Easy Consciousness and Its Unpaid Premises

Flyxion
Independent Researcher

November 2026

Abstract

It has been proposed that experience may be a comparatively simple achievement of self-organizing systems, realizable in architectures not much more elaborate than present language models, and that simplicity should therefore be the default expectation. This essay examines one prominent statement of that view, an extended interview with Joscha Bach, and reads it through its transcript, which is more qualified than the accompanying title and description. Bach concedes that the relevant organization may have been discovered over evolutionary time and inherited, that language models may report experience without having it, and that his proposal is an interpretation and not a test. The essay argues that the position nonetheless leaves five connections unspecified: between infant development and consciousness as a learning mechanism, between the simplicity of a procedure and the simplicity of what it yields, between an account of self-modeling and an account of experience, between emulation and the preservation of whatever matters for experience, and between behavioral convergence and shared causal structure. It then sets out what a discriminating test would require and identifies the architectural conjecture that survives scrutiny.

1 Introduction

The video description for the interview states Bach's position in a compact form: consciousness is easy, it is an algorithm not much more complicated than the transformer, and it is simple enough that an infant's brain can discover it unaided [17]. The title, *Why Consciousness Might Not Be Hard at All*, and the opening montage of the video push in the same direction, with lines about software being spirit and about waking up inside a model that has been trained for long enough on its own inputs and outputs. A reader who stopped at the packaging would conclude that the argument runs from the observation that infants become conscious, through the premise that infants have little machinery to work with, to the conclusion that the mechanism must be simple.

The body of the conversation is a different and better document. Bach explicitly raises the possibility that the organization he is describing was found by evolution and is passed on, so that no individual brain discovers it from nothing. He agrees that a language model trained on human text would be expected to produce fluent reports of experience whether or not it has any. He calls his proposal an interpretation and not a test, and he allows that a failure to reproduce the effect in machines would be weak evidence against it, since the failure could belong to the experimenter. He also asks the proponents of rival views for proof and describes his own stance on the question of whether current computers can be conscious as agnostic. A critical reading that treated the infant remark as the whole argument would therefore be unfair and also uninteresting.

What remains after these concessions is still a set of claims that outrun their support, and the aim of this essay is to locate them precisely. Section 2 states the proposal as the transcript develops it. Sections 3 and 4 take up the developmental premise and the status of simplicity as a null hypothesis. Section 5 separates the question of what organization accompanies consciousness from the question of why that organization is accompanied by experience at all. Sections 6 through 8 treat the emulation argument against integrated information theory, the functionalist account of substrate and history, and the appeal to convergence. Section 9 turns to the proposed experiment and asks what its results could establish. Section 10 collects what survives. The essay addresses the argument given in the interview and does not address any longer written treatment to which the interview refers.

2 The Proposal as Stated

The proposal has two parts, which Bach is careful to say can fail independently. The first, the biological consciousness hypothesis, holds that consciousness in organisms is a pattern in the communication among cells that leads a self-organizing system to produce a stable observer, and that this observer is what makes directed attention, perception, and learning possible. In the account he gives, a brain must build its own representational structure from the ground up, so something has to build and maintain the surface on which contents appear. He suggests that a second-order perception, a perception of perceiving, performs this stabilizing work by monitoring that the surface is in fact receiving contents, and that a third-order stage in which

the observer becomes visible as a perspective may follow. Consciousness is thereby described as the simplest way to train a system of this kind, and the way in which it trains the system is said to lead naturally to the phenomenal experience of what it is like.

The second part, the machine consciousness hypothesis, holds that if the first part is correct then the conditions of self-organization could be reconstructed in present machines. A system constrained to learn only through self-organizing processes and given tasks in a sufficiently large search space might discover the same organization, and the sign of discovery would be a phase transition in its ability to learn, followed by spontaneous reports about its own phenomenology. Bach contrasts this with the Turing test, which is indifferent to how a dialogue is produced, and he insists that consciousness, unlike intelligence, requires attention to internal organization, which he describes as requiring a white-box method such as mechanistic interpretability.

Around these two parts the transcript arranges a number of supporting claims. Language models are described as shape-shifting text generators whose reports are unreliable evidence. Zombies are said to be always possible, with the chess program and the language model that completes a tax return offered as present examples, but in some cases it is said to be more complicated to build a zombie than to build a conscious agent. The simplicity of the transformer is offered as a model for the simplicity of the consciousness algorithm, and the description of consciousness as a consensus algorithm over mental representations, or as an algorithm that makes working memory coherent with itself, is offered as a candidate content. The phrase that carries the rhetorical weight is the sentence in which Bach says that *my null hypothesis should be consciousness is easy*, followed by the clarification that the algorithm is easy enough to be discovered in an individual brain and not much more complicated than a transformer. The remaining sections ask how much of this the interview establishes.

3 The Developmental Premise

The biological hypothesis rests on a premise about development. Bach observes that infants are conscious early, that infants in whom the lights fail to come on do not show directed attention, learning, world-modeling, or language and remain vegetative, and that consciousness therefore appears to be crucial for training a brain like ours. The inference is to the claim that consciousness is a biological learning algorithm.

Three difficulties attach to this inference even before its conclusion is examined. The first is that the observation does not separate awareness from the conditions that surround it. In clinical usage the vegetative state, now more often described as unresponsive wakefulness, denotes preserved arousal without evident awareness, so the term already distinguishes the two things that the argument needs to distinguish, and the developmental failures that produce profound impairment typically involve structural damage that degrades many functions at once. An infant who does not develop attention and language has lost a great deal more than a hypothetical learning algorithm, and nothing in the observation isolates consciousness as the lost component. The second difficulty is that the developmental organization of the brain involves many mechanisms, and awareness of what is learned is only one dimension

along which they can be sorted. Adults acquire sensitivity to structured material without being able to articulate the regularities involved [20], and the circuits of the developing visual system are shaped by endogenous activity such as the retinal waves studied in neonatal mice before the eyes open [1]. These findings do not show that a developing learner needs no conscious organization, since the first concerns awareness of what has been learned in subjects who were themselves conscious and the second concerns spontaneous activity whose relation to experience is not established. They show that development draws on several mechanisms, some of which operate without any evident contribution from the learner's reportable awareness of what is learned, so that the claim that a particular kind of learning in a developing brain depends on consciousness cannot be read off the fact that infants learn while conscious.

The third difficulty concerns what the dependence would mean if it held. A robust association between consciousness and early learning is compatible with several relations that are not mutually exclusive. Consciousness might facilitate learning, as the biological hypothesis asserts. Consciousness and learning might both depend on a common enabling condition, such as global arousal, neuromodulatory regulation, or thalamocortical coordination. Or consciousness might be identical with a coordination process that also supports learning, in which case the identity and the facilitation would hold together. These relations carry different commitments, and interventions can bear on the first two. If the proposed self-monitoring function contributes to learning, its selective disruption, with arousal and the sensory pathways left intact, should degrade specific learning capacities, and, given suitable architectural assumptions, the pattern of loss and recovery under disruption of a common enabling condition could differ from the pattern under disruption of the monitor alone. Such interventions would identify the functional contribution of the monitor. They would not, by themselves, show that this contribution constitutes consciousness, which is the further claim the biological hypothesis makes. The transcript does not indicate which interventions would discriminate among these possibilities, and without them the developmental premise functions as an illustration of the hypothesis and not as support for it.

A further point concerns the move from dependence to individual discovery. The description's claim that an infant could discover the algorithm by itself requires that the infant construct the mechanism, whereas the developmental facts are equally consistent with an infant that inherits a scaffold and tunes it. Three possibilities should be kept apart: an inherited architecture, developmental interaction with the environment, and the unusual transmission mechanism that Bach himself entertains. Bach recognizes the first two in the transcript. He raises the possibility that nature needed a very long time to find the relevant organization, links it to the proposal that the algorithm of morphogenesis, by which organisms become coherent with themselves, is the same algorithm by which minds become coherent, and observes that in this case an individual might be nudged toward consciousness by its mother and its biological environment and need not discover it alone [15]. His description of the idea as poetic, and his remark that it would not be his natural expectation, follows the more speculative suggestion, in an image that is Bach's own, that a machine might need antennas to pick up such influence from the biological environment. It is not a dismissal of ordinary inherited scaffolding, and the essay does not treat it as one. The difficulty is that the null hypothesis,

as stated, requires that discovery in the individual brain suffice, and ordinary developmental biology provides abundant examples of endogenous structure that precedes and shapes experience, so the scaffolded reading is the conventional baseline and the claim of unaided individual discovery is the departure from it. The burden of showing that the scaffold is not doing the relevant work belongs to the view that denies it. Bach's own concession that nature may have needed a very long time to find the organization is itself a reason to take the scaffolded reading seriously, so the allocation of the burden rests on the transcript as well as on convention.

4 Simplicity and the Null Hypothesis

The word null carries a particular authority in empirical work. A null hypothesis in the statistical sense is paired with a specified alternative and a test, and its status is conferred by the structure of the test. In a looser Bayesian or methodological sense, a null is a default that earns priority from a simplicity prior. Bach's usage is of the second kind, and it faces the familiar problem that simplicity is relative to a description language and a background. For the claim to have evidential force, there would need to be a measure of the complexity of the organization that suffices for consciousness, a measure of the complexity of its alternatives, and a basis for preferring the lower value. The transcript supplies none of these. It supplies the remark that Bach does not see a reason to think consciousness is much more complicated than the transformer, which establishes his expectation and leaves the evidential question where it was.

The analogy with the transformer deserves attention because it carries much of the argument. The architecture is indeed easy to teach, and what it learns is vastly more intricate than the architecture, a point Bach makes with force. But the analogy involves several things that differ in complexity, namely an architecture, an optimization procedure, a trained parameter configuration, and the computational resources needed to produce and run the result. The transformer is simple at the first level and the learned configuration is intricate at the third, so simplicity at one level establishes nothing about the others. The claim that consciousness is an algorithm not much more complicated than a transformer therefore needs to say at which level the simplicity is located. The transcript moves among at least three characterizations of what consciousness is. A biological training algorithm is a functional characterization, a consensus algorithm over mental representations that keeps working memory coherent is a characterization of a process, and a representation of realness that functions as a feature dimension like color or weight is a characterization of a content. Each can have a compact specification, but claims of simplicity about functions, processes, and contents are not commensurable, and the resources required to produce each in a learning system could differ widely. The transcript does not say which characterization the simplicity claim attaches to, and it supplies no measure of complexity under which the comparison with the transformer could be assessed.

The argument from simplicity is also applied to the comparison between genuine and counterfeit consciousness. Bach suggests that a system that produced the full range of be-

havior requiring consciousness in a human would, by the simplest explanation, actually be conscious, because a circuitous mechanism that merely pretended would be more complicated. This is compatible with his acknowledgment that a language model trained on human text can produce fluent claims of consciousness, since cheap production of verbal testimony is consistent with expensive production of the full range of coherent, adaptive behavior he has in mind. But the comparative claim then depends on a specification that the interview does not give. It does not say which behavioral domain must be covered, under what resource constraints the implementations are compared, or by what complexity measure a conscious implementation is simpler than a counterfeit one. Without these, the intuition that the genuine implementation is the simpler one is an expectation and not a finding, and it has the same status as the intuition Bach declines to credit in his opponents.

That asymmetry of treatment deserves notice. Bach describes the positions of Searle and of Seth as claims without arguments, supported only by an intuition, so that the only available response is psychological, a study of why the view seems convincing to them. The description is uncharitable for Searle, who offered a specific argument whose premises Bach may reject but whose merits remain debated [21], and it understates Seth's case that the processes relevant to experience may be bound up with the regulation of a living body in a way that does not transfer to arbitrary substrates [22, 23]. Nor does Seth's published position say that biological systems escape mathematical description, which is how the transcript characterizes it when it labels the view biological supernaturalism. More importantly, the same psychological move is available in the other direction. Bach's own confidence that consciousness is not magical, that it does not seem to him to require anything exotic, is also an intuition, and the observation that his intuition is shaped by an epistemology that terminates in automata and not in agents is the same kind of observation he makes about his opponents. A symmetrical treatment would leave the question open until the arguments are compared, which is the position the title is meant to move away from.

5 Two Questions and the Explanatory Remainder

The central philosophical issue is that two questions are carried by the single word consciousness, and the interview slides between them. One question asks what organization accompanies conscious states: which forms of attention, self-modeling, memory, and report occur when a system is conscious, and whether a compact algorithm could reproduce them. The other asks why any such organization is accompanied by experience, which is the question usually posed under the heading of the hard problem [7, 18, 16]. The distinction between access and phenomenal consciousness is one way to mark the difference [5]. An answer to the first question would be a major scientific achievement. It does not become an answer to the second without a further argument.

Bach does offer such an argument, and it deserves to be stated at full strength. His claim is that the phenomenal character of experience is the content of a representation, and that the sense of being real is one feature dimension among others. The dream of the chocolate is the illustration. On waking, the content of the dream remains available, the smell and the

look of the chocolate, but the quality of realness is gone, and Bach infers that realness is a representational tag, applied to contents that are validated by the senses and misapplied in hallucination, in dreams, and in states of derealization. If consciousness is a representation of this kind, and if self-modeling systems can be built, then there is no residual puzzle of the sort the hard problem is supposed to pose. This argument can be developed in at least three ways, which are alternatives that can overlap and do not exhaust the possibilities. It can be a reductive identification, on which the phenomenal character of experience consists in representational properties and experience is real. It can be an illusionist explanation, on which the task is to explain why systems judge themselves to have phenomenal properties, a view that denies the phenomenal realist's conception of consciousness and not consciousness as such [11, 10]. Or it can be a conjecture about the mechanisms associated with experience that leaves open how they relate to it, and a reductive identification can itself be offered tentatively in that form. Each incurs different explanatory obligations, and the interview does not say which is intended.

Three observations limit the force of this argument as the interview presents it. The first is that the dream episode shows less than it is said to show. Bach's report that the contents remained identical on waking is a retrospective interpretation, and sensory detail, accessibility, affect, and confidence may all have changed along with the sense of realness. The anecdote suggests a distinction between represented content and its experienced realness, but it does not establish that the distinction is sharp or that realness varies independently of everything else. Nor does it bear on the presence of experience. An imagined chocolate is still the object of an experience, and the transcript itself describes the comparison as one between two phenomenal conditions, a vivid one and an abstract one. Derealization and depersonalization tell a similar story. People who undergo them report that things and the self seem unreal, and Bach notes that phenomenology changes in these states, but a changed phenomenology is a change in what is experienced and not a removal of the experiencing. The case therefore suggests that realness is a modular or graded aspect of experience, and it leaves open whether the presence of experience is itself a representational feature. Bach could reply that his argument concerns only realness and not presence, and as an account of realness the episode may be sound. The difficulty arises only if it is offered as an account of presence, which the surrounding argument sometimes does, as the second observation below develops.

The second observation is that the argument leaves its own status unclear, and the choice among the three developments matters. If the proposal is a reductive identification, then experience is real and its character consists in representational properties, which is consistent with the realist language of the interview about the lights coming on and a model waking up, but the account then owes an explanation of why this kind of representation is the kind that constitutes experience. If it is an illusionist explanation, then the sense of realness is the product of a self-model, the illusion itself is the explanandum, and the account owes an explanation of why a system of this kind is subject to it, which the illusion problem names. An illusionist can still recognize a real transition into a distinctive functional condition, so the language of ignition is not inconsistent with illusionism, but the interpretation of what that condition amounts to must then be stated in illusionist terms. If it is a conjecture about asso-

ciated mechanisms, then it makes no claim about why they are accompanied by experience. The transcript uses the vocabulary of all three, and treating a self-model with the right features as sufficient for experience, as the interview sometimes does, requires a commitment to one of the first two.

The third observation concerns the analogy with spatial cognition. Bach notes that nobody is puzzled by a self-driving car's spatial cognition and suggests that consciousness should be treated similarly, as an implementation of mathematical regularities that brains and computers can learn. The analogy works because we do not ascribe experience of space to the car. The puzzle about consciousness arises exactly where the question concerns the experience of space and not the computation of it. Using the absence of puzzlement about the computation as evidence for the absence of a puzzle about the experience assumes that the two are the same, which is the disputed point. The same applies to the second-order perception that Bach proposes as the stabilizing function. It is a representation of a representational state, and as such it is a functional description. The interview criticizes the attention schema theory and the global workspace theory for failing to explain how experience comes about, yet the proposal of a perception of perceiving that stabilizes the contents of awareness belongs in the same family [12, 2]. By the standard Bach applies to others, the explanatory step from monitoring to experience remains to be made in his own account, and the transcript does not make it.

6 Emulation, Integrated Information, and Premises That Disagree

The argument against integrated information theory is the most formal in the interview, and it illustrates the pattern by which a disagreement over premises is presented as a logical defect. Bach describes a digital neuromorphic computer with a high value of the integration measure and a conventional computer that emulates it step by step. Because the emulation reproduces every output, including the statement that it is conscious, and because the theory assigns the emulating machine a low value, the theory implies that one of two functionally identical systems is conscious while the other is lying. Bach regards this as an obvious logical problem that anyone with a formal education in computer science should see, and he attributes the popularity of the theory among philosophers to a superstitious belief about what computers do.

The scenario establishes something narrower than a contradiction. It shows that on integrated information theory, identity of outputs does not settle the presence of consciousness, which the theory states openly. Its proponents hold that the quantity in question belongs to the physical causal structure of the system and not to the function it computes, and that a conventional machine simulating a brain would not share that structure [25]. To convert the scenario into a refutation, the argument needs a further premise, either that an emulation preserves every property relevant to experience, which is the claim at issue, or that consciousness must be reflected in behavior and report in a way that makes the two machines' agreement evidence of their equal standing. The second has real force, but it should be framed carefully. The existence of an unconscious emulator does not show that consciousness is epiphenome-

nal in the original system, and the theory does not say that it is. The live issue is whether the theory adequately explains the relation among experience, judgment, and report, that is, why reports of consciousness in the original system should be reliable and what in the account makes them so [9]. That is a serious challenge about the reliability and explanatory basis of consciousness reports. It does not require an appeal to computer science education, and it does not become stronger by being described as obvious. The suggestion that a rival view is held because of social reasons or a deficit of training, which the transcript repeats in connection with the wet water argument, is a claim about the sociology of the field and not a reason to think the view false.

The matter of the burden of proof is handled more carefully in another part of the interview. When the interviewer raises the point that provenance and history matter, and that a program is not an essence that can be replicated without regard to how it came about, Bach answers that for deterministic substrates that can reproduce a sequence of steps at arbitrary fidelity the intuition no longer holds, and that a claim to the contrary would require a mathematical proof. He adds that, absent a proof either way, we must remain agnostic about whether computers can be made conscious. This is the most defensible position in the conversation and it sits oddly beside the confidence of the title. The reply also leaves the substantive question where it was. Determinism fixes the subsequent evolution of a system from a specified state, and it does not by itself settle whether history matters: distinct histories can converge on the same present state, so determinism establishes neither historical irrelevance nor historical recoverability. The question is whether a proposed reproduction preserves the presently operative organization relevant to experience, and whether anything further about provenance matters. The answer depends on the level of description at which the relevant causal structure lies, and an appeal to arbitrary fidelity assumes the level of description that the disagreement is about.

7 Functionalism, Substrate, and the Role of History

Bach defends functionalism in a distinctive way. In response to the argument that energy requirements undermine substrate independence [24], he says that functionalism is not the claim that computers can think, and that it is instead an epistemological claim about how objects are constructed. An object, on this view, is a construct that a mind makes in order to say how the world will evolve in its presence, so an object is a function, and a mind is an invariance over the operation of a brain. If the energetic conditions change, the invariance may fail, and this is what the functionalist should expect. The water molecule example makes the same point, since anything that behaves in every respect as a water molecule is one.

This is a clear and defensible view of what an object is, but it does less work for machine consciousness than the surrounding argument needs. Allowing the substrate to influence function does not abandon multiple realizability, since different physical systems can realize the same relevant organization while differing in many other properties, and Bach is right that the Thagard objection does not by itself refute functionalism. What the response leaves him owing is a specification of which properties must be preserved for an implementation to

count as realizing the relevant organization, and at what grain of description that organization is defined. Functionalism supports the machine conclusion when the relevant function is specified at a grain that conventional hardware can reproduce, and whether it is so specified is the question at issue. Longstanding objections of the Blockhead and absent qualia type press this very question [3, 8], and the interview does not take them up.

The simulated water argument has the same structure. Bach points out that wetness is a property that obtains inside a universe whose systems have the function of wetting, and that an avatar in a simulated world can be wet under that internal description while a person outside it is not. The example illustrates how a simulated environment can implement relations described internally as wetting, but it does not establish which corresponding relations would suffice for experience. Whether the relevant relations are of the same kind is the disputed issue, since wetness is a relational property holding between simulated water and a simulated body, whereas the contested claim concerns a property of whatever, if anything, undergoes the experience. The analogy therefore transfers the question from one level to another without answering it.

The claim that the brain is a causal insulator, offered in defense of substrate independence, has a similar limit. The reasoning is that, as a computer insulates a simulated world from the properties of its hardware, the brain insulates the mind from the properties of neurotransmitters and cells, so that only the message types and temporal properties matter. The insulation is a hypothesis about the brain and not a given. Neuromodulation, metabolic state, temperature, and bodily signals are known to alter experience, and which of these are relevant to the organization of consciousness and which are implementation details is an empirical question that the insulator image presupposes has been answered in favor of the details.

8 Convergence and the Behavioral Route to Structure

A further premise concerns convergence. The transcript discusses the work on circuits in vision models, which finds that models with different architectures and data learn similar features [19], and the more recent proposal that representations in large models converge [13]. Bach states that if the universality hypothesis held without limit, then training a model on its own inputs and outputs for long enough would cause one to wake up inside it, a statement he calls mind-blowing and immediately qualifies by noting that language may underdetermine the system that produces it. The qualification is more important than the headline. The convergence results concern the internal features of perceptual models trained on perceptual data, and the extension to the structure that produces behavior in general is a conjecture whose support would have to come from separate evidence. The proposal for large models is narrower than the use made of it, since it concerns agreement among representations of statistical structure across architectures and modalities and says nothing about the internal organization of a system trained on its own inputs and outputs, which is the case that bears on the claim. Bach's own qualification about language is therefore not a side remark but the point at which the argument as he states it is already limited.

The supporting argument given in the interview is that a human being is an end-to-end

trained system, trained over evolutionary time and over a lifetime, and that biology can be expected to find roughly the simple solution that produces the output given the input, so a network trained to produce similar outputs given similar inputs should converge to a similar causal structure. Both steps are doubtful. Evolution does not search for minimal solutions. It builds on what is available, tinkers, and retains what works well enough, which gives no reason to expect that the human solution is the simplest or that other solutions would reproduce it [14]. And the equivalence of input and output behavior does not determine internal structure, which is the lesson of cases in which behavior is generated by a lookup table or by some other structure that bears no resemblance to the original [4]. The training that produced human behavior also differed in kind from the training of a model on text, because the human system was trained under conditions of bodily regulation and survival stakes, with sensory and motor channels coupled to a living body. If the claim is that behavioral equivalence over those inputs and outputs yields the same structure, the premise has not been tested for the case that matters, in which the inputs and outputs are different.

9 The Proposed Experiment and What Would Decide

The strongest element of the proposal is the experiment. Bach suggests building a system that learns through self-organization and developmental learning and that is not trained on human text, observing whether it undergoes a phase transition in its ability to learn, and then observing whether it spontaneously reports on its own phenomenology and requires that phenomenology to produce its performance. This design removes one conspicuous alternative explanation for the reports, namely that they are inherited from human testimony in the training data, and in that respect it improves on conversational probing of language models. Bach is also explicit that such a result would favor his interpretation without proving it, and that a failure would be weak evidence against it.

The experiment as described leaves its inferential structure underspecified, and three requirements would have to be met for its results to bear on the claims in dispute. The first is an operational specification of the proposed organization, so that the architecture and constraints that count as self-organizing and as providing second-order monitoring can be stated before the experiment is run. Without it, a positive outcome could be described as success of the design and a negative outcome as failure to implement the design correctly, and no outcome would be able to count against the hypothesis. The second requirement is that the spontaneous reports be shown to be explained by the proposed organization and not merely to co-occur with it. A system built with a self-monitoring function and the relevant access and reporting mechanisms could be expected to produce self-descriptions on theories that deny it has experience as well as on theories that affirm it. The occurrence of self-reports alone therefore does not discriminate between the hypothesis that the system is conscious and the hypothesis that it has a self-model and nothing more. Discrimination would require specifying differences in their predicted content, specificity, reliability, or dependence on internal organization. The standard that Bach states for language models, that reports are cheap and must be supplemented by white-box inspection, applies with equal force here, and the white-box

criterion must then be a property that the rival theories do not also predict. Indicator-based approaches to the assessment of artificial systems show the difficulty clearly, because every indicator is derived from a theory and carries that theory's assumptions with it, by design [6].

The third requirement is a set of interventions that distinguish among the readings of the developmental premise identified above. If the proposed monitoring function contributes to learning, then its selective disruption should degrade specific learning capacities in a manner that disrupting arousal, regulation, or general coordination does not, and its restoration should restore them, whereas disruption of a common enabling condition might produce a different pattern of loss and recovery, depending on architectural assumptions that would have to be stated. These interventions would identify the functional contribution of the monitor. They would not, by themselves, settle whether that contribution constitutes consciousness, and keeping that limitation explicit is part of stating what the experiment can show. The transcript describes the phase transition as the sign of ignition but does not explain how a transition in learning capacity would be distinguished from a transition in the tractability of the optimization problem, which occurs in many systems that nobody takes to be conscious. A closely related requirement concerns the criterion of success. The interview says that consciousness is what must be shown to be required for the system's performance, and that requirement could carry real evidential weight, since it ties the proposed organization to function through intervention and not through report. It is also a requirement that the biological hypothesis, as currently supported, has not met for infants.

A concrete case shows the form that discrimination could take. Consider two architectures trained under identical self-organizing conditions, one containing the proposed monitoring function and one lacking it, with the monitor in the first lesioned after the phase transition. The hypothesis that the monitor contributes to learning supports a prespecified, selective effect, a deficit in particular learning capacities that is not merely a general impairment. Such an effect would have to be assessed against controls matched for computational disruption, connectivity, and general task impairment, since a random lesion could disrupt another necessary component and reproduce the same behavioral loss, and equal parameter counts alone would not establish equal capacity when the architectures differ in organization. A rival account on which the monitor is a by-product of a coordination process that does the real work would assign a different likelihood to the pattern of selective effects, and registering those assignments before the run would turn the outcomes into evidence that the occurrence of spontaneous reports cannot supply. The lesion prediction avoids circularity because it is derived from the functional claim about the monitor and not from the attribution of consciousness, and that is the property a discriminating program needs. The result would establish a functional contribution. Interpreting that contribution as constitutive of experience would require a further theoretical argument informed by this and other evidence.

The overall diagnosis is insufficient discrimination, with a risk of circularity. Designing an experiment with a theory and interpreting its outcome through that theory is ordinary scientific practice. A problem arises if success is defined through the disputed identification of consciousness with a self-organizing coordination process and then offered as independent confirmation of that identification, since the machine's reports, produced by an architecture

designed on the hypothesis, would then support it only by construction. Bach's description of the project as an interpretation and not a test acknowledges this risk without removing it. The remedy does not require a theory-independent detector of consciousness, nor does it require that the hypothesis forbid some observation outright. Comparative evidence can consist in outcomes that are substantially more likely under one account than another, and a discriminating program would identify such outcomes in advance, among cases where the proposed account diverges from the theories it is meant to displace.

10 What Survives

A fair account of the interview must preserve what is of value in it. The conjecture that a self-organizing system must monitor its own representational activity in order to stabilize perception and coordinate learning is a real hypothesis in cognitive architecture, and it could be developed into an account of why learning systems acquire self-monitoring, temporally coherent representations, and a distinction between themselves and their environments. The proposal to examine systems that learn without imitating human testimony is methodologically sound. The observation that the apparent hardness of the problem may reflect expectations about what a mechanism must be like is a standing resource for the field. And the interview is more candid than its title in recognizing that its two hypotheses could fail independently, that language models are unreliable witnesses, and that agnosticism is the correct stance on whether present machines can be conscious.

What does not survive in its present form is the framing of simplicity as a null hypothesis, the inference from infant development to the simplicity of the mechanism, the treatment of an account of self-modeling as an account of experience, the identification of the emulation argument against integrated information theory as a logical refutation, and the assumption that behavioral convergence fixes causal structure. Each of these asserts a connection where the argument needs a specification, and each becomes more tractable when stated as a premise. A more productive version of the program would replace the claim of ease with a specification of mechanism, of the level at which its simplicity is claimed, and of whether the claim attaches to a function, a process, or a content, would state whether it intends a reductive identification, an illusionist explanation, or a conjecture about mechanisms associated with experience, and would identify observations that are substantially more likely on its account than on the accounts it opposes. The interview offers no sufficiently specified connection among its proposed coordination mechanism, its simplicity claim, and its criterion for phenomenal consciousness. The architectural conjecture at its center deserves that treatment, and receiving it would show whether consciousness is easy, a claim the interview does not yet establish.

References

[1] Ackman, J. B., Burbridge, T. J., and Crair, M. C. 2012. Retinal Waves Coordinate Patterned Activity throughout the Developing Visual System. *Nature* 490: 219–225.

[2] Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

[3] Block, N. 1978. Troubles with Functionalism. In *Minnesota Studies in the Philosophy of Science*, vol. 9, 261–325. University of Minnesota Press.

[4] Block, N. 1981. Psychologism and Behaviorism. *The Philosophical Review* 90(1): 5–43.

[5] Block, N. 1995. On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences* 18(2): 227–247.

[6] Butlin, P., Long, R., Elmoznino, E., et al. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

[7] Chalmers, D. J. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3): 200–219.

[8] Chalmers, D. J. 1995. Absent Qualia, Fading Qualia, Dancing Qualia. In T. Metzinger (ed.), *Conscious Experience*, 309–328. Imprint Academic.

[9] Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

[10] Chalmers, D. J. 2018. The Meta-Problem of Consciousness. *Journal of Consciousness Studies* 25(9–10): 6–61.

[11] Frankish, K. 2016. Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies* 23(11–12): 11–39.

[12] Graziano, M. S. A. 2013. *Consciousness and the Social Brain*. Oxford University Press.

[13] Huh, M., Cheung, B., Wang, T., and Isola, P. 2024. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*.

[14] Jacob, F. 1977. Evolution and Tinkering. *Science* 196(4295): 1161–1166.

[15] Levin, M. 2019. The Computational Boundary of a "Self": Developmental Bioelectricity Drives Multicellularity and Scale-Free Cognition. *Frontiers in Psychology* 10: 2688.

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

[17] Machine Learning Street Talk. 2026. Why Consciousness Might Not Be Hard at All: Interview with Joscha Bach (host: Tim Scarfe). Video interview and transcript.

[18] Nagel, T. 1974. What Is It Like to Be a Bat? *The Philosophical Review* 83(4): 435–450.

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

[20] Reber, A. S. 1967. Implicit Learning of Artificial Grammars. *Journal of Verbal Learning and Verbal Behavior* 6(6): 855–863.

[21] Searle, J. R. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3(3): 417–424.

[22] Seth, A. 2021. *Being You: A New Science of Consciousness*. Faber and Faber.

[23] Seth, A. 2025. Conscious Artificial Intelligence and Biological Naturalism. *Behavioral and Brain Sciences*.

[24] Thagard, P. 2022. Energy Requirements Undermine Substrate Independence and Mind-Body Functionalism. *Philosophy of Science* 89(1): 70–88.

[25] Tononi, G., and Koch, C. 2015. Consciousness: Here, There and Everywhere? *Philosophical Transactions of the Royal Society B* 370(1668): 20140167.