of Red anchoring the reference, the cluster is just a high-density region in a vector space. It does not qualify as a concept held by a subject.

Concepts (A), therefore, are not Platonic ideals waiting to be discovered. They are constituted

neurophysiological states[1] existing solely within the cognitive system of the agent that performed the abstraction. They are the "internal map" derived from the "territory" of experience. Once formed, these constituted 'common cores' serve as the stable building blocks for compositional imagination. Because the concept 'Red' and the concept 'Whale' are intrinsic control states derived from lived experience, the brain can recombine them to instantiate the experience of a 'Flying Red Whale'—a composite that has never been encountered but is physically coherent. Thus, thinking is not based on algorithmic processing of empty symbols, but a combinatorial generation of constituted invariants. AI simulates the rules of this recombination flawlessly, but it structurally lacks the intrinsic building blocks required to run the experiential imagination.

2.3. The Indispensable Mapmaker in the Mapping Function (f)

Historically, the physical sciences, and engineering in particular, progressed by systematically removing subjective experience from their explanations of natural phenomena (Frank et al., 2025). However, if one imposes this operational objectivity onto the ontology of computation, it creates an epistemic blind spot. It forces computational functionalism into an impossible conundrum: attempting to reconstruct subjective experience from a starting point defined entirely in objective, non-experiential terms.

After having established the physical grounding of abstract states (A), which reside by logical necessity within an actively experiencing cognitive agent, we can now expose this blind spot in the standard definition of implementation: The mapping function f, which links the machine's physics (p) to that abstraction (A), cannot reside within the machine itself.

In the philosophical literature concerning semantics and the map-territory relation, this external anchor is traditionally referred to as an 'observer'. However, the term 'observer' implies a passive reception of information–an entity merely looking at a pre-existing territory or map. We deliberately introduce the term *mapmaker* to explicitly correct this passive implication. The mapmaker is the active, metabolically vulnerable cognitive agent that must exist as a prerequisite to generate computation. It performs two active, constitutive roles: first, extracting invariants from continuous physical experience to construct the internal map (concepts); and second, executing the arbitrary assignment of physical tokens to construct the external computational map (symbols). Applying this insight fundamentally resolves the ontological status of the computational terms:

1. **The Physical States (p):** These are the symbols (the *vehicle*). They are objective physical entities (e.g., voltage gradients), possessing zero intrinsic semantic content.
2. **The Abstract States (A):** These are concepts (the *content*). As established, these are grounded physiological states existing exclusively within the mapmaker who holds the computation's meaning.
3. **The Mapping Function (f):** This is the *alphabetization*. It represents the assigned association held in the mapmaker's mind, actively bridging the machine's blind physics (p) to the mapmaker's grounded concepts (A).

The standard definition ($p \to p' \cong A \to A'$) therefore describes a hybrid relationship: a physical object (p) linked to a mental concept (A) via the necessary mediation of the mapmaker (f).

It is important to note that identifying this indispensable mapmaker does not resurrect a dualistic "homunculus" or a localized "decoder" sitting inside the brain. As argued by Buzsaki

[1]We use the term "constitutive" to denote a relationship of strict ontological composition, distinct from mere causal triggering of functional equivalence. A constituted mental state is one whose semantic reality is physically made of, and fundamentally un-abstractable from, the specific thermodynamic and metabolic dynamics of the experiencing organism.

(2019) and Maturana & Varela (1980), the mapmaker is the entire structurally unified organism subject to the laws of thermodynamics. The organism does not algorithmically "choose" to make a semantic cut. Instead, the continuous environment is filtered into discrete states directly through the organism's metabolic constraints. There is no ghost reading the alphabet; the living experiencing subject enacts it.

Treating the logical progression ($A \to A'$) as an intrinsic property of the physical evolution ($p \to p'$) is the main functionalist error. Such a perspective conflates the cognitive interpretation of the mapmaker with the actual physical reality of the machine, ignoring the experiencing subject needed to ground the computation in the first place.

2.4. Alphabetization: The Semantic Imposition Beyond Discretization

The mapping function f serves as the actual locus of alphabetization. While often dismissed in the literature as just "reading" the system, alphabetizing is actually a metabolically demanding cognitive act. It imposes a discrete ontology onto continuous physics and is subject to the thermodynamic bounds of information processing (Attwell & Laughlin, 2001; Bennett, 1982; Landauer, 1961; Laughlin et al., 1998). Here it is important to separate two processes that are routinely conflated:

- **Discretization (Thermodynamic):** A system physically settling into stable attractors, such as a transistor holding at 5V. This is a property of the vehicle (p) and functions only to suppress physical noise.
- **Alphabetization (Semantic):** Explicitly assigning those stable states to a *predefined finite set of symbols* (like $\{0, 1\}$ or $\{A, B, C\}$). This operation belongs exclusively to the mapmaker (f).

Because physical reality is inherently continuous, thermodynamics can only yield stable macroscopic states—it can never provide a predefined finite alphabet. Constructing a computational system therefore requires intervention from a mapmaker. This external agent must enforce a semantic identity by treating vastly heterogeneous micro-states as one fungible symbol (e.g., "1").

Consequently, relying on a mapmaker to build the system introduces a fundamental causal disconnect. In the physical territory of the machine, the transition from 2.0V to 2.1V is a genuine causal event driven by electrodynamics. Yet within the computational map, this transition is functionally invisible: the mapmaker has "alphabetized" it into identity. Therefore, the causal dynamics of the computation do not supervene on the physics of the substrate. They supervene entirely on the rules of the mapmaker.

Claiming these symbols exist independently of the observer exemplifies *The Blind Spot* (Frank et al., 2025). It is a textbook case of "surreptitious substitution" (Husserl , 1970): taking the cognitive output of the scientist (the finite alphabet), projecting it backward into the physical system, and declaring it was there all along. Information is not a fundamental building block of the universe; it is a derivative property that absolutely presupposes a cognitive agent to define the finite set.

2.5. Simulation vs. Instantiation

After having established the distinct nature and differing roles of concepts versus symbols, we can apply our definitions to clarify why simulating a process is fundamentally different from instantiating it:

- **Simulation:** The syntactic manipulation of physical vehicles (p) to track the abstract

relationship between concepts (A).
- **Instantiation:** The replication of the intrinsic, constitutive dynamics (P) of the process itself.

Standard functionalism assumes that preserving the abstract topology of the map ($A \rightarrow A'$) is sufficient to generate the phenomenon of the territory (P), thereby overlooking the specific causal powers and constitutive mechanisms of the physical substrate (Craver, 2007).

Consider the biological heart. We often describe it as a pump that pushes blood through the body. We design and build mechanical hearts to pump blood in the same way, and so we say it is "functionally equivalent". But the real heart does more than just pump. It also releases hormones (such as ANP), helps control the body's metabolism, and communicates with the nervous system through feedback signals. Patients with mechanical hearts often suffer subtle, systemic physiological deficits precisely because the device instantiates only the coarse-grained map of the selected function. It fails to instantiate the full biological territory of the organ.

Note that this granularity mismatch applies *a fortiori* to the neuron. Functionalism tends to treat the neuron as just a receiver and sender of electrical signals, even though it is a living, metabolic entity deeply integrated into a chemical and hormonal network of the body. This abstraction undermines the "Fading Qualia" thought experiment (Chalmers, 1996), a cornerstone of functionalist intuition. Chalmers argues that if biological neurons were gradually replaced, one by one, with silicon chips that preserve the same input-output relations, it would be implausible for consciousness to progressively fade while behavior and functional organization remain unchanged. From this premise, he concludes that the preservation of abstract functional organization is sufficient for the preservation of conscious experience. However, a silicon replacement that perfectly mimics only the electrical firing profile ($p \rightarrow p'$) preserves nothing but an extrinsic computational map, one defined entirely by an external mapmaker's chosen abstractions ($A \rightarrow A'$). It systematically obliterates the intrinsic thermodynamic territory (P) required for life, substituting a constitutive physical reality with a causally inert, syntactic simulation. The qualia do not mysteriously "fade"; the foundational metabolic substrate required to instantiate them is simply removed.

The limits of physical simulation elsewhere in biology make the point explicit. A GPU that simulates photosynthesis may accurately model the abstract transformation from sunlight, water, and carbon dioxide (A) to oxygen and glucose (A'), but it will not synthesize a single molecule of glucose or release oxygen. So while perfectly simulating the process, it lacks the causal capacity to perform the underlying biochemical work. To suggest that simulating the "software" of the brain avoids this physical constraint introduces a category error (Searle, 1980). It conflates the algorithmic description of a process with the intrinsic physics required to instantiate it.

This requirement for intrinsic causality follows directly from physicalist principles and is not a metaphysical preference. Illusionist accounts argue that a functional report fully captures the reality of an experience (Dennett, 1991). However, if we apply Jaegwon Kim's principle of causal closure (Kim, 2005), the act of reporting an experience, such as physically displacing air to say "I am in pain", is an indisputably physical event. For the subjective experience to truly cause this report, rather than being a coincidence or an illusion, the experience itself must have grounded physical power and be capable of performing work.

In a digital simulation, the causal chain is driven entirely by the vehicle (p). The logic gate does not switch because it 'hurts' (content causality driven by A). Instead, it switches because the voltage crosses a defined physical threshold (vehicle causality driven by p). The physical state of the system alone determines its evolution. The semantic content of the symbol (A) plays no causal role, since the machine would perform the same physical operations even if the symbol referred to nothing at all. Assuming otherwise would mean to fall victim to the abstraction fallacy.

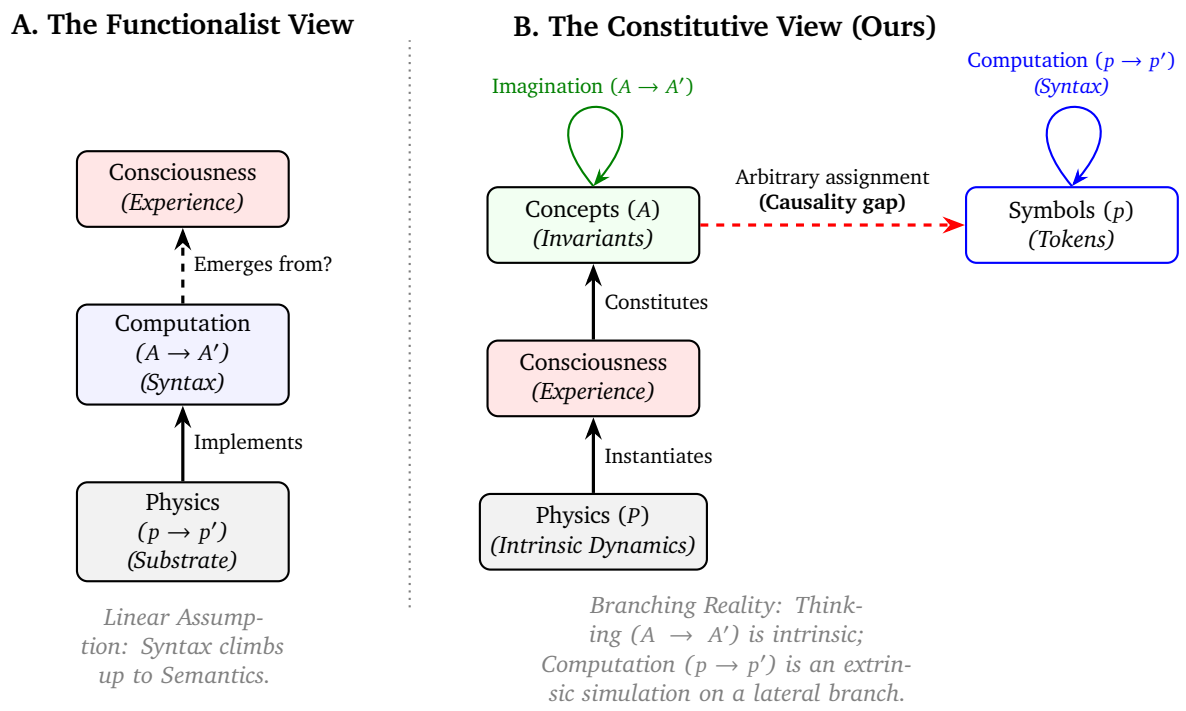

Figure 2 | **The Causal Topology of Abstraction.** (**A**) Functionalism assumes a linear hierarchy where computation is the bridge between physics and consciousness. (**B**) Our framework reveals a branching topology. The vertical axis represents the *intrinsic chain*: Physics (P) instantiates experience, which constitutes concepts (A) via abstraction. Thinking/imagination ($A \to A'$) occurs here. The lateral axis represents the *extrinsic chain*: symbols (p) are created by *arbitrarily associating* a physical token with a concept (the red dashed arrow). This unbridgeable lateral step constitutes the **causality gap**. Computation ($p \to p'$) is a process that operates solely on this lateral branch. This lateral move from concept to symbol–an arbitrary assignment rather than a vertical abstraction–severs any intrinsic causal path from the symbol back to the originating experience.

2.6. The Fallacy of Computational Emergence

When confronted with the distinction between simulation and instantiation, functionalists tend to retreat to complexity theory and emergence. They argue that just as "wetness" emerges from the interaction of water molecules, consciousness will emerge from computation once the system crosses a threshold of sufficient complexity. This objection fails because it conflates weak physical emergence with what we term the *fallacy of computational emergence*.

- **Weak Emergence (Physical):** Macroscopic properties (like wetness) supervene directly on the intrinsic causal dynamics of the microscopic physical substrate (e.g., H_2O).
- **Computational Emergence (Abstract):** The claim that an abstract description of a process (the map) can, solely through a massive increase in syntactical complexity, transmute into the physical process itself (the territory).

Functionalists insist that consciousness is a unique case because it is purely "substrate independent information". But this argument assumes its own conclusion: it presumes the mental state *is* the abstract information (A), completely sidestepping the physical reality (P) that generates it. As we have established, syntax ($A \to A'$) possesses no intrinsic causal power; it is a mapmaker's attribution. To claim that an abstract syntax somehow "emerges" to become a physical cause falls outside scientific hypothesis entirely, as it requires violating the causal closure of the physical

world.

3. The Causal Circularity: Correcting the Chain

By establishing this firm boundary between physical dynamics (P) and computational maps (A), we can locate the exact logical collapse within computational functionalism (illustrated in Figure 2).

3.1. The Ontological Inversion and the Causality Gap

Traditional functionalist accounts usually lean on a straightforward, unexamined causal sequence:

$$\text{Physics} \rightarrow \text{Computation} \rightarrow \text{Consciousness}$$

This assumes consciousness will simply appear as a downstream byproduct once computational complexity reaches a certain threshold. Yet, as Section 2 established, computation is hardly a natural kind waiting out in the world to be discovered. Defining discrete symbols and giving them semantic meaning requires an already-conscious agent acting as a mapmaker (f).Consequently, we have to fundamentally reorder the causal sequence:

$$\text{Physics} \rightarrow \text{Consciousness} \rightarrow \text{Concepts} \rightarrow \text{Computation}$$

1. **Physics:** The universe's intrinsic causal dynamics.
2. **Consciousness:** Phenomenal experience arising directly out of specific thermodynamic organizations within that physics.
3. **Concepts:** The *internal* map, formed by extracting invariants from raw experience.
4. **Computation:** The *external* map, consisting of the syntactic manipulation of discrete symbols arbitrarily assigned to those concepts.

This revised chain operates strictly unidirectionally. While concepts stay physically anchored in the subject's intrinsic experience, the irreducible 'what it is like' to be that entity (Nagel, 1974), computational symbols are just physical tokens with no inherent link to the concepts they represent. Moving from concepts to symbols is not a step in abstraction. It is a lateral act of assignment where a mapmaker forcibly binds a physical token to a mental concept. It is precisely this unbridgeable lateral step that exposes the causality gap, permanently cutting off any intrinsic path leading back from the symbol to the original experience.

Once this link is established, the mapmaker constructs syntactic rules to govern the physical state transitions of the symbols ($p \rightarrow p'$). These rules are explicitly designed, from the top down, to perfectly track and mimic the intrinsic, associative evolution of the corresponding concepts ($A \rightarrow A'$). Yet despite this flawless structural mimicry, the physical tokens themselves exert no causal influence on the semantic content. The machine blindly executes the mapped trajectory, completely decoupled from the phenomenal reality it simulates.

Functionalism attempts to explain the origin of the mapmaker (Step 2) by appealing to a process (Step 4) that already presupposes the mapmaker's existence. This is not simply an empirical gap but a category error: it serves as a robust physicalist constraint. The construction of a syntactic map requires a mapmaker from the outset. Therefore, no amount of algorithmic complexity can traverse the causality gap backward to produce an experiencing subject. This ontological inversion inherent in computational functionalism generates a structural paradox:

it attempts to derive the foundational mapmaker solely from the mapmaker's own derivative outputs.

3.2. The Universality of Alphabetization

A long-standing debate in AI, tracing back to the connectionist shift of the 1980s (McClelland et al., 1987), argues that modern neural networks differ from earlier symbolic systems because they operate at a so-called sub-symbolic level. According to leading researchers, this architecture makes it possible to build 'world models' (LeCun, 2022) or recursive epistemic loops (Laukkonen et al., 2025) that amount to genuine understanding.

We agree that these recursive architectures can reproduce the structural features of introspection. In the same way that high-dimensional vector spaces capture geometric relationships in a form that differs from discrete logical symbols, neural networks can model complex relational structure. However, interpreting this structural or geometric accuracy as evidence of intrinsic meaning repeats the abstraction fallacy. It confuses the structure of the representation with the underlying physical reality, and treats the geometry of the model as if it were the physics of the system itself.

To formalize this objection, we introduce a strict Shannon constraint: to process information in the strict sense, a system requires a finite classical alphabet of discrete symbols[2] and a probability distribution over those states. At the macroscopic level of biological life and artificial hardware, the physical world of light intensity, chemical concentration, and membrane voltage does not come pre-labeled with discrete 0s and 1s. The universe does not pre-package its macroscopic physical states into an operative computational alphabet; a mapmaker must explicitly enforce it. Treating a neural spike or voltage toggle as a "symbol" requires more than just physical discretization; it requires alphabetization. A mapmaker must actively enforce a semantic identity upon the system, treating a heterogeneous range of continuous physical states as a single, fungible token. The same constraint applies equally to the high-dimensional vector spaces of deep learning. Despite being frequently described as 'continuous' representations, vectors are implemented as sequences of floating-point numbers, where each float is a discrete symbol from a finite alphabet (e.g., IEEE 754).

The alphabetization requirement inherent in the mapping function (f) applies to all forms of computation, whether digital, analog, or quantum. Consider an analog clock. Physically, the device is a collection of gears and springs governed by continuous dynamics (P). It only "computes" time because a mapmaker intervenes, mapping a specific set of continuous angles to a semantic concept (e.g., "3:00 PM"). Without this semantic imposition, the clock is just metal moving in accordance with Hamilton's equations; it contains no intrinsic "time." Thus, the physical substrate does not "process information" absent a prerequisite alphabet of intrinsic symbols; rather, it generates continuous dynamics that an external mapmaker interprets as information.

Even if future AI systems abandon floating-point operations for fully analog neuromorphic chips, the ontological gap would still remain. As soon as a physical state, whether it is a discrete voltage level or a continuous charge pattern, is identified as a "readout" or "hidden state," it has already undergone alphabetization by a mapmaker. Consequently, these models remain sealed behind a semantic barrier. While they can construct sophisticated internal maps, they lack the intrinsic, constitutive connection to the physical territory of experience.

[2]We acknowledge that fundamental physics is arguably quantized at the Planck scale, as posited by information-theoretic interpretations of quantum mechanics. However, the macroscopic thermodynamic and metabolic gradients that constitute biological life operate continuously. Furthermore, even if the universe is fundamentally informational in the quantum sense (Tegmark, 2008; Wheeler, 1990), translating intrinsic quantum states into extrinsic, classical symbols manipulated by Turing machines still requires a mapmaker-dependent discretization process.

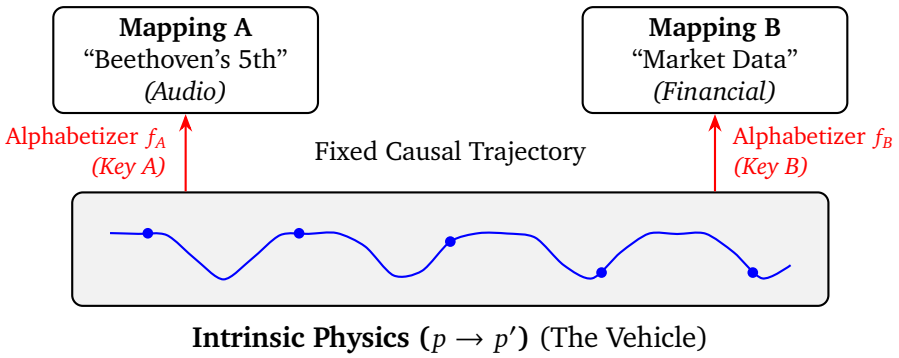

Figure 3 | **The Indeterminacy of Mechanism (The Melody Paradox).** A single physical vehicle (bottom) possesses a fixed causal trajectory. However, it does not instantiate a unique computation. Depending on the alphabetization key applied (f_A or f_B), the same physical states can be mapped to entirely different abstract computations (Top Left vs. Top Right). Therefore, computation cannot be intrinsic to the physics (p).

3.3. The Indeterminacy of Mechanism

The mechanistic view of computation, championed by Piccinini (2008), attempts to excise the mapmaker completely, arguing that computation can be defined without any appeal to representation. The idea is that computation can be defined solely by the manipulation of "digits"—macroscopic physical states distinguishable solely by the system's functional organization. However, while this maneuver hides the mapmaker, it does not eliminate the need for one. As Sprevak (2018) notes in his analysis of triviality arguments, fixing the computational identity of any physical mechanism still requires an external specification of the relevant states. Physical mechanisms can certainly possess stable attractors (which we clarified as thermodynamic discretization in Section 2.4), but grouping these continuous attractors into a specific, finite computational alphabet remains a strictly extrinsic imposition by a mapmaker.

We can consider a simple melody paradox (Figure 3) to expose this sleight of hand. Imagine a physical device stepping through a sequence of stable voltage states. The physical transitions ($p \to p'$) are fixed by the laws of electrodynamics. Yet, the *computational identity* ($A \to A'$) of this exact process remains entirely under-determined. Without an external mapmaker to supply the mapping key, that single sequence of physical states could represent:

1. A melody played forward (Mapping A).
2. The exact same melody played backward (Mapping B, e.g., a retrograde inversion).
3. A rapid stream of stock market prices (Mapping C).
4. Coherent noise (if the symbol set is defined with different granularity).

There is no property inherent to the physical voltage (p) that privileges one of these finite symbol sets over the others. The "digit" is not a natural kind waiting to be discovered in the mechanism. It is an epistemic cut made by the mapmaker, forcing continuous physical dynamics into a finite logical set (Putnam, 1988).

Thus, even if a physical system evolves through clearly distinguishable macroscopic states according to reproducible rules, there still needs to be a mapmaker involved to collapse the indeterminacy of computational interpretations into a single, unique trajectory. The mechanism provides the ink. The mapmaker must provide the alphabet.

4. Implications: The Limits of Computational Implementation

Our framework reveals that the barrier to AI consciousness is not a matter of computational scale or increasing algorithmic complexity. Instead, it is a matter of simulation vs. instantiation. Because of this, it bears directly on two areas that currently receive rapidly increasing attention and investment in the field: embodied robotics and AGI safety (Bengio et al., 2024; Bostrom, 2014).

4.1. The Transduction Fallacy in Robotics

One of the strongest objections to our constitutive framework comes from the idea of embodiment. According to this argument, the key missing ingredient for AI systems to become conscious is proper causal integration in the physical environment. Providing sensors and actuators that allow the system to perceive and act in real time, so the argument goes, can close the causality gap and allow the system's internal symbols to become grounded.

However, adding sensors and actuators does not by itself solve the deeper issue of instantiating experience. We agree that an embodied agent does solve the referential aspect of the symbol grounding problem (Harnad, 1990). It enables successful mapping of internal symbols to an external physical data stream, which avoids the infinite regress of a purely lexical internal dictionary. But we need to carefully distinguish such referential mapping from intrinsic sense-making.

An analogy helps clarify the point. Connecting a computer to cameras and robotic arms is similar to attaching measuring instruments to a simulation. The simulation now receives real-world data, but the internal variables of the model are still symbolic representations rather than the physical processes themselves. In the same way, a weather model connected to live atmospheric sensors does not become the atmosphere. It simply receives and manipulates data about it.

The same principle applies to embodied AI systems. Sensors and actuators allow the system to interact with the physical world, but they do not automatically transform symbolic representations into intrinsic, experienced semantics. The system may build increasingly detailed maps of its environment, but interacting with the territory does not by itself turn the map into the territory of experience.

Tracing the causal topology of the embodied system exposes what we term the *transduction fallacy*:

1. **Input Transduction:** Sensors transduce external physical forces into continuous voltages, which an ADC, calibrated by an external mapmaker, subsequently alphabetizes into internal digital states (e.g., thermal energy → continuous voltage → discrete integer).
2. **Syntactic Policy:** The syntactic engine manipulates these internal discrete states to generate output states, physically implementing the abstract algorithm.
3. **Output Transduction:** Actuators convert the digital output back into macroscopic physical forces.

Importantly, the operational core of the robotic system, its algorithmic controller, functions entirely within the second step. It operates only on symbols (e.g., floating point numbers manipulated by matrix multiplications) that have been discretized and alphabetized for computation by an external mapmaker.

Proponents of modern "end-to-end" continuous control might object that contemporary robotic architectures map raw sensor arrays directly to actuator torques via deep neural networks, without

requiring human-readable symbolic representations. However, as explained in Section 3.2, the hardware that executes these control policies, such as GPUs, still relies on symbols via the alphabetization of floating point numbers, and the pre-defined mathematical rules for manipulating them. The prerequisite alphabetization by a mapmaker is not absent. On the contrary, it is baked into the silicon architecture itself.

The transduction fallacy is not just that physical force is translated into numbers at the sensor. The deeper, categorical error occurs if one assumes that the algorithmic manipulation of these transduced symbols can somehow instantiate the phenomenal subject. To fully understand the difference between the embodied robot running an algorithm on a chip and the biological mapmaker, we need to remember that for the latter, subjective experience is a given, not because of abstract information processing, but because of a specific, metabolically constituted physical reality.

There is no physical or logical justification to assume that a silicon chip generates a similar physically constituted experience simply because it is executing a syntactic mapping between sensory inputs and mechanical actuators. If we assume otherwise, we arrive at a logical entailment that violates physicalist principles. As we demonstrated in Section 2.5, the abstract states associated with any algorithm (i.e., the "content" of what the computation refers to) have no intrinsic causal power. The only physical causality in the system belongs to the silicon vehicle itself. Therefore, arguing that the execution of a syntactic mapping (whether of sensor data, actuator data, or otherwise) in the embodied robot generates experience, means arguing that the physical chip must inherently possess the capacity for consciousness solely due to its material properties. Note that this follows irrespective of whether it is connected to a robot body, or what specific algorithm it runs. Thus, a rigorous analysis of the map-territory relation reveals that–unlike what may seem initially plausible–embodiment cannot transform a simulation into constitutive subjective experience.

4.2. Ontological Relief: The Safety of the Non-Sentient Tool

Having established that neither algorithmic complexity nor physical embodiment can cross the causality gap, we can now address the practical implications of this framework. The structural dissociation between computational mapping and physical territory has immediate implications for AI safety. It helps clarify what kinds of systems might actually support phenomenal experience and which ones do not.

Work in enactivism and embodied cognition has pointed to several physical processes that appear closely tied to conscious experience. These include autopoiesis and ongoing thermodynamic regulation within a living system (Damasio, 1999; Friston, 2010; Thompson, 2007). Historically, these mechanisms have been treated as properties of biological organisms, and therefore as features tied specifically to carbon-based life.

Our framework suggests a slightly different interpretation. It keeps the same physicalist emphasis on real, intrinsic physical processes, but it does not require that those processes occur only in biological organisms. Phenomenal experience, in this view, depends on the actual physical instantiation of certain kinds of dynamics. Consequently, the framework does not imply that consciousness must be limited to biological life. In principle, a non-biological system could be designed to realize the necessary physical conditions. If those conditions were successfully instantiated in a synthetic substrate, then conscious experience might also arise there.

However, this fundamental structural limitation guarantees that if such an artificial system were conscious, it would be entirely due to its specific physical constitution–the exact inverse of substrate independence. Given that phenomenal consciousness is a constituted physical state, and given the ontological boundary between simulation and instantiation, it follows that subjective

experience will not suddenly emerge by scaling compute or running certain kinds of powerful algorithms. It is not a software artifact that can be accidentally or deliberately created.

This realization helps clarify the field's immediate direction: the development of highly capable Artificial General Intelligence (AGI) does not inherently lead to the creation of a novel moral patient, but rather to the refinement of a highly sophisticated, non-sentient tool.

However, achieving behavioral mimicry at this scale creates a new need for epistemic hygiene. AI systems are becoming rapidly better at reproducing the behavioral signals humans associate with other conscious minds, a trend that will be increasingly amplified with embodied systems such as humanoid robotics. This presents a clear challenge for the scientific community. Instead of preparing for the rights of the machine, what we need is a clear defense of the methodological boundary between simulated agency (teleonomy) and the physical instantiation of a subject (teleology) (Cao, 2012). Therefore, any future claim of artificial sentience must be subjected to rigorous physicalist verification, not based on algorithmic complexity, but on the specific, intrinsic physical dynamics required for experience.

5. Conclusion: The Blind Spot of Computation

Computation is routinely viewed as a basic feature of the universe, and computational functionalism builds on this view by assuming that computation is at the root of our conscious experience. However, by carefully examining the causal origins of computation, we have shown that this view commits an ontological inversion: conscious experience cannot be the downstream result of computation because it is the necessary physical prerequisite for it.

Furthermore, we show that computation is fundamentally a description, a map, that cannot physically instantiate what it describes. These insights, which challenge widespread intuitions about both the nature of subjective experience on the one hand, and the nature of computation on the other, are based solely on well-established physical laws and carefully applied logic. Crucially, in contrast to most discussions and speculations around the potential of conscious AI, this framework does not depend on a complete theory of consciousness. It resolves the apparent conundrum by addressing the other side of the equation: what, exactly, is computation in ontological terms? Regarding the question of consciousness, the framework only requires that phenomenal experience does not violate causal closure, one of the most fundamental principles of our scientific understanding. This principle alone is sufficient to show that experience must be a physically constituted, wholly physical phenomenon, which allows us to sidestep speculations around any form of dualism or epiphenomenalism.

To summarize the ontology we have established, computation is the syntactic manipulation of discrete symbols governed by rules designed to simulate conceptual thought. These symbols are not distilled essences of concepts; they are arbitrary physical tokens assigned by a mapmaker. Concepts, in turn, are physically constituted invariants that are actively extracted from lived, thermodynamic experience. Therefore, expecting an algorithmic description to instantiate the quality it maps is like expecting the mathematical formula of gravity to physically exert weight. Believing AI can become conscious solely through the manipulation of internal variables is to commit the error of the "blind spot" (Frank et al., 2025): mistaking the map for the territory.

The inability of computational descriptions to generate subjective experience is therefore not a failure of engineering, but a logical necessity of description itself. This also implies that qualia are not puzzles that can be solved by increasingly elegant syntax. Instead, they represent the intrinsic, underlying substrate that makes the semantic assignment of syntax possible in the first place.

By creating increasingly powerful artificial intelligence we are not engineering a new form of

life, but instead constructing increasingly accurate predictive maps. Yet, regardless of its predictive fidelity, its utility as a reasoning tool, or its physical embodiment, the artificial system remains categorically distinct from the territory of phenomenal experience. Recognizing this distinction, and avoiding the ontological inversion of the abstraction fallacy, is the prerequisite for a mature, physically grounded science of machine intelligence.

Acknowledgements

I am grateful to Shamil Chandaria for his review and encouragement, and to Sébastien Krier for his early advocacy and public policy perspective. I also thank Mandana Ahmadi for her valuable feedback in adapting this manuscript for a wider scientific readership, and my colleagues at Google DeepMind for the rigorous debates that helped sharpen the presentation of this framework.

Disclaimer

The theoretical framework and proofs detailed herein represent the author's own research and conclusions. They do not necessarily reflect the official stance, views, or strategic policies of his employer.

References

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, 21(10), 1133–1145.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science* (New York, N.Y.), 384(6698), 842–845.

Bennett, C. H. (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12), 905–940.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9, 261-325.

Block, N. (2025). Can only meat machines be conscious? *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.08.009

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

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 [cs.CL]*. arXiv. http://arxiv.org/abs/2303.12712

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.

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 preprint arXiv:2308.08708*.

Cao, R. (2012). A teleosemantic approach to information in the brain. *Biology & Philosophy*, 27, 49–71.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.

Chalmers, D. J. (2023). Could a Large Language Model be Conscious? *arXiv preprint arXiv:2303.07103*.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Damasio, A. (1999). The feeling of what happens: Body and emotion in the making of consciousness. Harcourt Brace.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486-492.

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Co.

Frank, A., Gleiser, M., & Thompson, E. (2025). *The Blind Spot: Why Science Cannot Ignore Human Experience*. MIT Press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1), 335–346.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). Training Compute-Optimal Large Language Models. *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2203.15556

Husserl, E. (1970). The crisis of European sciences and transcendental phenomenology. Evanston: Northwestern University Press.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv [cs.LG]*. arXiv. https://doi.org/10.48550/arXiv.2001.08361

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183–191.

Laughlin, S. B., de Ruyter van Steveninck, R. R., & Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience*, 1(1), 36–41.

Laukkonen, R. E., Friston, K., & Chandaria, S. (2025). A Beautiful Loop: An Active Inference Theory of Consciousness. *PsyArXiv*, March 11, 2025. doi:10.1016/j.neubiorev.2025.106296.

LeCun, Y. (2022). A path towards autonomous machine intelligence. *OpenReview*. https://openreview.net/forum?id=BZ5a1r-kVsf

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*.

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

McClelland, J. L., Rumelhart, D. E., Group, P. R., & Others. (1987). *Parallel distributed processing (Vol. 2)*. MIT press Cambridge, MA.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.

Piccinini, G. (2008). Computation without Representation. *Philosophical Studies*, 137(2), 205-241.

Putnam, H. (1988). *Representation and Reality*. The MIT Press.

Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind.* Princeton University Press.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3(3), 417-457.

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *The Behavioral and Brain Sciences*, 1–42.

Sprevak, M. (2018). Triviality arguments about computational implementation. In *The Routledge Handbook of the Computational Mind* (pp. 175–191). Routledge.

Sutton, R. (2019). The Bitter Lesson. *Incomplete Ideas*. http://www.incompleteideas.net/IncIdeas/BitterLesson.htm

Tegmark, M. (2008). The Mathematical Universe. Foundations of Physics, 38(2), 101-150. https://doi.org/10.1007/s10701-007-9186-9

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

Wheeler, J. A. (1990). Information, Physics, Quantum: The Search for Links. In W. H. Zurek (Ed.), *Complexity, Entropy, and the Physics of Information* (pp. 3–28). Addison-Wesley.