A Critique of Causal-Functional Equivalence in Mind Uploading

Daniel Burger[1]

[1]**Eightsix Science Ltd**
daniel@eightsix.science

September 27, 2025

Abstract

This essay critiques the functionalist foundations of mind uploading, arguing that the conflation of operational equivalence with causal identity constitutes a fundamental category error that undermines the project's philosophical coherence. The analysis demonstrates that neurocomputational functionalism, whilst avoiding the pitfalls of dualism, commits an equally problematic ontological slide from representation to instantiation. By engaging with contemporary theories of consciousness, the argument establishes that behavioural and even predictive parity are insufficient criteria for the preservation of consciousness.

Table of Contents

1 Introduction

The project of mind uploading, when premised upon neurocomputational functionalism, proceeds from a foundational category error: the conflation of operational likeness with causal identity. It rests on the flawed thesis that a simulans, by merely replicating the simulandum across arbitrarily rich behavioural batteries, thereby instantiates the counterfactual-supporting causal organisation that is the necessary ground for experience, whereas such mimicry may yield nothing more than a hollow operational counterfeit. This essay argues that the functionalist approach to mind uploading fundamentally misunderstands the nature of consciousness and the requirements for its preservation.

2 The Functionalist Foundation

2.1 The Category Error of Operational Equivalence

A more sophisticated proposal suggests that if a model can predict not merely behaviour but also the underlying neural dynamics with sufficient accuracy, it must have captured the essential causal structure. Yet this too is a precarious leap of faith, for it mistakes the conditions of possibility for access consciousness with the fundamental conditions for phenomenal consciousness, that is, for subjective experience itself. This functionalist argument rests upon an implicit, monumental assumption: the Physical Church-Turing thesis (PCT), which posits that the evolution of any physical system, including a brain, is computable by a Turing machine. What must be instantiated, however, is not a function, but a system's intrinsic cause-effect structure: the complete, maximally irreducible web of causal powers it exerts upon itself.

The very term "whole brain emulation" contributes to the confusion, for its etymology is taken from software engineering and evokes the notion of a virtual machine reproducing an instruction set, whereas its scientific ambition ought to include any physical realisation that instantiates the requisite organisation. Treating "emulation" as purely symbolic simulation (even one that generates predictively accurate dynamics) risks reifying the idea that a sufficiently detailed description of causal relations is equivalent to the presence of those relations in a real, physically closed system. This is an ontological slide from representation to instantiation, which must be resisted if the term is to avoid collapsing into a category error.

2.2 Dennett's Framework and Its Implications

Dennett's demolition of philosophical zombies already removes the lazy escape route: one cannot coherently claim a perfect functional duplicate lacks consciousness without smuggling in a hidden inner theatre. Yet his critique does not entail that every system which passes a behavioural equivalence test must therefore instantiate consciousness, even if that test includes predicting internal neural states. The logical question is whether the candidate system truly possesses the same integrated, temporally distributed causal organisation that the brain exhibits, rather than whether some ethereal quale might be missing.

Dennett's multiple drafts model frames consciousness as a process, a rolling pattern of content fixation and competitive revision distributed over time and space. Yet herein lies the crux: to reproduce this pattern is not merely to compute the right outputs, nor even to predict the right sequence of internal states, but to build a system whose internal web of mutual constraints is comparable in its counterfactual profile to that of the biological brain. A predictive automaton, however flawless, models the explanandum; a conscious system is the phenomenon.

3 The Causal Structure Problem

3.1 Architectural Mismatches

Many digital architectures struggle here because their design relies on feedforward modules, computational bottlenecks, and serial arbitration that are fundamentally alien to the brain's integrated nature. This failure is not merely a matter of high-level topology but extends down to the most granular levels of mechanism. Contemporary theories increasingly point to subcellular processes, such as the non-linear integration in the apical dendrites of pyramidal neurons, as a critical nexus for binding top-down context with bottom-up data. A standard simulation merely calculates the result of such dendritic events; it does not instantiate a physical system in which two distinct causal streams actually converge.

3.2 The Granularity Problem

This integrated causal organisation is the dense mesh of dependencies in which many elements are each other's difference-makers. The notion that an emulator's capacity simply increases with the granularity of its data is a dangerous oversimplification. It assumes consciousness is a smooth function of data resolution, when it may in fact depend on a phase transition enabled by the specific causal powers of the physical substrate. Worse still, formal results from computability theory, such as Rice's theorem, establish that for any non-trivial property of a program, such as "instantiates consciousness", it is undecidable whether an arbitrary program possesses that property. We are therefore not merely empirically uncertain as to the correct level of granularity; we are faced with a fundamental, logical barrier to ever knowing if a simulation meets the criteria.

A weather model, even one predicting atmospheric dynamics, predicts the hurricane but does not exert aerodynamic forces. Likewise, an algorithm that computes the state transitions of a cortical column does not instantiate that column's intrinsic cause-and-effect structure; its own structure is merely that of the processor executing the code.

4 Computational Limitations

4.1 The Abstraction Problem

The assumption that consciousness is captured at some high level of functional abstraction is an empirical hypothesis that must survive scrutiny, not a given. Many prominent theories, while disagreeing on specifics, converge on the necessity of recurrent processing

and integration. However, they remain underconstrained regarding the precise physical conditions required. A simulation might satisfy the abstract requirement for "recurrence" by passing a value from one time-step to the next in a buffer, yet this is a causal shadow of the biophysical reality of reverberating, mutually constraining feedback loops.

4.2 The Self-Simulation Fallacy

Recursive self-simulation does not bridge this gap. A simulator that includes a detailed model of its own functioning adds more layers of description but does not generate the basal causal closure that would make those descriptions operative. Dennett's rejection of inner observers generalises here: mirrors nested inside mirrors never produce the fire they represent.

5 The Self-Simulation Paradox

Furthermore, the functionalist hope implicitly requires a Reverse Physical Church-Turing thesis (RPCT): that our universe can physically host a computer of universal power. If we grant both theses, a startling consequence follows, not of simple identity-transfer, but of ontological catastrophe. Computer science proves that a universe obeying both PCT and RPCT can, in principle, perfectly simulate itself. This "self-simulation lemma" reveals an infinite regress: within our universe is a computer simulating our universe, which contains an identical copy of us running an identical computer, which in turn contains a simulation of our universe, and so on, ad infinitum. Within this framework, we are simultaneously the simulator and the simulated. The very notion of a singular, originary "self" to be preserved dissolves into an infinite cascade of causally interdependent, ontologically equivalent instances. We are denied any basis for distinguishing the "real you" from the "copy", for each instance is as physically real as the next.

6 Towards a Processual Account

6.1 Reframing the Problem

The claim that an upload is conscious should therefore be reframed as a demand for equivalence of intrinsic cause-effect structure, rather than a simple assertion of behavioural or even predictive mimicry. Behavioural parity can be obtained through lookup tables, and predictive parity through statistical modelling, both of which may be implemented on a fragmented architecture that eliminates the very counterfactual richness that makes the biological system what it is. To argue that an imperfect simulation is "good enough" because biological systems are also noisy is to commit a profound category error. The noise in a brain is a property of a physical system; the "noise" in a simulation stems from modelling imperfections. Conflating their statistical profiles with ontological parity is to mistake the map's smudges for the territory's terrain.

Dennett's framework, properly understood, sharpens this demand. His rejection of the inner theatre removes the possibility of asserting that a transcript of events is enough. What

must be replicated is the sprawling, decentralised coalition of processes. An emulation that collapses this coalition into a central, predictive bottleneck has not reproduced the brain's functional organisation but replaced it with a fragmented artefact whose own cause-and-effect structure is trivial, and entirely dissociated from the structure it simulates, even if it is computationally capable of self-simulation.

6.2 The Synconetics Approach

At this juncture, the inquiry benefits from the conceptual machinery of Synconetics, which seeks to formalise consciousness not as a static property of states but as a processual trajectory through a four-dimensional causal manifold. This requirement enforces not merely computational equivalence but strict physical realisability and substrate-invariant functional isomorphism, excluding solutions that rely on mere symbolic mimicry. Incorporating this invariance criterion into the problem of mind uploading reframes the goal as one of dynamical continuation rather than digital reproduction. The central question is not whether a simulation produces functionally similar outputs but whether it traces a process that is causally isomorphic to the biological trajectory it aims to preserve.

7 Conclusion

The term "whole brain emulation" should therefore be expanded to include not merely software descriptions but physically instantiated systems that reproduce the relevant causal nexus across all causally potent scales. Only when we build a neuromorphic substrate that achieves comparable concurrency, spatial locality, and subcellular non-linearities, and can demonstrate that it constitutes a single, integrated causal complex, does the claim of consciousness in silico gain empirical standing. Until then, the ontological status of the simulans is not mysteriously indeterminate but epistemically underdetermined and undecidable in principle.

Such a reframing forces the debate into a much more stringent register. It demands that mind uploading be judged not by its surface resemblance to cognition but by its capacity to preserve the ongoing process of identity with interventionally robust fidelity. Current von Neumann and Turing-style architectures, with their fragmented causal structures, appear ill-suited to sustain this kind of processual continuity. To present purely digital whole brain emulation as already sufficient is not optimism but ontological hubris. The burden of proof lies with the emulation community to demonstrate that the simulated process is not merely a record but a causally closed continuation of the original world-line. Until that case is made, claims that uploading will trivially deliver personal survival remain promissory gestures, and the conflation of computational fidelity with ontological sufficiency risks producing not preserved persons but exquisitely detailed counterfeits trapped within an inescapable, undecidable logic, where no one is home to inhabit.