Could Computer Programs Become Conscious?

Cvijetić Dušan, School of Basic Sciences
Kolly Florian, Neuro-X Institute
Sobotka Jan, School of Computer and Communication Sciences
Zhang Michael, School of Basic Sciences

Project SHS 1st year master

Supervised by

Esfeld Michael-Andreas, Philosophy of Science
Cucu Alin, Philosophy of Science

Date of last update: 29.05.2025
Word count: 7009

Lausanne, academic year 2024 – 2025

EPFL

Abstract

This essay investigates whether computer programs can achieve consciousness, navigating the intersection of philosophy, cognitive science, and computer science. Beginning with a historical overview of consciousness from Plato to Chalmers' "hard problem", it contrasts physicalism, which grounds consciousness in physical processes, with dualism, which views both mental and physical as fundamental. Adopting physical causal closure and functionalism, the essay introduces the concept of simulated dualism, explaining why standard dualism seems natural. Then, it describes and argues for Virtualism, a computational theory that views consciousness as a representation of oneself experiencing something. In turn, Virtualism is used to address common arguments like the Knowledge and Conceivability Arguments and defend the possibility of computer consciousness against objections based on biological naturalism and embodiment. Finally, this essay reframes and answers positively to the question "*Can computer programs implement consciousness*?", arguing for why we should reconsider the mind-machine boundary.

Introduction

Behind the question of whether computer programs can become conscious lies a more fundamental problem: defining what consciousness is. Once we establish a clear definition of consciousness, figuring out whether a program—or any other entity—could be conscious becomes far simpler, as knowing the necessary conditions for consciousness would allow us to determine whether it could be replicated or emulated within a computational framework. The concept of consciousness, however, seems to be rather elusive. Not only is it challenging to study the properties and behaviors of a conscious mind, but the object of the research itself is very hard to pin down. This difficulty leaves us grappling with a central problem: before we can determine whether machines can be conscious, we must first understand what it means to be conscious at all.

The inquiry into consciousness has evolved dramatically across the history of Western philosophy. Ancient Greek thinkers established foundational approaches: Plato conceived consciousness through his theory of an immaterial soul capable of apprehending eternal truths, while Aristotle offered a more naturalistic view, describing the soul as the "*form of the body*" and connecting perception to physical processes (*De Anima*, III.5).

Medieval philosophers like Augustine and Aquinas integrated these classical ideas within Christian theology, positioning consciousness as a bridge between material existence and divine reality. This framework was radically transformed during the Early Modern period when Descartes (1641) established his influential dualism, separating the

mental realm (*res cogitans*) from the physical (*res extensa*). This division created the mind-body problem that continues to shape contemporary discussions.

More recently, the 19th and 20th centuries brought diverse approaches to consciousness. William James (1890) characterized it as a continuous "*stream*", while phenomenologists like Husserl (1913) examined consciousness's intentional structure. The rise of cognitive science and neuroscience introduced materialist theories suggesting consciousness emerges from neural processes, yet dualist challenges persisted—notably in Chalmers' formulation of the "*hard problem*" (Chalmers, 1996), questioning how physical processes generate subjective experience.

Today, there are multiple competing approaches in trying to explain consciousness. The two most prominent schools of thought are that of physicalism and dualism. Based on the empirical premise of the completeness of physics, physicalist theories claim that consciousness is—or supervenes on—physical states or processes. A closely related view, which differs only in that it does not include physical objects such as energy, space, and time, is materialism. Materialism holds that everything that exists, including mental phenomena like consciousness, is entirely explicable in terms of material components and their interactions. On the other hand, dualists argue that the mental and the physical are both equally fundamental and stand in a causal relation. (Morch, 2023)

The deep rift between these two approaches is best reflected in what Chalmers (1996) calls the epistemic gap between the easy and hard problems of consciousness. He proposes that it is easy to describe the functional and structural properties of the mind. That is, even though answering these questions might prove to be quite technically involved, there is a clearly defined research program to be followed. This line of inquiry often yields what is named functional consciousness, a definition of consciousness based on its functions and physical properties. Tackling the hard problem means explaining phenomenal consciousness. Phenomenally conscious states are characterized by the fact that there is "*something that it's like for a creature or entity to be in them*" (Morch, 2023). This poses a conceptually much harder task, as it becomes necessary to explain how a physical system can give rise to a conscious experience.

The main dualist arguments are the knowledge, the conceivability, and the explanatory argument. The knowledge argument states that consciousness cannot be deduced from any physical knowledge, and therefore it is not physical (Jackson, 1982). The conceivability argument, most famously presented by Chalmers (1996) through the zombie thought experiment, claims it is conceivable for consciousness and the physical to exist independently, therefore implying that consciousness is non-physical. Finally, the explanatory argument, made famous by the Leibniz's Mill, states that consciousness cannot be explained in physical terms, so it cannot be physical itself (Leibniz, 1714).

A good theory of consciousness needs to provide an answer to the fundamental question of what consciousness is and how it arises, while at the same time supporting or fending off existing arguments that dominate contemporary discourse in the philosophy of mind.

Constructivist path toward computer consciousness

Before we explore whether computer programs could become conscious, we must first address metaphysical questions about existence itself and clarify our stance on the topic of this essay.

Our starting point is the axiom stating that existence is the default. Reminiscent of Descartes's way of argumentation, this is reasonable as the idea of existence itself is self-proving since it could not be entertained without such an axiom. The only question remaining is: The existence of what?

According to Papineau (1993), causal closure of the physical (PCC) follows from the completeness of physics. He considers physics "*is complete, in the sense that all physical events are determined, or have their chances determined, by prior physical events according to physical laws*" (pp. 16). The completeness thesis considers a future complete theory of physics, and is based on two assumptions: 1. that such a theory is feasible and, 2. that such a theory will make no use of psychological categories.

PCC has been under much fire from the philosophical community. For example, Kim (1989) noted that reductionist theories were seldom popular among philosophers, and accepting both PCC and grounding of mental states in the physical necessarily leads either to a reductive or eliminative theory. A more concrete attack on PCC is given in a yet unpublished manuscript by Cucu (2025). Specifically, Papineau's claim of completeness of physics can be attacked in two ways. One is to reject the first premise, questioning whether a complete theory of the physical is even attainable. But even entertaining the possibility of complete physics, Cucu argues that PCC should refer to sufficient causes, which need not be necessary. That would mean physics can remain complete and explain physical effects by physical causes, while still allowing for mental causes, which would lie outside its domain of interest. Such a view is compatible with Noether's theorem and allows for a completeness of physics compatible with the dualist approach to the theory of mind. It is worth noting that such a view is directly compatible with Papineau's completeness of physics thesis, but not with his definition of PCC, which posits physical causes as exclusive.

The situation is mostly different in the scientific community, where PCC is almost unanimously taken as a premise in scientific practice. Most of the arguments about PCC find their support in the empirical evidence, either when referring to the conservation

of energy, completeness of physics, or any of the many other approaches to the subject. Since, at its root, PCC seems to be an empirical question, and since no strong empirical evidence has been found against it, we shall take the stance that PCC holds and develop our theory from there. It is, however, worth noting that not much experimental work has been put into trying to (dis)prove PCC (Cucu, 2025). Nevertheless, since experiments in physics and neuroscience seem consistent with PCC, there is not much empirical justification in doubting it. Moreover, we would argue that the absence of **expected** evidence of non-physical forces is a strong indicator that the initial expectation is flawed. For example, if extensive investigations on Earth had failed to find any evidence of dragons (e.g., bones, nests, ...) in areas where they would logically be expected, the lack of evidence would support the idea that dragons are fictional.

Taking PCC to be true, the argument is then simply that, if all physical effects are exclusively due to physical causes, then anything that has a physical effect must itself be physical (Papineau, 2001). This will be the stance we take for the rest of the paper.

Now, what could a satisfactory physicalist theory of consciousness, that would help answer the overarching question, look like? To be satisfactory, it would need to have empirical support—be it a first-person reflection with sufficient similarity across people or evidence from natural sciences. This discards theories that posit internal essences of things since those are not observable. In light of this fact, functionalism emerges as the most sensible and practical stance, as it directly grounds consciousness in observable causal roles. It says that object categories, including mental states, are defined not by their physical composition or underlying substance but by their **causal relationships** to sensory inputs, other mental states, and behavioral outputs and effects. For example, functionalism would assign the category of water (H_2O) based on the material's interactions: it forms hydrogen bonds with each other, creating its fluid yet tense surface; it boils and freezes; on touch, it leaves our hands wet, etc.

Computational functionalism takes this a step further and says that consciousness depends on computational processes implemented by our brains, rather than, for instance, the fact that a specific subnetwork of neurons is firing. It uses the fact that computation describes an arbitrary system through its states and a deterministic or probabilistic transition function between these states, and lifts consciousness from a functional unit of a brain (be it biological or in silico) into a computational framework. Therefore, computational functionalism posits that *simulating* the same computational process (i.e., sequence of state transitions) that underlies the mind is necessary and sufficient to make a system conscious. In the following text, we will argue for this perspective and revisit its nuances. For now, it is important to clarify what a *simulation* is: Simulation is a process with equivalent structural and dynamical properties as the thing being simulated, but with different causal underpinnings (Bach, 2021). In simpler terms, a simulation is a model of some domain on a possibly different substrate.

A common class of objections to (computational) functionalism is that a simulated hurricane is not a real hurricane. For example, Searle (1980) writes: "*No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched*". In our view, Searle's objection makes the error of mixing the causal structure of the simulation itself with the causal structure of the thing being simulated. For example, when software engineers say that they simulated a rainstorm in their computer game, they mean that the variables of interest within the simulation match the variables that would have been measured had the storm really existed in the base (physical) domain. In simpler terms, the software engineers mean that any character **inside that game** will get drenched. This is the fundamental idea of functionalism: The simulated rainstorm has the same causal relationships to other entities as the *real* rainstorm (the one in the base physical reality to which Searle referred).

Building on physicalist computational functionalism, we begin by proposing a unifying view on dualism and physicalism, and why physicalism, in particular, might not seem enough to explain all mental phenomena, such as consciousness, and a way to resolve it. This, in turn, will allow us to fully address the question of whether computer programs could become conscious.

Simulated dualism

Gottfried Leibniz, with his Mill Argument against materialism, asks us to imagine a hypothetical machine with the ability to think and sense as we do, enlarged to such a scale that a person could walk inside it. He argues that even if we were to inspect every part of this vast apparatus, we would observe only things pushing and pulling one another, but nowhere would we find feelings or conscious perception. His conclusion is that such phenomena reside in the "*simple substance, and not in the composite or in the machine*" (Leibniz, 1714). (Leibniz's point is that perception and consciousness arise from indivisible, immaterial entities he calls "*monads*", the true units of reality, rather than from any mechanistic arrangement of matter). Leibniz's thought experiment nicely illustrates the distinction between physical and psychological reality: While the subjective feeling of a cold breeze or the redness of an apple **exists** within the mind (psychological reality), physical reality knows only of the **existence** of electromagnetic waves and atoms pushing and pulling one another.

This sharp distinction leads to substance dualism, which puts the mental substance on equal footing with the physical substance, setting both the physical and the psychological realities as fundamental. However, physicalists, including ourselves, would reply that since substance dualism posits an interaction and an inter-causation between the two realities, the mental might as well be part of the physical as it is the causally closed lower layer of the universe, backed by strong empirical support from the natural sciences. In other words, if the mental has causal power over the physical, then,

by definition, it is part of the physical and should not be taken as another reality but merely as a class of representations that physical entities (e.g., humans) create.

Our theory starts with the position that the physical reality is being simulated within a physical brain as a representation, along with the subject (the *self*) being modeled therein. This simulation, akin to a game engine, is part of what we call the psychological realm.

In this psychological realm, anything representable, such as magic or phenomenal experience, is possible. In this realm, mental phenomena such as feelings and consciousness can be portrayed to be experienced by the *self*—the main character modeled as a physical part of the simulated physical reality. We are taking the first-person perspective of this character. This also explains why dualism feels natural to many philosophers: The brain creates representations of the mental and physical realities next to each other as separate substances, creating a form of *simulated dualism*.

However, for the psychological reality to exist, it needs to be implemented in some way. The completeness of physics discussed earlier implies that the psychological reality is implemented in the physical reality or as a part of a causally ineffective soul. The latter is most commonly called epiphenomenalism and has mostly been abandoned by the philosophical community. For this reason, and for our stance that the mind has in some way causal effects (discussed later), we will proceed with the first option: The psychological reality is implemented in the physical brain of a primate. Building on this, we now turn to computation as a powerful tool to describe any such physically implementable system by its states and deterministic or probabilistic transitions between these states.

In modern days, the most well-known physical incarnation of a computational description is a (digital) computer. Just as the brain—a physical *thing*—constructs a mental image of the physical world based on its sensory signals, computers construct virtual snapshots of their input signals. Computer programs, also sometimes called software, form the intangible parts of computers, while hardware forms the tangible: the circuits, wires, and processors. The connection between the two is that software is the sequence of state transitions that governs the hardware's behavior. As philosopher Joscha Bach put it, software is a *physical law*: If one arranges transistors (the matter) on a computer in some particular way, some specific sequence of state transitions will occur (Bach, 2024a). In this respect, one cannot assign a fixed identity to software. For example, Microsoft Word running on one computer is not the same as Microsoft Word running on another computer with a possibly completely different architecture. Only the pattern—the evolution of state transitions—is the same. Bach has also suggested that software might be likened to a *spirit*, an organizing principle that animates the machine, much as the mind animates the body (Bach, 2024b). This analogy reveals the

21st-century version of the mind-body problem: the psychological reality (the mind), like software, seems to exist as a dynamic process, while the body, like hardware, provides the physical substrate.

Through this lens, the disagreement between the major schools of thought becomes clear. Eliminative materialism views everything as a part of the physical reality, trying to set fixed physical identities to everything. According to eliminative materialism, software is hardware. The opposite monist view, idealism, acknowledges that we cannot access the substance of physical reality and that it is merely within our minds. Thus, idealism posits running software without any underlying hardware. Property dualism keeps the physical reality and views the psychological merely as a set of irreducible properties. It acknowledges that the software runs on some hardware, but does not attempt to deduce the computer code (which is written in memory blocks of the hardware) from the running software, denying its existence. Finally, while property dualism acknowledges that the psychological reality cannot exist without the physical one, substance dualism says that the two realities can exist independently: software and hardware running independently but interacting with one another.

The computational lens also lets us explain why physicalism, as a broader class of theories, appears not to be enough to explain consciousness: When physicalists attempt to explain consciousness, which exists solely within the psychological reality, they can only describe the physical implementation of this consciousness software, the neural hardware and its processes. Yet, the human mind, armed merely with knowledge of the hardware (e.g., the charge at neuronal membranes, the specific interconnections between neurons), cannot transform that detailed symbolic manual into a running program capable of experiencing phenomenal consciousness, such as the felt quality of redness. Consequently, physicalism seems to struggle to bridge the explanatory gap—the divide between physical processes and subjective experience.

Virtualism

With the aim to bridge the explanatory gap and explain how subjective experience can arise from physical processes, a theory of consciousness called Virtualism, sometimes also referred to as the Cortical Conductor Theory (Bach, 2017), combines physicalist functionalism with the computational framework. It applies the representational view of the mind, which we discussed in the previous chapter, to consciousness itself.

Virtualism can be described by starting from the following proposition: "*If you do not remember what you paid attention to, you couldn't be conscious*". In other words, one could not say that they are conscious unless they view themselves as being aware of or perceiving something in the immediate past. Such a representation of themselves in the past could be called a *memory*, which, however, is not completely truthful. The reason being that when the psychological reality changes the self to think that it is conscious

(i.e., includes phenomenal experience as part of its perspective), it needs to create an image of that which has been experienced, and such an image is purely imaginary with no requirements to adhere to what actually happened. A well-known example of such false inner pictures that brains can create are optical illusions.

The consequence of the above proposition is that consciousness is not possible at a single time point in physical reality, but is only possible as a sequence of time steps in psychological reality. Hypothetically speaking, if we were to freeze time, consciousness would stop. It is also what functionalism suggests—the causal role that consciousness plays would halt. Even though all physical materials would be the same, consciousness would be nowhere to be found. In this respect, **consciousness exists only in the eye of the conscious observer**. The perspective that we are not actually conscious in a particular moment, but merely remember as being conscious, is supported by inconsistencies in our subjective experiences. For example, time dilates subjectively as we go through more or less intense events, and some events are falsely experienced as continuous in time.

The view that phenomenal experience is a certain representation of the self as viewed from the first-person perspective is what gives Virtualism its name. It posits that consciousness is *virtual* (i.e., appearing to exist but not existing as part of the material world). Stated differently, the (virtual) consciousness is a representation of what it would be like if the simulated *self* was phenomenally experiencing something. As Joscha Bach, the main proponent of Virtualism, puts it:

> *The reason why we experience things in a particular way is the same why a character in a novel does: because the contents of our experience and the fact of the experience itself are written in exactly this way by its author. Like a character in a novel, we generally also don't notice that we are not real, as long as the author does not write the discovery that we are not real into our story.* (Bach, 2019)[1]

This idea also allows for answering questions such as "*What is it like to be a bat?*", famously put forward by Thomas Nagel (1974). Granted that bats have a rich mental life, our brains would need to run similar software to them to answer such a question. It would need to instantiate a similar causal structure of mental processes as bats have, ultimately restructuring our psychological realm into what bats possess. However, just as characters inside a computer game cannot change the game itself any way they would like, we cannot fully control or completely overwrite our *cognitive software stack*, rendering Nagel's question impossible to answer through these means.

[1] The word "*author*" in this passage refers to the physical brain.

The fact that our brain is organized in such a way that our symbolic knowledge does not have direct access to the implementation of the psychological reality also addresses the well-known Knowledge Argument, exemplified by the Mary's Room thought experiment (Jackson, 1982). This experiment imagines Mary, a scientist who has read everything there is about color vision and gained all the symbolic knowledge about it, but who is confined to a black-and-white environment. When she finally sees color, the crucial question arises: does she learn something new, the *subjective experience*, or *what it's like* to see red? If she does, it suggests that her knowledge of all the physical facts did not grant her access to this aspect of the psychological reality, thereby challenging the physicalist view that all facts are reducible to physical facts. However, in our view, this is not a refutation of the physicalist approach to consciousness, but instead an observation about symbolic knowledge and the limitations of our brains. Specifically, even if Mary reads about someone else experiencing the redness of an apple, including all the neural processes behind it, she cannot turn this symbolic knowledge into her own phenomenal experience—she cannot change the source code running on her brain. Therefore, she cannot change what the psychological reality presents her with. When she steps out of the black-and-white room, her psychological reality gets richer as a result of some physical brain circuits being stimulated in a new way and learning to create a new feature dimension of her virtual experience.

This apparent gap between symbolic, third-person knowledge and first-person phenomenal experience is also what the Global Workspace Theory (GWT) (Baars, 1988), favored by many scientists, attempts to explain within a physicalist framework. According to GWT, consciousness is conceived as arising when information is broadcast within a central global workspace, making it widely available to numerous unconscious specialized cognitive processes (language, motor control, etc.). From this perspective, Mary's predicament can be reinterpreted. Before her release, her comprehensive symbolic knowledge about color resided within specialized knowledge systems but was never integrated with the relevant perceptual input and broadcast globally as a conscious percept. While she knew about the neural correlates of seeing red, this information itself was not the kind of information (sensory data processed and attended to) that can gain access to the global workspace to become a conscious experience of redness. Only when she finally saw red, the sensory information was processed, and gained access to the workspace. What she learned, therefore, is not necessarily a non-physical fact, but rather the functional consequence of this specific information type being globally available—the state of the brain when the sensory stimulus of redness is being broadcast. The implementation of the psychological reality, in GWT terms, involves this dynamic process of information broadcasting, which her symbolic knowledge alone did not replicate.

As we can see, Virtualism takes a strong computational stance toward explaining consciousness, and its resulting explanations coincide with more well-known theories such as GWT. Indeed, terms used by Virtualism for understanding consciousness, such

as "*representation*" and "*simulation*", are well-known concepts to computer scientists. What is important for the purposes of this essay, however, is that all the terms and processes used by the theory are implementable on digital computers. This last step from biological brains to computers will be our focus in the next section.

Computed consciousness

Virtualism views consciousness as a representation of what it would be like if the *self* had a phenomenal experience. Brains, as the only evidence of consciousness, are clearly capable of creating such inner simulations, for otherwise we would not say we are conscious. Now, as far as the science of the last century knows, brains are physical. The quest of neuroscience—the objective study of the brain—is to map out this physical apparatus. Since all the findings so far about the brain take the form of algorithmic procedures (i.e., how the brain's components interact, what cause-effect relationships there are), there does not seem to be any empirically supported reason why the brain mechanisms could not be functionally replicated on a silicon-based computer. We will revisit some of the objections later, but for now, it is important to note that a physical object—the brain—is capable of creating consciousness as a virtual property of the story that the brain tells itself. Indeed, consciousness being virtual implies that it has no restrictions on the substrate, and hence can be instantiated as a program on any sufficiently powerful computer. We could, in principle, simulate a fictional world—the psychological world similar to what our mind presents to us—on a computer. If we connected the computer to external sensory inputs, this psychological world could also simulate the physical world, where it could place the main character. To make the main character conscious, we could endow it with a strong self-reinforcing belief in their phenomenal experience, a representation of themselves being in a conscious state. Based on how Virtualism explains consciousness, such a being, nested in multiple levels of simulations, would pass the necessary and sufficient conditions for consciousness: it would represent itself as having experienced something, and portray it as a fully truthful and undeniable memory on which it would be able to report. With this functional description in hand, let us now turn to possible counter-arguments against the claim that such programs could instantiate phenomenal experience.

A first objection comes from perspectives emphasizing the biological grounding of consciousness, such as that articulated by Anil Seth. According to Seth, consciousness is inextricably linked to the organism's fundamental biological imperative of staying alive. In this view, subjective experience arises not from computation alone, but from the brain's continuous process of predicting sensory inputs and updating these predictions based on actual sensory information, a process deeply rooted in the organism's overall biological regulation and survival needs (Seth, 2021). Silicon-based machines would lack the biological and physiological context of the organism, suggesting that merely replicating functional roles in a non-biological substrate might produce a functionally equivalent zombie, but not genuine conscious experience. Seth

therefore argues that the functions or computations implemented by conscious biological systems may not be separable from their material basis.

The approach favored by Anil Seth can be seen as an evolved, more teleologically-driven version of biological naturalism, originally proposed by John Searle (1980). In its original formulation, biological naturalism proposes that conscious states are higher-level and irreducible properties of lower-level neurobiological states. Searle advocates that consciousness exists because biological brains have the *right causal powers*, and not because of the functions they produce. A notable challenge to this view is Haugeland's *Searle's demon* critique (Haugeland, 1980), questioning why biological causation is somehow special. The critique goes as follows: Imagine a demon inside Searle's brain who manually implements every synaptic firing and neural interaction exactly as the neurons would, without any consciousness of its own. The brain behaves identically, but the demon does not understand anything; it is just recreating every interaction. If consciousness depends on biological causation, as Searle argues, then the demon-operated brain is not conscious. However, the demon is performing the entire causal process of a biological brain, effectively simulating the causal powers of biology. It is unclear, then, why biological causation is special and forces Searle to accept the uncomfortable position of carbon chauvinism.

Seth's position modifies this somewhat. For him, consciousness serves a purpose in biological self-regulation. By arguing for his view, he makes a distinction between the types of computation that a brain and a computer can do. Standard Turing computation (abstract model of modern-day computers) is immortal, that is, its existence outlasts the existence of any specific instance of hardware. Biological brains cannot implement immortal computations, meaning that the computation cannot be separated from the hardware that implements it. As a result, his position places constraints on the substrate independence of mortal computations, such as consciousness. In particular, the substrate independence required for conscious programs is unlikely to hold because those programs are based on an implementation paradigm that assumes computational immortality.

Here, Seth argues against the underlying framework upon which Virtualism is based, computational functionalism. Specifically, Seth disagrees with the idea that mental states do not depend on their specific underlying substrate (i.e., substrate independence), but solely on their function within it, and that any such function is describable in computational terms. However, a well-known argument supporting substrate independence is the Neural Replacement thought experiment (Chalmers, 1995): Imagine that, one by one, each neuron in a person's brain is replaced with a functionally equivalent silicon counterpart that perfectly replicates the input-output behavior of the original biological cells. If this replacement process were carried out seamlessly, the person's behavior, cognition, and subjective experience would remain

unchanged, demonstrating that consciousness depends on functional structure rather than on the biological substrate itself.

Critics of functionalism have offered alternative outcomes to the Neural Replacement experiment. For example, Block (1978) points out that the functional organization of the brain could, in principle, be instantiated by the population of China, and questions the plausibility of attributing mental states to such a system. Chalmers (1995) replies to this *absent qualia* objection by arguing that, if *absent qualia* are possible, then a phenomenon he coins as *fading qualia* is also possible, but that such *fading qualia* are deeply implausible. More concretely, imagine again the replacement of neurons with functionally equivalent silicon counterparts, but suppose this time that, once all neurons have been replaced, there is nothing it is like to be the system. Then either consciousness gradually fades over as more neurons are replaced, or there is a point at which consciousness suddenly disappears. The second case seems unlikely, as it implies that replacing a single neuron could trigger a complete collapse of experience. Moreover, one could also start to replace biological matter at the molecular level in this specific neuron, finding a single molecule that could switch between experiencing and not experiencing. If we then suppose that consciousness is fading as replacement progresses, the system will have increasingly dimmed, degraded or partial experiences. Its cognitive functions and behavior, however, remain the same, as we suppose a seamless replacement that preserves the functional organization. Chalmers argues that this fading scenario is implausible because it would require the system to be massively mistaken about its own conscious experience while remaining fully rational and behaving normally.

A final reply to Anil Seth and to those who tie consciousness to quantum mechanisms in biology (e.g., Roger Penrose and Stuart Hameroff) is to note the structural and functional similarities between the operations in computer programs, particularly in deep learning systems, and biological brains (Yamins, 2016; Kubilius, 2019; Hosseini, 2024; Kell, 2018). While these analogies do not imply equivalence, they support the plausibility of certain computational systems developing forms of representation and processing that echo those found in biological minds.

Now, even taking computational functionalism as the starting premise, Kleiner (2024) shows that computational functionalist theories imply that consciousness is a mortal computation (as defined earlier). However, for his proof, he must assume the existence of an organism that is capable of conscious experiences that cannot be programmed in today's framework of Turing computation. We would argue that the two Kleiner's assumptions are incompatible: Namely, if he assumes computational functionalism, then there cannot exist an organism that could not be programmed. This is because stating that it is not (Turing) programmable but still within the computational functionalist framework leaves us only with two options: (1) the organism's behavior involves randomness of a kind that cannot be probabilistically modelled, or (2) the

behavior could be computational via non-Turing computations, such as hypercomputation. To argue against (1), suppose such an organism exists. As consciousness depends on the causal organization, it is unclear how stable and continuous conscious experiences could arise from the incomputable randomness. Additionally, physical processes, including quantum phenomena, exhibit randomness that is statistically modelled, contrary to the organism. Postulating the existence of a novel and unobserved kind of physical law introduces additional complexity and violates principles of ontological economy. Finally, one might object that even if the organism is not programmable within the Turing framework, it could still be computable under some broader notion (option 2), for example analog computation or hypercomputation. To the extent that non-Turing computational models invoke infinite resources, infinite precision, or idealized mathematics, they fail the test of physical realizability.

Another way to argue against the possibility of computer programs being conscious is to point out that such programs are disembodied. Disembodiment here is not meant as being immaterial, but is meant in the sense that they do not interact with the world in continuous time through a unified sensorimotor system. Whereas humans (and other animals) possess a spatially confined body that serves as the hub for perception and action, computer programs rely on distributed and distant racks of physical computers and interact with the world only indirectly through networks, sensors, and interfaces. Embodiment is thought to be closely linked with consciousness through what Hurley (2001) describes as a "*perspective*": One's experiences and perceptions must systematically depend on one's actions, and vice versa. Since consciousness arguably requires a unified, situated perspective, the lack of a cohesive embodiment presents a serious challenge to conscious computer programs.

This challenge is, however, mitigated by the idea that embodiment does not necessarily have to be physically instantiated. Computer programs could be made such that they model the causal structure of the world, thus forming internal representations of how their actions will impact their perceptions of the world. This simulated perspective is in line with the view of Virtualism and of the *phenomenal self* of Metzinger (2003), where our selves are merely representational appearances from the computational processes in the brain. As per substrate independence, computer programs could as well as us model such representational spaces, thus creating for themselves a sense of embodiment. Moreover, in case the above argument does not prove to convince the reader, a more traditional reply to the embodiment argument is to simply state that there is no barrier in giving computers embodiment, that is, allowing them to interact physically and directly (as in, not through a distant interface) with their environment. Robots in the common sense are controlled by computer software and operate in the physical world, and thus possess an embodiment.

Finally, let us entertain and reply to the Philosophical Zombie thought experiment. To remind the reader, this thought experiment presents a hypothetical scenario of microphysically identical entities to human beings, called zombies, that lack phenomenal consciousness. The argument goes as follows: (1) We can imagine such zombies (conceivability). (2) If zombies are conceivable, then they are metaphysically possible. (3) If zombies are possible, consciousness is nonphysical, rendering physicalism false.

Our reply to this goes back to our discussion about *simulated dualism*. Specifically, this thought experiment confuses the psychological reality, where anything representable, such as philosophical zombies, can be part of the simulated physical world, with the real physical world, where only entities that adhere to physical laws are possible. This also applies to any other universe that one could imagine, and where the mind could be conceived of as a separate substrate. In other words, the thought experiment, entertained in the psychological realm, presupposes its conclusion by mere imagination—by simulating dualism.

It is important to note, however, that even if the argued implication in the Philosophical Zombie thought experiment was invalid, it would not imply that consciousness is physical. Consciousness is itself immaterial in the sense that software is immaterial: It makes sense to talk about software or about consciousness only as some specific sequence of state transitions. At a single point in time, the software would be only the arrangement of transistors (i.e., hardware), just as consciousness would be only the arrangement of the brain. Eliminative materialists (strict identity theorists) would argue that this is all there is, challenging the existence of consciousness. However, no two physical states are ever **exactly** the same, and hence reducing everything to physical states of a single point in time does not make sense from an epistemological perspective. Moreover, it is not clear how eliminative materialists can think that they are referring to the same thing as everyone else when they say that consciousness does not exist: if it were just a snapshot of the physical arrangement of the brain, they could never talk about the same thing in retrospect. Additionally, the neural replacement argument for functionalism applies.

Note that our claim that consciousness is immaterial does not imply standard dualism. The crucial distinction lies between *being* material and *supervening* on the material. Phenomenal experience—ultimately implemented by physics—has causal power over the physical reality through its physical implementation. Note that this does not imply causal overdetermination (both the physical and the mental states being sufficient for the physical effect): The mental state cannot interact with the physical reality through other means than through its physical implementation, and hence it is not sufficient **on its own** to physically cause anything (and is never on its own).

Let us now turn back to the initial goal of this essay. The original question, "*Could computer programs become conscious*?" asks whether computer programs (subject) could gain the property of being conscious. This, in our view, is rather unfortunate phrasing. We argued that psychological reality (the mind) is itself a computer program where anything representable, such as magic or philosophical zombies, is possible. Virtualism took this a step further and explained consciousness as being virtual: A computer program that represents what it would be like if the simulated self was phenomenally experiencing something. On this account, we have replied to counter-arguments and provided justification for answering positively to the question: "*Could computer programs implement consciousness*?"

Conclusion

In this essay, we have explored the topic of computer consciousness, discussing intersections of philosophy, cognitive science, and computer science. Our journey began with an examination of the metaphysical foundations of existence, adopting the axiom that existence is self-evident and grounding our inquiry in the debate between physicalism and dualism. We established physical causal closure as a guiding principle, supported by in-so-far empirical evidence for the completeness of physics in disfavor of dualism. Furthermore, we established functionalism as our stance, motivated by the fact that internal essences are not observable and that any two entities exhibiting the same sequence of state transitions and causal interactions are necessarily the same things.

We then discussed the nature and origins of consciousness, where the explanatory gap between physical processes and subjective experience seemed as a barrier to many physical accounts of consciousness. To address these limitations, we explored a physicalist computational functionalist theory of consciousness called Virtualism, which views consciousness as a virtual simulation of a self experiencing phenomena within the psychological reality.

Virtualism posits that consciousness emerges as a sequence of representational states, akin to a computer program running on a physical substrate, such as the brain. This framework allowed us to reconcile the psychological and physical realities, addressing classic objections like Leibniz's Mill Argument and the Knowledge Argument by clarifying that phenomenal experience is a representational construct, not a separate substance. We further defended Virtualism against biological objections, such as those from Anil Seth and Roger Penrose, by highlighting structural and functional parallels between biological brains and existing computer systems, and supporting the idea of substrate independence using the neural replacement thought experiment. Then, we countered the embodiment challenge by suggesting simulated embodiment or physical robotic embodiment of computer programs.

Finally, we argued against the implications of the Philosophical Zombie thought experiment by exposing its confusion between psychological and physical realities, reinforcing that consciousness, while immaterial in the sense of software, supervenes on physical processes without requiring dualist assumptions. By reframing the essay's central question to "*Could computer programs implement consciousness*?" we aligned it with Virtualism's computational stance, concluding that consciousness is a process that can be instantiated in any sufficiently powerful computational system.

In conclusion, this essay affirms that computer programs can, in principle, implement consciousness by simulating the causal structures of phenomenal experience, as described by Virtualism. This perspective bridges the explanatory gap, integrates empirical and philosophical insights, and challenges the boundaries between biological and artificial minds.

Bibliography

1. Aristotle. [ca. 350 BCE] 1931. "De Anima." Translated by J. A. Smith. Oxford: Clarendon Press.
2. Baars, Bernard J. 1988. "A Cognitive Theory of Consciousness." New York: Cambridge University Press.
3. Bach, Joscha. 2017. "The Cortical Conductor Theory: Towards Addressing Consciousness in AI Models." Paper presented at the IJCAI-17 Workshop on Architectures for Generality & Autonomy (AGA 2017), Melbourne, Australia, August 19–20. Accessed April 30, 2025. http://cadia.ru.is/workshops/aga2017/proceedings/AGA_2017_Bach.pdf
4. Bach, Joscha. 2019. "Phenomenal Experience and the Perceptual Binding State." In Proceedings of the Toward Conscious AI Systems Symposium (TOCAIS 2019), AAAI Spring Symposium Series, Stanford, CA, March 25–27. Accessed April 30, 2025. https://ceur-ws.org/Vol-2287/paper29.pdf
5. Bach, Joscha. 2021. "Virtualism as a Perspective on Consciousness." YouTube video. Posted 2021. Accessed April 30, 2025. https://youtu.be/b6oekXIQ-LM
6. Bach, Joscha. 2024a. "Consciousness, Artificial Intelligence, and the Threat of AI Apocalypse." YouTube video. Posted August 5, 2024. Accessed April 30, 2025. https://youtu.be/XcNlv9gp20o
7. Bach, Joscha. 2024b. "Cyber Animism." YouTube video. Posted 2024. Accessed April 30, 2025. https://youtu.be/YZl4zom3q2g
8. Block, Ned. 1978. "Troubles with Functionalism." In Minnesota Studies in the Philosophy of Science, vol. 9, 261–325. Accessed April 30, 2025. https://web.archive.org/web/20110927150409/http://w3.uniroma1.it/cordeschi/Articoli/block.htm

9. Chalmers, David J. 1995. “Facing Up to the Problem of Consciousness.” Journal of Consciousness Studies 2 (3): 200–219.
10. Chalmers, David J.. 1996. “The Conscious Mind: In Search of a Fundamental Theory.” New York: Oxford University Press.
11. Cucu, Alin. 2025. “The Many Inadequate Justifications of the Causal Closure Principle.” Unpublished manuscript.
12. Descartes, René. 1641 [1996]. “Meditations on First Philosophy.” Translated by John Cottingham. Cambridge: Cambridge University Press.
13. Eghbal A. Hosseini et al. 2024. “Artificial Neural-Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training.” Neurobiology of Language 5 (1): 43–63. https://doi.org/10.1162/nol_a_00137
14. Haugeland, John. 1980. “Programs, Causal Powers, and Intentionality.” Behavioral and Brain Sciences 3 (3): 432–433.
15. Hurley, Susan. 2001. “ Perception and action: Alternative views”. Synthese 129 (1): 3-40. https://philpapers.org/rec/HURPAA
16. Jackson, Frank. 1982. “Epiphenomenal Qualia.” The Philosophical Quarterly 32 (April): 127–136.
17. Kubilius, Jonas et al. 2019. “Brain-like Object Recognition with High-Performing Shallow Recurrent ANNs.” In Advances in Neural Information Processing Systems 32: 12805–12816.
18. Kell, Alexander J. E. et al. 2018. “A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy.” Neuron 98 (3): 630–644.e16.
19. Kim, Jaegwon. 1989. “The Myth of Nonreductive Materialism.” Proceedings and Addresses of the American Philosophical Association 63 (3): 31–47. Accessed April 24, 2025. http://links.jstor.org/sici?sici=0065-972X(198911)63:3
20. Kleiner, Johannes. 2024. “Consciousness qua Mortal Computation.” arXiv preprint arXiv:2403.03925. https://arxiv.org/pdf/2403.03925
21. Leibniz, Gottfried W. [1714] 1898. “The Monadology.” Translated by Robert Latta. Oxford: Clarendon Press. Online version: http://home.datacomm.ch/kerguelen/monadology/
22. Mørch, Hedda H. 2023. “Non-Physicalist Theories of Consciousness.” Cambridge: Cambridge University Press.
23. Nagel, Thomas. 1974. “What Is It Like to Be a Bat?” The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450.
24. Papineau, David. 1993. “Philosophical Naturalism.” Oxford: Blackwell.
25. Papineau, David. 2001. “The Rise of Physicalism.” In Physicalism and Its Discontents, edited by Carl Gillett and Barry Loewer, 3–36. Cambridge: Cambridge University Press.
26. Searle, John R. 1980. “Minds, brains, and programs.” Behavioral and Brain Sciences 3(3): 417-424.

27. Seth, Anil K. 2021. “Being You: A New Science of Consciousness.” New York: Dutton.
28. Yamins, Daniel L. K., and James J. DiCarlo. 2016. “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex.” Nature Neuroscience 19: 356–365. https://doi.org/10.1038/nn.4244