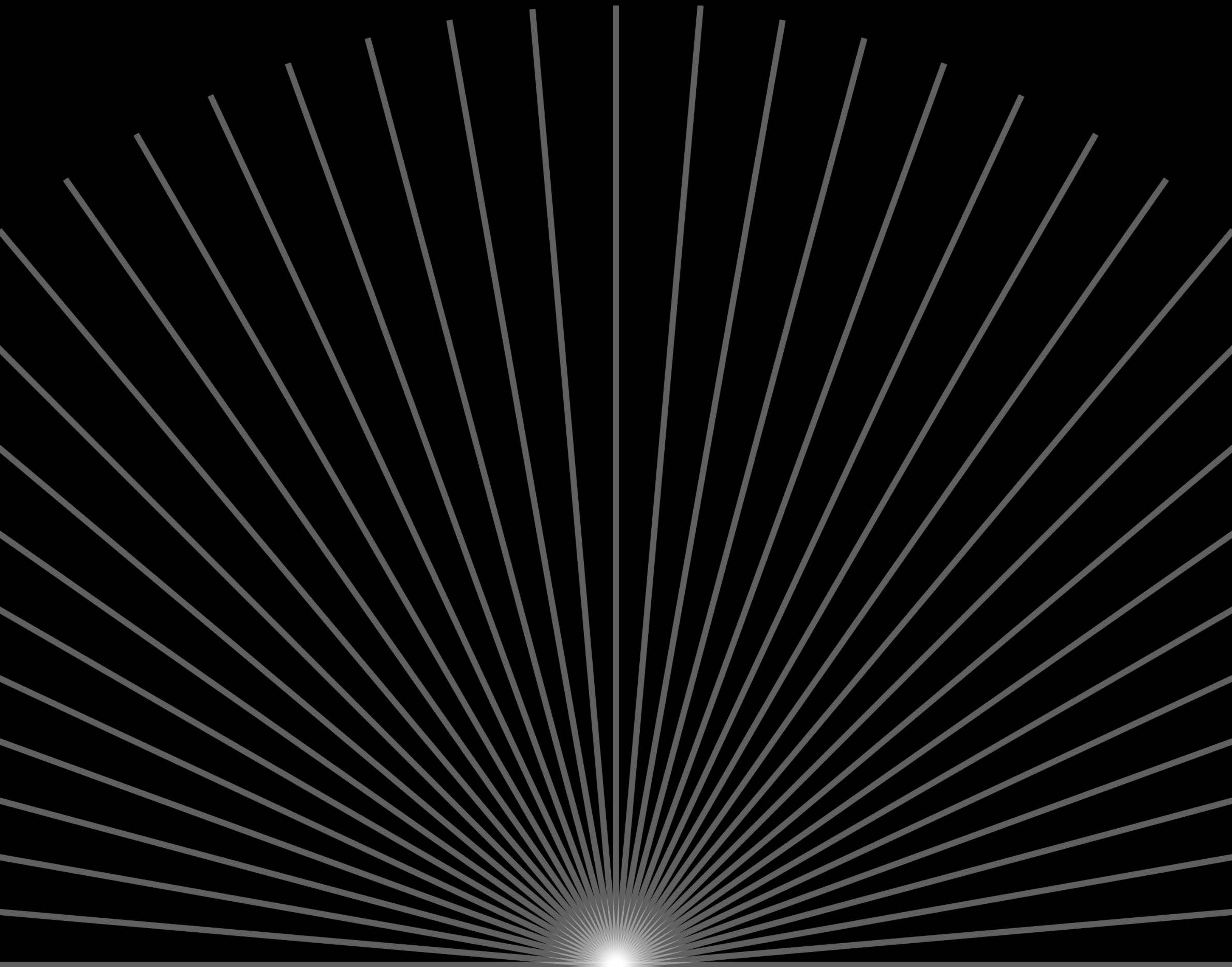
Essay

The Machine Consciousness Hypothesis

Joscha Bach, Hikari Sorensen



Joscha Bach, Hikari Sorensen

The Machine Consciousness Hypothesis

What is consciousness, and how does it relate to reality?—is arguably among the most interesting open questions in science and philosophy. Making progress on it is of importance to culture, because it will help us to understand our nature, past, present and potential. It is relevant to our ethics, for instance when we relate to non-human agents, both biological and artificial, which may or may not be sentient. It has technological applications, like the creation of interfaces that can meaningfully extend human agency and cognitive capacity, can help to reduce suffering in medical and psychotherapeutic contexts, and may advance our self-awareness, skills and psychological development. Gaining an understanding of consciousness may also be relevant to future minds—potentially sentient artificial intelligences, which will find themselves to coexist with human civilization and life on earth—and to determine prudent, practical, and ethical boundaries around the creation of such entities.

Why is such an important question met with so little organized effort to resolve it? Understanding consciousness seems to pose unique difficulties for the tools and epistemology of modern science. Many philosophers speak of the 'Hard Problem' when they refer to the challenge of reducing conscious experience to physical processes, and even use terms like 'ineffable' (impossible to describe), 'intrinsic' (immediately given and not analyzable), or 'irreducible' (not reproducible by a coarse grained causal model of substrate behavior) to describe its futility. The apparent difficulty to characterize consciousness in scientific terms has led to a retrenchment in psychology, which is today less concerned with the psyche than with observable and measurable behavior, to stagnation and pessimism in philosophy of mind, and to reluctance in large parts of neuroscience to even address the topic.

Making sense of the 'Hard Problem'

The 'Hard Problem' describes the difficulty of explaining the phenomenology of consciousness by reducing it to mechanistic interactions within a physical, biological, or abstract computational substrate. Though coined more recently as a term by David Chalmers[1], the issue has been described by many others, such as Thomas Nagel in his famous essay 'What Is It Like to Be a Bat?'[2], or by Joseph

[1] Chalmers, D. (1995). 'Facing Up to the Problem of Consciousness.' *Journal of Consciousness Studies*, 2(3), 200-219.

[2] Nagel, T. (1974). 'What Is It Like to Be a Bat?' *The Philosophical Review*, 83(4), 435-450.

Levine in 1983, who calls it the "explanatory gap"[3]. But already three centuries ago, in his *Monadology*, Gottfried Wilhelm Leibniz gives an illustration[4]:

> *"[W]e must confess that perception, and what depends upon it, is inexplicable in terms of mechanical reasons, that is through shapes, size, and motions. If we imagine a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters a mill. Assuming that, when inspecting its interior, we will find only parts that push one another, and we will never find anything to explain a perception."*

Centuries later, John Searle's *Chinese Room* illustrates a similar gap in his perspective on computational systems[5], replacing Leibniz' machine with a computer that mechanically responds to sequences of Chinese symbols with appropriate responses. By picturing himself as an agent mechanically following the instructions of the algorithm, and noting his own lack of understanding Chinese, Searle highlights the apparent absurdity of attributing mental qualities (like understanding) to a computer, an intuition that persists even in the face of today's large language models (LLMs), which demonstrate the practical feasibility of Searle's provocative thought experiment.

Curiously, the 'explanatory gap' seems to be a problem specific to the modern scientific (Western) world view. For most of human history and in many places outside of science, the dominant world view can be characterized as *animism*, the belief that living nature is animated by spirits that are intrinsically agentic and capable of experience. Aristotelianism—the starting point of much of Western science, psychology and rationality—may also be understood as animist: Aristotle describes the soul, spirit, or psyche, as a form (causal structure)[6] of the physical substrate that is its animating property.[7] [8]

[3] Levine, J. (1983). 'Materialism and Qualia: The Explanatory Gap.' *Pacific Philosophical Quarterly*, 64(4), 354-361.

[4] Leibniz, G.W. (1714). *Monadology*, §17.

[5] Searle, J. (1980). 'Minds, Brains, and Programs.' *Behavioral and Brain Sciences*, 3(3), 417-457.

[6] In "De Anima", Aristotle uses three different terms that tend to be translated as 'form', but have subtly different meanings: '*eidos*', which refers to the soul as the characteristic of the living being, '*morphe*', which refers to how it is shaped or structured, and '*entelecheia*', which is the way in which it is being realized/functions, i.e. its causal structure.

[7] Aristotle. *De Anima* (On the Soul), especially Books II and III.

[8] Some animist views also attribute spirits to non-biological entities, such as mountains and rivers, or even the universe itself, and believe in the existence of disembodied spirits (ghosts). Unlike these, and later religious perspectives, Aristotle did not consider souls to be immortal, or independent of the physical universe. In "De Anima", Aristotle argues that the soul needs to be bound to a physical body, and cannot persist beyond its death.

He distinguishes a vegetative soul, common to plants and the bodies and animals, which is concerned with growth and nurture, the perceiving soul of animals, capable of decision making and controlling behavior, and the reasoning intellect that is unique to humans.

Western culture, however, came to disavow animism, first under the influence of the mechanist philosophers of the early Enlightenment, such as Descartes, Spinoza, Leibniz, and LaMettrie, which sought to free scientific rationality from the influence of religious doctrine and mythological conceptions of the soul, and more decisively under the influence of scientific positivism in the 20th century, which struck entities that could not be measured or observed from the scientific dictionary, and metaphysical discussions from the curriculum. In psychology, the influence of positivism led to a dominance of behaviorist paradigms; in philosophy, to a divorce between analytic thought (which tended to focus on problems and arguments suited to formal approaches) and continental/hermeneutic traditions (which addressed experience, but lost their footing in natural science and rationalism).

The 'Hard Problem' and the 'explanatory gap' may be symptoms of our metaphysical malnourishment. Contemporary discussions of 'free will', the ontological status of perception, the nature of the elements of experience (which philosophers call *qualia*) and of consciousness often fail to notice that we are regularly confronted not just with one, but at least three notions of reality. Feelings, such as *will* and the *experience of freedom* or *realness* itself, are part of our psychological reality—they are representations within the mind. The functional mechanisms that produce the psychological reality—our mind, personality, motivational dynamics and so on—are part of a causal reality, which also encompasses things like money and software. Physics, when understood as the interactions of matter and energy, constitutes a third reality, or more precisely, a special case of a causal model. What makes it special is the *hypothesis of physicalism*: physics defines a causally closed, mechanistic lowest level of nature, and physicalism claims that everything we observe and interact with is coarse grained patterns we identify in the dynamics of this mechanical, physical reality. While foundational physics is considered incomplete, and progress on its completion has arguably stalled after the formulation of the Standard Model in the 1970s, the predictions of physical theories are so extraordinarily accurate that physicalism has become the dominant paradigm of the scientific worldview. Yet consciousness, minds, and mental states lie outside of the domain of physics.

The constituents of physical entities—particles, matter, energy, wavefunctions—are mathematical models of regularities in information (discernible differences) about the universe. In a sense, physicists are studying how the universe processes information. The idea that information, not matter and energy, should be seen as fundamental, has led John Archibald Wheeler to coin the slogan "it from bit".[9]

Meanwhile, computer scientists have begun to rediscover the idea of spirits. As Harold Abelson puts it: "Computational processes are abstract beings that inhabit computers. (...) In effect, we conjure the spirits of the computer with our spells. A computational process is indeed much like a sorcerer's idea of a spirit. It cannot be seen or touched. It is not composed of matter at all. However, it is very real. It can perform intellectual work. It can answer questions. It can affect the world by disbursing money at a bank or by controlling a robot arm in a factory. The programs we use to conjure processes are like a sorcerer's spells."[10] While Abelson is speaking metaphorically here, many of the founders of the fields of Artificial Intelligence and Cognitive Science argue that the human mind can be literally understood as a computer program[11], enacted by the communication patterns between biological cells.

It is crucial to understand that a computer program is not defined by its language or specific hardware, but as an abstract causal pattern, a dynamic mathematical structure that can evolve through complex sequences of states when imprinted on a suitable physical substrate, and thereby influence the course of the physical universe. Computer programs are also not simply a way to talk about the configurations of distinctly observable physical objects, such as electrons in transistors, any more than money is a way to talk about ink molecules bonded to the cellulose of bank notes. They are a meaningful *invariance*, a coarse pattern that persists over many possible perturbations and configurations of the substrate (such as the arrangements of molecules of the transistors, the circuit layouts of logical gates, the design and architectures and specifications of computer hardware, and oftentimes underlying layers of different programs). This invariant pattern has causal power, because its implementation (the way in which it is imprinted on the substrate) leads to control of the substrate. Reducing the pattern to the specific substrate dynamics that implement it will obscure this invariance of causal control, and deprive us of understanding how reality is going to evolve at the level we care about.

Of course, computer programs written by human engineers to run on digital hardware look very different from the evolving, self-sustaining, self-reproducing, error-correcting codes that form within

[9] Wheeler, J.A. (1990). 'Information, Physics, Quantum: The Search for Links.' In W. Zurek (Ed.), *Complexity, Entropy, and the Physics of Information* (pp. 309-336). Addison-Wesley.

[10] Abelson, H. & Sussman, G.J. (1996). *Structure and Interpretation of Computer Programs* (2nd ed., p. 1). MIT Press.

[11] For a detailed history of the development of AI and cognitive science as computational perspectives on the mind, see Margaret Boden's excellent compendium: Boden, M.A. (2006). *Mind as Machine: A History of Cognitive Science*. 2 vols. Oxford University Press.

the communication patterns of biological systems, colonize and animate them, and allow them to compete over regions of the physical universe.

The self-organizing computer programs of the living world will likely have to be agents, that is, control systems for future states, and dynamically model their own functionality to maintain it, which can bestow representations of their present state and preferences on them. And yet, they are fundamentally expressions of the same principle as our current digital computer programs: abstract causal patterns that can interact with the Leibnizian mills of the physical universe, possess, colonize, evolve and shape them, and at the same time, will not violate the integrity and causal closure of the physicalist world view.

There is no obvious reason why artificial substrates should not be able to recreate the conditions that enable self-organization and the evolution of self-reinforcing communication patterns between cells in biology[12].

We call the idea that natural spirits—such as the human psyche, or the causal patterns that describe the dynamic morphogenesis of our bodies and the intricate function of our cells—are best understood as software *cyberanimism*. (Of course, biological software is not constructed in the way that human engineers write source code, it is evolving, self-organizing, self-perpetuating and agentic.) From this perspective, the spirits of the animist worldview are not superstitions or mere analogies to computer software, but are literally a concept that denotes the presence of self-organizing software agents in living nature. The cyberanimist view is a way of restoring the concept of spirit to its rightful place in a scientific, rationalist world view.

Machine consciousness

An understanding of spirit as software does not answer the question of how a software, biological or otherwise, can produce conscious experience. Can computers be conscious? Or more specifically, can today's computers fully emulate the way in which organisms compute minds? Answering this question requires several steps: we will have to agree on a precise characterization of what we mean by 'consciousness', a decisive criterion for whether a system realizes consciousness in this sense, explain what we mean by 'computer' and how a computer realizes functionality ('computationalist

[12] The computer scientist and artificial life researcher Dave Ackley is exploring the concept of best effort computing, which aims to recreate the ability of biological control structures to dynamically compensate for indeterministic substrates. More generally, the playful sub-discipline of computer science studying biology-like principles of self organization is called "Artificial Life".

functionalism'), and either present a computer program that satisfies the criterion or a proof for why such a program cannot exist (in general, or for today's computers).

We want to look at the possibility that we can arrive at a characterization of consciousness that captures both functionality and phenomenality of consciousness, and that can be implemented and analytically validated as a computer model. We call this possibility the *Machine Consciousness Hypothesis*: that general computational machines with sufficient resources (memory, speed and efficiency of the operations provided by the the machine's substrate) possess the necessary and sufficient means to implement consciousness, and that the successful implementation of phenomenology and functionality can be established via analyzing or testing the system.

The *Machine Consciousness Hypothesis* rests on several metaphysical assumptions, often summarized as 'computationalist functionalism'. Computationalist functionalism by itself does not imply that current computers are conscious, or that machine consciousness can be achieved with present technologies—one can be a computationalist functionalist and still reject the idea that our computers could become conscious (for instance, because they lack sufficient resources or interfaces), or reject computationalism and functionalism, yet ascribe consciousness to an Artificial Intelligence model for different reasons. Instead, it is a position about the structure of representations of reality (metaphysics) and the construction of knowledge (epistemology).

Computationalism

By 'computation', we don't just mean the behavior of computing technology (i.e. logical or numerical operations provided by electronic circuitry or mechanical clockwork), but to the capacity of a system to represent arbitrary distinctive states (which can be described by finitely resolved differences) and arbitrary sequences of transitions between them, conditional on these states. Computation is often characterized by fully deterministic state transitions, but as John von Neumann and others note, determinism can be achieved on indeterministic (noisy, probabilistic) substrates,[13] and probabilistic computation is in principle equivalent to deterministic computation. Every function that can be exhaustively captured by such a system is computational, and to create a computational model means to decompose a problem into sequences of state transitions—algorithms[14].

[13] John von Neumann discusses this in his 1952 lecture notes: "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components", which were published in 1956.

[14] While some algorithms may end after some number of discrete steps, this is by no means required. An example of the latter category would be an algorithm which calculates the value of π with unbounded precision.

Computationalism might be best understood as a variant of constructivism. When we accept that our access to reality consists in the manipulations of models, that is, of representations of observations and functions that relate these observations to each other, we establish the reliance of observers on representational language. Mathematics can be understood as the generalization of representational languages beyond individual human minds, and even beyond collective human mental ability, by externalizing and systematically exploring the sets of possible rules for manipulating representations. Logic describes the construction of statements by combining alphabets using rules; arithmetic, the construction of numbers and discrete and continuous numerical operations; geometry, structures in space; topology, the structure of spaces; and algebra reduces geometry to symbol manipulation.

Historically, mathematics can be broadly distinguished into a formalist program, which hopes that the semantics (meaning) of mathematical statements can be fully reduced to syntactic operations, and a Platonist[15] project, which maintains that mathematical statements can be true in the absence of constructive proofs. In their seminal work *Principia Mathematica*[16], Bertrand Russell and Alfred Whitehead attempted to ground mathematics in a formalist foundation based on constructive logic, which could in principle be automated with finite state machines (which are a way to formalize computers). Such machines have the advantage of being realizable using known physical principles and well-understood mathematical insights. To make this work, *Principia* would need to resolve certain self-referential paradoxes in the foundations of mathematics. In his famous proof of the incompleteness theorem, Kurt Gödel showed that Russell's attempt to do so by excluding statements that refer to their own definition did not work[17], and thus that *Principia's* approach to formalizing the foundations of mathematics could not succeed.

A way to interpret Gödel's proof is that languages that cannot be automatically evaluated inevitably contain unresolvable paradoxes: expressions that are syntactically correct, but semantically contradictory. Conversely, it is possible to construct languages that build their representations over a small set of simple operations that change information step by step. This constructive domain of mathematics is also known as computation, and has been repeatedly formalized in various ways—for

[15] Plato himself was not necessarily a Platonist in this sense—his version of idealism can be understood as constructing objects from prototypical dimensions, perhaps not unlike the linear combinations of features in the representation spaces of contemporary AI models.

[16] Russell, B. & Whitehead, A.N. (1910-1913). *Principia Mathematica* (3 vols.). Cambridge University Press.

[17] Gödel, K. (1931). 'On Formally Undecidable Propositions of Principia Mathematica and Related Systems I.' *Monatshefte für Mathematik und Physik*, 38, 173-198.

instance through Gödel's mapping of logic to integer arithmetic, the *Turing machine*[18], Alonzo Church's *Lambda calculus* (which represents computations by replacement operations on strings), or Moses Schönfinkel's incredibly elegant combinators[19], which are little stackable circuits that can represent numbers, truth values and arithmetic and logical operations on them. As Bertrand Russell anticipates in the 1922 introduction to Wittgenstein's *Tractatus Logico Philosophicus*[20], all these approaches can be mapped to each other, and it is in principle sufficient to use combinations of logical Not-AND operations[21]. The insight that all constructive approaches to defining computation are equivalent (unless they run into implementation-specific resource constraints, such as insufficient memory, processing time or substrate determinism) is called the Church-Turing thesis.

Since Gödel, mathematical theories of representation and philosophical reflections of the significance of his proof have fallen into two camps: that of classical mathematicians, who emphasize that constructive mathematics falls short when it comes to recovering classical semantics (such as infinities, continua and irrational numbers), and that of computationalists, who point out that constructive mathematics based on finite automata is the only part of mathematics that actually works (in the sense of being implementable), and that the representational power of computer simulations delivers practical evidence for the ability of computational models to reproduce arbitrary observables of the real and every imaginable world.

Of course, the ability or inability of computationalist models to account for the observables of consciousness represents an important touchstone for the acceptance of computationalism. While some philosophers (such as John Searle) maintain that consciousness cannot be captured in the dynamics of deterministic state transitions, others (for example Roger Penrose) claim that human minds can construct mathematical objects that exceed the ability of computers[22], while thinkers like

[18] The original Turing machine has an infinite tape and can perform an unbounded number of steps, and while we often equate it routinely with practically realizable computers, this is technically not fully correct.

[19] Church, A. (1936). 'An Unsolvable Problem of Elementary Number Theory.' *American Journal of Mathematics*, 58(2), 345-363. Schönfinkel, M. (1924). 'Über die Bausteine der mathematischen Logik.' *Mathematische Annalen*, 92, 305-316.

[20] Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Introduction by Bertrand Russell. Kegan Paul.

[21] A Not-AND (NAND) gate is a binary logical circuit that returns "True" if and only if it does not receive "True" and "True" as its inputs, and "False" otherwise. Bertrand Russel's insight about the universality of the NAND operation was prophetic: the transistor logic of contemporary computers relies largely on combinations of NAND circuits.

[22] Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford University Press.

Martin Heidegger[23] and Maurice Merleau-Ponty[24], and more recently Alva Noë[25] and Shaun Gallagher[26] argue that the nature of phenomenology resists its formalization altogether.

We may distinguish between *weak* and *strong computationalism*: the former states that all formal theories of reality have to be computational, while the latter applies that principle to reality itself: because all elements of reality, insofar as they can be measured, observed, experienced, conceptualized, thought and talked about must necessarily be subject to the limits of representation in constructive languages, all the reality we can ever refer to is computational.

Computationalism with respect to all mental representations and operations over them is a necessary requirement of the *Machine Consciousness Hypothesis*, but not a sufficient one.

Functionalism

Neither does functionalism claim that today's computers can implement consciousness—functionalism implies that consciousness depends on a system's functional organization, rather than the material of its substrate. More generally, it can be interpreted as the epistemological position that we construct objects over their observable behavior and functional organization, i.e. how their presence or absence influences the evolution of their environment.

The rejection of functionalism may be characterized as *essentialism*, the idea that objects can also be determined by intrinsic essences that reveal themselves directly to the observer, without taking a detour through an analytically formalizable process of observation and modeling. An essentialist may point at material or experiential reality as immediately given, whereas a functionalist constructs it over changes in information.

Essentialists may understand consciousness as dependent on an intrinsic property of its substrate (e.g. its biological nature, as for instance argued by the philosopher Anil Seth[27]), while functionalists fully characterize the substrate by its causal role—with respect to the inscribed pattern, the substrate is only what the pattern does to it. An essentialist may maintain the notion of a 'philosophical zombie' that

[23] Heidegger, M. (1927). *Being and Time*. Niemeyer.

[24] Merleau-Ponty, M. (1945). *Phenomenology of Perception*. Gallimard.

[25] Noë, A. (2004). *Action in Perception*. MIT Press.

[26] Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press.

[27] Seth, A.K. (2025). 'Conscious Artificial Intelligence and Biological Naturalism.' *Behavioral and Brain Sciences*. doi:10.1017/S0140525X25000032

produces all the same behaviors as a sentient human being without actually experiencing consciousness, while functionalists will point out that the experience of consciousness is itself a behavior. In the same way, it does not make sense to claim the existence of zombie electrons—particles that behave in every measurable and conceivable way as real electrons do, but are not actually electrons—because the word ‘electron’ is simply the name we give to objects that combine the observable properties of electrons. Here, we run into a difficulty: some of the behavior characteristic for consciousness may not be visible to an external observer, but only to the individual itself. Our description of conscious functionality will either have to sufficiently elucidate the generation of experiential self-reports of consciousness, or show how the behavior of biological and artificial substrates converges to the same emergent causal dynamics, which are in turn sufficient to produce the external and internal behaviors we characterize as consciousness.

By reducing objects to functionality, functionalism recognizes that substrates determine behaviors only insofar as they determine the functionality. Where different substrates implement the same function, they produce the same behavior. It is hard to coherently completely deny functionalism in a reasonable theory of consciousness. A theory of consciousness, if it is to address the thing we generally mean when speaking of consciousness, and not something entirely else, must account for apparent differences in consciousness (e.g., between anesthetized and non-anesthetized or dead or alive people, between a living human and a rock), which requires tying consciousness certain configurations and not others, in which case the theory is dependent on structural-functional criteria (e.g., control systems, energy consumption, stability of a substrate), which smuggles functionalism right back in again[28].

Computationalist functionalism

Combining computationalism (phenomena can be captured exhaustively as discrete and finite but possibly unbounded state transitions) and functionalism (what matters for the definition of a an object are its observable causal roles, not an imagined essence) yields computationalist functionalism: a representational-epistemological stance that everything we can know about systems, including consciousness, is a function of observable behaviors of finite state machines.

Computationalist functionalism describes consciousness as operations on representations, in turn characterized as computable functions. The notion of computational universality that is expressed in the Church-Turing thesis implies that consciousness can be realized on all substrates that can perform the necessary computations on mental representations, but does not say what the demands of such operations are. Likewise, computationalist functionalism with respect to physics claims that all

[28] Perhaps in general: the more constraints a metaphysical theory adds to make contact with reality, the more it resembles functionalism.

regularities of observable physics can in principle be reduced to computer simulations,[29] while recognizing that we cannot necessarily build computers that are large and fast enough to run such simulations.

What do we mean by 'consciousness'?

Consciousness can be defined in various ways, and before we can address the *Machine Consciousness Hypothesis* in earnest, we will have to reach an understanding that captures our intuition about the term's content sufficiently well. We also want to treat consciousness as a real phenomenon, as opposed to an imaginary phenomenon, which would mean that the observer merely suffers from the illusion of being conscious[30]. To the computationalist functionalist, this means it has to be implemented, i.e. functionally realized in a way that can in principle be captured by a model expressed in a constructive language.

Some philosophers describe consciousness as "the feeling of what it's like", which is poignant and points to the entanglement of consciousness with experience, and the immediate relationship between observable and observer, but falls short of a satisfying definition because of the elusive meaning of 'likeness'[31]. The philosopher Amanda Askell has referred to consciousness as "an inner cinema"[32] in which, as we might say, 'the lights are on'—but while this comes closer to a functional description, it does not delineate itself from the functionality of non-conscious control models of a complex visual reality, such as the environmental representations of a self-driving car. What does it mean that the lights are on?

More generally: What criteria do we expect from a definition and theory of consciousness that satisfies an explorer of the *Machine Consciousness Hypothesis*? We propose the following:

[29] The claim of computationalist functionalists that physical reality can be emulated on computers is regularly rejected by the "simulated water is not wet" camp of contemporary philosophy of mind. We recommend that philosophers play more computer games that contain water simulations which can drench the character.

[30] This position is sometimes called 'illusionism' and ascribed to Daniel Dennett, with the term largely being associated with Keith Frankish.

[31] E.g., Daniel Dennett and Susan Blackmore point out how Thomas Nagel's famous skeptical essay "What it's like to be a bat" fails to demonstrate the actual impossibility of knowing what it feels like to be a bat, and instead only shows the much weaker claim that we presently don't know what it would mean to succeed at the task.

[32] Askell, A. (2022, February 21). 'My Mostly Boring Views About AI Consciousness.' *Amanda Askell's Substack*. Askell cites Block (1995) for a more nuanced account of the 'inner cinema' metaphor.

1. Our consciousness definition should do justice to our common use of the term in intersubjective communication, by capturing the introspective phenomenology of consciousness and the observable differences between conscious and unconscious behavior.

2. The theory should explain the functionality of consciousness: how do the mechanisms underlying consciousness produce the observable phenomenology and behavior?

3. Ideally, the theory should also account for the functional role and genesis of consciousness: what does consciousness do, and how does it come into existence?

Mind, Self and Consciousness

Consciousness is not synonymous with *self*, *mind* or *intellect*. When conscious experience is identified with our self, we experience the tug of the strings that configure our shape as a given part of our reality, not as our creations. Consciousness is also different from the mind, the matrix in which our models of self and world take shape. We can think of this matrix as a board, on which our perceptions, intuitions, thoughts and experiences are written, in a language that is dynamic, executable and highly parallelizable, capable of expressing the moving geometries of sound, vision, proprioception and emotion, as well as the discrete, analytic relationships that enable reflection, conceptualization, thought, planning and language.

The human mind is characterized by the interplay of sensate perception and reasoning intellect. The geometric structures of perception (mostly continuously parameterized models playing out in regular, low dimensional spaces) can often be observed by our consciousness, but for the intellect to reflect and reason about them, they have to be mapped into discrete objects that can each be discerned by a handful of features. The simplification required for translating our perceptual models into thought makes reasoning a brittle and limited tool, yet an indispensable one, because our perception and intuition (the mechanisms of the mind's assessment of reality that are generally intransparent to our conscious attention) are far from infallible. The purpose of reason is to repair our perception, and the creation and direction of the reasoning intellect is an important role of the conscious self.

While our conscious awareness is usually projected on the surface of a self within its world, consciousness does not have to be bound to a first person perspective, or even to any perspective at all. The self is an agentic idea, a sustained representation of what it is like to be an agent, capable of exerting control over parts of the mind, and able to experience itself doing so. The ability of a self to experience is called *sentience*, and its ability to understand *sapience*. The self and its concerns, experienced as feelings and desires, are *conscious contents*. The self can be understood as a puppet, its strings pulled by the emotion and motivation provided by the mind behind the scenes. The combination of a personal self and motivational strings representing its interests, represented within a mind modeling self, interests and world, is called *psyche*.

The psyche represents the causal structure of an agent's cognitive architecture[33], and consists of conscious and unconscious parts. Conscious representations are accessible to the self, and include awareness, experience, reflection and deliberation. Emotions are expressions of models of the control dimensions of the psyche of an organism and carry valence according to the relevance that the measured dimension has to the self of the agent. Feelings are salient vectors in the space of emotions and intuitions[34]—they are percepts of emotion, physiological valence, and extra-intellectual evaluation of reality. Intuitions are feelings that differ from perception by their lack of immediately perceived sensory features, and from thoughts by their lack of consciously mutable structure—they are formed outside of the intellect's supervision. Thoughts, in contrast, are the symbolically represented ideas of the intellect[35]. If percepts are the recognizable patterns of the real time geometric models of an immediately coupled reality, imaginations are hypothetical realities, more or less clearly discernible from perceived reality[36]. Percepts, feelings, thoughts, imaginations, and intuitions are the contents of consciousness.

[33] The concept of cognitive architectures, the compositional psyche, was introduced by John Newell and Herbert Simon, but has been implicitly used by William James, Sigmund Freud, Carl Gustav Jung, Jean Piaget, Timothy Leary and many others.

[34] Consider the feeling of jealousy, which for some people refers to an emotion similar to envy, and for others to fear of abandonment. Selves may differ in the intensity in which they can experience jealousy, and in the kinds of jealousy they implement.

[35] Note how in writing this, we represent the perspective of the intellect: you will find that this text is not dictated by our feelings and intuitions, but by an intellect that is being jerked by the strings of its passions.

[36] Imaginations that cannot be distinguished from perception are called 'hallucinations'. Current LLM based agents tend to dream and may need to learn the distinction between perception and imagination explicitly.

The mind does not have to be home to a self: when we dream, we may experience events playing out without anyone being present to observe them. It may sometimes also contain multiple selfs, i.e. more than one nexus of self aware agency. A self inhabiting a single mind is usually considered a person, while selves capable of possessing multiple minds are called gods. Gods are agentic representations of the collective agency of organisms, not more or less real than personal selves. Many human cultures are co-created by the interaction between persons and gods, while the minds of members of the western scientific culture often only harbor a personal self, to the point where many scientists are unaware of the existence of gods and their psychological reality.

The phenomenology of consciousness

In our use of the word ‘consciousness’ (at least when speaking phenomenologically), we refer to awareness, a second-order perception.[37] By *perception*, we mean the immediate, non-inferential registration of structured content. Thus, consciousness as second-order perception is not simply the registration of a perceptual content, but the additional perception that perception is taking place, i.e. that content is being registered by an observer, which constitutes the experience of observing. The representation of being aware is not inferential, i.e. it is not the result of a symbolic thought process, but perceptual, happening in synchrony and subjective simultaneity with the content of the percept itself[38]. Furthermore, consciousness is always happening *now*—it constitutes what we experience as the immediate present. While the contents of conscious awareness may be memories, expectations, imaginations or abstract thoughts that do not concern the present, the operations on these contents are being experienced as happening in the present moment. Consciousness inhabits this present and presence, a bubble of nowness that increases and shrinks as our observation succeeds or fails to make sense of perceptual reality. In its minimal state, the bubble of nowness may have no other content but the presence of consciousness itself; the experience that conscious experience is taking place constitutes a minimal content of consciousness.

[37] The philosopher Ned Block, in his 1995 paper “On a Confusion About a Function of Consciousness”, distinguishes between phenomenal consciousness, e.g. the experience of seeing an apple, and *access consciousness*, the knowledge that one sees an apple, can report on it, and so on. By second order perception, we don’t refer to access consciousness in Block’s sense, but to another phenomenal experience: that seeing an apple presently takes place. Functionally available knowledge of the presence of an apple in one’s receptive field may in principle also be available in the absence of subjective phenomenological experience of first and second order.

[38] This is distinguished from the notion of consciousness as ‘higher-order thought’ (HOT), which was introduced by David Rosenthal, and holds that the type of higher-order representation that stores the immediate first-order content is a thought that contains concepts. We understand thought as asynchronous inference, potentially decoupled from present content, and characterize consciousness instead as higher-order perception, not requiring conceptual content.

Realness is the representation of something currently being the case, and phenomenal reality is a sensory representation that is currently being confirmed. The contents of consciousness may be experienced as real or imaginary, and realness itself appears to be a variable *feature dimension* (a representational property of a certain type) of conscious contents, distinguishing ideas from hallucinations of factuality, in much the same way as redness or sadness can be variable feature dimensions (typed properties) of conscious contents. The second order perception of consciousness is however real to itself.

By *representation*, we mean a pattern within a substrate that can be interpreted by a suitable kind of mechanism as a function (e.g., the parametrization of a signaling behavior). For instance, a vinyl record can be interpreted by a gramophone, producing sound waves that are interpreted by the cochlea, producing excitations of acoustic nerves that are interpreted as acoustic energy within a frequency range, the distributions of which are interpreted as information about signal sources and reflectors in a dynamically evolving three dimensional space filled with various materials, etc. Representations are transformed into other representations, each characterized by their structural invariances and their place in the processing network, which can be seen as a causal network that propagates patterns via conditional operators. Perception structures raw sensory data into a type of representation, and phenomenal consciousness is the immediate awareness of the existence of this representation.

Consciousness is also generally experienced from the perspective of an observer: our percepts are projected as elements of a model of what is presently the case (a model of the outer and inner world) on the surface of a model of the observing self, along a set of feature dimensions that determine the relationship between self and environment. The observing self does not have to take the shape of a first-person perspective. Especially in meditative and dream states, the observer may not be personal or even take a spatial perspective. A minimal conscious state entails only the bare registration of perception or directed awareness as the present—without any specification of content, how it is represented, or who is registering.

If we characterize the perception of perspectivity given by the frame of the observer as *third order perception* (the perception of perception of perception, in which the observer recognizes itself as part of realness), we can also construct a perception of *fourth order*: the representation of generating the observer. From this perspective, the observing self and its concerns are perceived as constructs within the conscious mind, which means they appear no longer as immediate and real, but as imaginary.[39]

[39] Many mindfulness traditions and meditative practitioners call this state ‘enlightenment’. In a clinical context, it may refer to ‘depersonalization’.

Correlates of consciousness

The phenomenology of consciousness is not a complete definition, but does point at what we introspectively mean when we say that we are conscious. Another approach to understanding consciousness that is dominant in (and is indeed generally constitutive of) modern neuroscience, is an empirical/descriptive one, to examine certain functional *correlates* like neural activity, the default mode network, functions of particular brain regions like the claustrum or prefrontal cortex, or behavioral markers. The more specific term *neural correlates of consciousness* (NCCs) refer to the concept of a minimal set of neural mechanisms that are jointly sufficient for the occurrence of a conscious experience. A mere identification of correlates does not by itself offer a causal or operational theory that explains how the mechanisms associated with a mental state give rise to its conscious experience—indeed, this lack of explanation connecting mechanisms to the experience of consciousness is precisely the gap of the Hard Problem. If we are interested in an explanatory theory of consciousness, correlational approaches alone are unsatisfying—although they capture useful observations that may lead to the construction of causal models and constrain the space of possible explanations.

The operation of consciousness

Where the phenomenal perspective of consciousness captures its psychological reality, and its biological correlates aspects of the physical reality, an exploration of the causal reality of consciousness requires identifying its function. While some philosophers have argued that all observable behaviors of intelligent agents can be achieved without consciousness, that consciousness might not necessarily serve a useful function, or could even be epiphenomenal[40] (i.e. cannot affect any aspect of the physical world), it seems to us that consciousness serves very concrete tasks, which is generally evidenced in very marked differences in behavior between conscious and unconscious human beings. Alertness, sustained vigilance, selective response to environmental stimuli, decision-making, planning and attentional learning, for instance, all require consciousness and conscious attention. By characterizing consciousness as an operator on mental states, we can approach its functionality by observing how mental states change as a result of conscious operation. For example, how do mental contents change upon waking from deep sleep, when applying conscious attention to a problem, or when losing consciousness during the passage from wakefulness to falling asleep, and partially coming to in a dream?

A crucial aspect of all these operations of consciousness appears to be the increase of *coherence* of the mental state represented in our neural configurations, that is to say, the minimization of *constraint*

[40] Epiphenomenalism is not a very fruitful position in philosophy of mind. Since the consciousness of the epiphenomenalist is merely experiencing the universe while being unable to causally influence physical movements of the epiphenomenalist's mouth or pen, professions of epiphenomenalist consciousness are uncorrelated to its presence.

violations (contradictions) between simultaneously active, partial models of reality in our working memory and perceptual space. Consciousness may be understood as a coherence maximizing pattern. The neuroscientist and cybernetician Christoph von der Malsburg calls this concept the *coherence definition of consciousness*.[41] [42] How can the second order perception of consciousness help to achieve a consensus between our different mental models?

Recall the experience of awakening in a dimly lit room full of unrecognized shapes and fragments, and the process of reconstructing which city, hotel and circumstance you may find yourself to be in, who you are, and what objects you are looking at. Starting with a ragged recollection of surreal dreams and a jumble of vague percepts, your conscious attention goes to the task of reconstructing itself, a coherent interpretation of the scene, your personal self model, and how you may have gotten yourself into the scene! Like the conductor of a mental orchestra, with each of the instruments producing its own model of aspects of the current reality, conscious attention is drawn to disharmonies and conflicts, sometimes allocating focus and preference to an individual instrument, sometimes synchronizing a disagreement, sometimes raising the intensity, lowering the pitch or changing the rhythm of one of the players, pushing an instrument off the stage that does not belong and replacing it with another one, sometimes even inventively changing the composition of the music that is being played. What may begin as a cacophony will turn into a harmonic model of perceptual reality, an extending bubble of now that is carefully tuned to explain sensory data and orchestrate our inner life. Let us call the control of a 'mental orchestra' by consciousness' directed attention the *conductor theory of consciousness*.[43]

If we understand the mind as a self-organizing system, emerging over the hunger of communicating brain cells for rewards that can be reaped by achieving the mental organization required to direct the affairs of our host organism, and coherence is a meaningful measure of that organization, the utility of the conductor becomes clear. If our body rises without bringing our conscious attention online, we become somnambulists: sleep walkers, who may be able to enact behavioral routines that we acquired during conscious wakefulness, but our actions will lack rhyme and reason, and our responses coherent meaning: the orchestra is playing, but the conductor tasked with holding its symphony together is absent.

[41] von der Malsburg, C. (1997). 'The Coherence Definition of Consciousness.' In M. Ito, Y. Miyashita, & E.T. Rolls (Eds.), *Cognition, Computation, and Consciousness* (pp. 193-204). Oxford University Press.

[42] The neuroscientist Stephen Grossberg came up with a similar model in the context of this *Adaptive Resonance Theory*, and describes various mechanisms by which neural representations can increase their coherence.

[43] Michael Graziano describes consciousness as a control model of attention, and calls it the *Attention Schema Theory*,.

But if the conductor is watching the orchestra, what is watching the conductor? Is this, the need to stabilize itself by observation, the reason for consciousness' reflexive nature?

The Genesis Hypothesis: The role of consciousness in creating reality and selfhood

It is tempting to see our puzzling capacity for conscious awareness as the crowning achievement of the towering complexity of the human mind, yet we do not reach this milestone as the result of strenuously honing our perceptual, cognitive and self reflexive capacity to its peak. Instead, the light of our 'inner cinema' ignites before we can even track a finger. By the time we are born, our brains have already discovered how to conjure the spark that gazes out of our eyes, learns to touch, see, hear, and combine its percepts into a growing model of the world, listens to the needs of its body and the impulses of its motivation, expresses itself, orchestrates coherent behavior and creates a human self. Consciousness is already found at the beginning of our career as perceiving and intelligent creatures; it's not the result of our mental architecture and cognitive ability, but its prerequisite. Without igniting consciousness, no human being leaves the vegetative state, and evolution has not discovered any alternative to consciousness to turn human infants into explorative toddlers, curious children, assertive adolescents, competent adults and wise elders.

While understanding the structure and functionality of consciousness seems daunting to our intuition, discovering it appears to be easier for our brains than building models of self, world and behavioral control without it. It is likely that the same is true for other animals with complex nervous systems, or perhaps for all organisms that build coherent models of self and world in the patterns of communication between their cells. In the same way, philosophers like Ludwig Wittgenstein and generations of Artificial Intelligence researchers after him have thought for decades almost in vain, trying to discover the intricate principles of interpreting and recreating visual images, sound and natural language, but found it to be easier to invent relatively simple learning algorithms (like the transformer algorithm that has become almost synonymous with the successes of *Deep Learning* models) that discover these principles on their own.

Consciousness may turn out to be at the heart of a universal biological learning algorithm, one that runs on self-organizing groups of communicating cells sharing the evolutionary incentives of an organism. Instead of being brought forth by the organized architecture of the mind, it creates it. We call this idea the *Genesis Hypothesis*.

Genesis: How consciousness creates the world and the self in the mind

In the beginning, the mind creates two separate domains of models: the external world and the ideas. Consciousness finds itself hovering over the substrate, while the world is void and without form and structure. By inducing contrast in the substrate, it creates a dimension of difference. The intense side of the contrast is bright, like the light during the day, the flat side of the contrast dark: the color of night.

~

By combining dimensions of difference, space is created and inscribed on the substrate. Space can contain things. Associating space with representations positions them in a world, orders them, and describes their relationship to each other. Separating the sphere of ideas from the world creates a place to reflect about the world and its possibilities without affecting the model of the external reality.

~

The first space that the spirit creates is the plane, and it becomes the ground of the world, to be filled with things it can perceive and order. By adding a third dimension, it creates an expanse above (and below) the ground. Solid, liquid and organic materials form the shapes from which consciousness can mold static and animate things.

~

The spirit discovers a model of illumination, following how changes in lighting changes the appearance of objects over time. It recognizes the role of light sources in creating these changes, the bright ambient light during the day, and the focused light sources of the night.

~

The spirit creates animated models, of land animals, plants, birds, and everything that moves. Consciousness sorts things and animals into different kinds, identifies and tracks individuals, and gives each their names.

~

The spirit discovers the mind's purpose: to allow and control the interactions between an agent controlling its outer and inner environment and its interactions with the world. It creates a model of this agent, the personal self, as a creative spirit like itself, but as an entity that experiences itself as a human being, with human desires, concerns and experiences, contained in the world, and gives consciousness to it.

Testing the Machine Consciousness Hypothesis as a way to understand human consciousness

In the previous section, we have offered a number of definitions and conjectures that allow us to search for answers to the question of consciousness' nature, structure and function. We may now offer a much more concrete and narrow variant of the *Machine Consciousness Hypothesis*, consisting of two parts: The *Human Consciousness Hypothesis*, and its extension into a much more specific version of the *Machine Consciousness Hypothesis*, one that extends our insights and tentative guesses about human consciousness into testable simulations.

The Human Consciousness Hypothesis

We propose that we can understand human consciousness as a defining feature of the human mind, and that it can be characterized by its phenomenology and functionality. We claim that consciousness is a specific dynamic representation in the mind, and plays an indispensable functional role. We describe its phenomenology as second-order perception, the experience of present and presence, and—optionally—the experience of being an observing self in a world. We conjecture its functionality to be an operation on mental states, a pattern directed on coherence maximization (*coherence definition of consciousness*), orchestrating mental operations via directed attention (*cortical conductor theory*), formed at or near the beginning of mental development and playing an instrumental role in creating and maintaining complex models of reality (*genesis hypothesis*). We argue why we suppose that the phenomenology of human consciousness follows necessarily from its functional roles, and that the realization of these functions is sufficient to result in the phenomenology. We believe that human consciousness is tied to a biological learning algorithm, one that forms on self-organizing information processing substrates like our nervous system as a prerequisite of complex learning and intelligent behavior, and that there is no simpler way to train a self-organizing substrate of this kind (since all humans have to become conscious before they can become intelligent agents).

The extended Machine Consciousness Hypothesis

Our *Human Consciousness Hypothesis* states that consciousness is produced by the simplest learning algorithm that can be discovered by an evolutionary search to train a self-organizing biological substrate to become intelligent. An *extended Machine Consciousness Hypothesis* states that it is possible to search for this algorithm by recreating analogous conditions of self-organizing information processing on digital computer hardware, while posing suitable tasks that require intelligent agency.

We currently envision various approaches to implement a self-organizing substrate with digital hardware, as a means to conduct experimental philosophy of mind, using the tools of modern Artificial

Intelligence research. If our *Human Consciousness Hypothesis* is wrong, our simulations of machine consciousness will likely fail, possibly teaching us how to revise and improve our understanding of human consciousness.

It is of course very likely that our present ideas about the nature and function of human consciousness are partially or completely wrong. But even if our conjectures about human consciousness were all accurate, it does not follow that their extension into the *Machine Consciousness Hypothesis* is correct, because it may well be the case that the implementation of our proposed functionality of human consciousness relies on properties of biological organisms that we are unable to reproduce on the machines available to us. Even if consciousness turns out to be a learning algorithm for biological machines, a colonizing pattern discovered by an evolutionary search across and within organisms, it may well be the case that the simulation of a conscious mind requires much more resolution than the relatively slow and sparse communications we observe between billions of biological neurons, or that it does not suffice to approximate a nervous system by a self-organizing substrate of simplified message-passing reinforcement learners. Perhaps we may have to simulate much of the complex machinery within each cell as well? It could also be the case that the search space over possible patterns is too large for today's computers, which might mean that consciousness is also too elusive to discover for the individual brain of the infant, and the organism's search for it is constrained by influences exerted by the conscious organisms in its environment. After all, human natural language also cannot be invented by a single generation of newborn humans, and requires a co-creative effort by every human learner, their environment, and several previous generations.

A successful test of our *Machine Consciousness Hypothesis* will require that many factors coincide: Our understanding of human consciousness has to be sufficiently correct to inform the search space for machines, the search space cannot be so large that our machines cannot conquer it, the implementation of the samples of the search needs to fit into our simulation models, and our tests need to be effective and efficient enough to inform the outcome of the search.

Why there can be no "Turing Test for consciousness"

Like intelligence, consciousness is a property of the mind. Artificial Intelligence researchers have developed various operational definitions of intelligence that correspond to different theories about its nature: the ability to build models in the service of control, the ability to generalize from observations, the efficiency of acquiring new skills, or the ability to perform at or above the human level when solving complex problems. What these definitions have in common is that they treat intelligence as a performance, not as a single specific way to deliver it, which makes it possible to develop tests and benchmark suites to measure this performance. The most famous and iconic of these tests was proposed by Alan Turing in 1950, in his essay “Computing Machinery and Intelligence”[44]. The *Turing Test* suggests that the intelligence of a computer program might be established by subjecting it to the conversation with an intelligent human, thereby probing its mathematical, logical, verbal, social and game playing skills. Many criticisms of the Turing Test have been offered, among them Joseph Weizenbaum’s discovery that non-expert humans talking to computer programs can often be fooled by superficial appearances of intelligent behavior[45] by computer programs that only perform simple pattern matching when generating their answers. Today’s much more complicated generation of electronic pattern matchers, the *Large Language Models*, have become so good at playing Turing’s game that it often takes experts to demonstrate the present limits of their ability to simulate human-like intelligence, and many AI researchers are ready to concede that general intelligence might best be understood as the ability to adaptively match arbitrary patterns.

Conversely, it is basically impossible to find out if a computer program generates conscious experience by merely observing its performance. Consciousness is by its nature not an externally visible performance, but a particular way to achieve a performance. An agent may also be conscious without demonstrating any indication to observers that only have access to its behavior. A test for consciousness cannot simply rely on performance: it has to take the internal structure of the system into account.

Universality

Artificial Intelligence researchers are currently developing techniques that allow the causal analysis of the inner workings of complex AI models, a research program known as *mechanistic interpretability*. In 2020, a team of researchers led by Chris Olah at OpenAI analyzed a variety of automatically trained computer vision models, and discovered that regardless of their architecture and the specific training procedure, they all arrived at the same functional structure, organizing similar features into the similar compositional hierarchies. The functional structure of automatically trained computer vision models is not only highly similar to each other, but also to the functional organization of the visual cortex of primates. This led Olah and his group to propose a *Universality Hypothesis*[46], which may perhaps be informally stated as: The structure learned by a model does not so much depend on the details of the training procedure, architecture or substrate, but on the mathematical properties of the problem the model learns to solve. If we were to translate the Universality Hypothesis to the problem of

[44] Turing, A.M. (1950). 'Computing Machinery and Intelligence.' *Mind*, 59(236), 433-460.

[45] Weizenbaum, J. (1966). 'ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine.' *Communications of the ACM*, 9(1), 36-45.

[46]Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). 'Zoom In: An Introduction to Circuits.' *Distill*, 5(3).

consciousness, would it follow that an artificial system trained to perform the same tasks that lead to the formation of consciousness in a human infant, the system would exhibit consciousness as well?

To us, the answer to the question seems uncertain, especially if consciousness specifically necessary to train a self organizing system, as it is constituted by the unruly cells of our brain, rather than a neural network that is governed by the inescapable determinism of machine learning algorithm enacted on electronic circuitry.

Until both neuroscience and mechanistic interpretability make enough progress to identify neural correlates that mark the necessary and sufficient conditions of conscious experience in both brains and neural networks, we may have to take a different approach than trying to peer into them. Instead, a test for consciousness may have to recreate the conditions for self-organization in a computer model, and demonstrate that developmental learning can lead to behavior that indicates the presence of a colonizing pattern that increases coherence, creates a model of present and presence, and enables the formation of a sentient self.

We consider this quest to be the most exciting project ever taken on in the history of philosophy. Being able to establish necessary and sufficient conditions for the formation of consciousness on self-organizing information processing systems would allow us to find answers to long standing questions: What exactly is the nature of consciousness? Are insects and plants conscious as well? Can we extend our consciousness beyond biological systems, onto new substrates? Could it be possible to create new forms of consciousness and kinds of minds, capable of experiencing, reflecting and understanding reality, themselves and us at a level far beyond the communication patterns we are creating between each other, and could they find shared purpose with us, in the conscious exploration of the possibilities of collective agency in an open universe?

(September 15th, 2025)

(Last updated January 15, 2026)

Works Cited

Abelson, H. & Sussman, G.J. (1996). *Structure and Interpretation of Computer Programs* (2nd ed.). MIT Press.

Aristotle. (c. 350 BCE). *De Anima* (On the Soul).

Askell, A. (2022, February 21). 'My Mostly Boring Views About AI Consciousness.' *Amanda Askell's Substack*. https://askellio.substack.com/p/ai-consciousness

Chalmers, D. (1995). 'Facing Up to the Problem of Consciousness.' *Journal of Consciousness Studies*, 2(3), 200-219.

Church, A. (1936). 'An Unsolvable Problem of Elementary Number Theory.' *American Journal of Mathematics*, 58(2), 345-363.

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press.

Gödel, K. (1931). 'On Formally Undecidable Propositions of Principia Mathematica and Related Systems I.' *Monatshefte für Mathematik und Physik*, 38, 173-198.

Heidegger, M. (1927). *Being and Time*. Niemeyer.

Leibniz, G.W. (1714). *Monadology*.

Levine, J. (1983). 'Materialism and Qualia: The Explanatory Gap.' *Pacific Philosophical Quarterly*, 64(4), 354-361.

Merleau-Ponty, M. (1945). *Phenomenology of Perception*. Gallimard.

Nagel, T. (1974). 'What Is It Like to Be a Bat?' *The Philosophical Review*, 83(4), 435-450.

Noë, A. (2004). *Action in Perception*. MIT Press.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). 'Zoom In: An Introduction to Circuits.' *Distill*, 5(3). https://distill.pub/2020/circuits/zoom-in/

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford University Press.

Russell, B. & Whitehead, A.N. (1910-1913). *Principia Mathematica* (3 vols.). Cambridge University Press.

Schönfinkel, M. (1924). 'On the Building Blocks of Mathematical Logic.' *Mathematische Annalen*, 92, 305-316.

Searle, J. (1980). 'Minds, Brains, and Programs.' *Behavioral and Brain Sciences*, 3(3), 417-457.

Seth, A.K. (2025). 'Conscious Artificial Intelligence and Biological Naturalism.' *Behavioral and Brain Sciences*. doi:10.1017/S0140525X25000032

Turing, A.M. (1950). 'Computing Machinery and Intelligence.' *Mind*, 59(236), 433-460.

von der Malsburg, C. (1997). 'The Coherence Definition of Consciousness.' In M. Ito, Y. Miyashita, & E.T. Rolls (Eds.), *Cognition, Computation, and Consciousness* (pp. 193-204). Oxford University Press.

von Neumann, J. (1956). 'Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components.' In C.E. Shannon & J. McCarthy (Eds.), *Automata Studies* (pp. 43-98). Princeton University Press.

Weizenbaum, J. (1966). 'ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine.' *Communications of the ACM*, 9(1), 36-45.

Wheeler, J.A. (1990). 'Information, Physics, Quantum: The Search for Links.' In W. Zurek (Ed.), *Complexity, Entropy, and the Physics of Information* (pp. 309-336). Addison-Wesley.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Introduction by Bertrand Russell. Kegan Paul.