Behavioral Evidence is Still Insufficient to Identify Consciousness

Maria Vorobeva (mariavorobeva@cmail.carleton.ca)
Department of Cognitive Science, Carleton, 1125 Colonel By Drive
Ottawa, ON, K1S 5B6, Canada

Eilene Tomkins-Flanagan (EileneTomkinsFlanaga@cmail.carleton.ca)
Department of Cognitive Science, Carleton, 1125 Colonel By Drive
Ottawa, ON, K1S 5B6, Canada

Mary Alexandria Kelly (Mary.Kelly4@carleton.ca)
Department of Cognitive Science, Carleton, 1125 Colonel By Drive
Ottawa, ON, K1S 5B6, Canada

Abstract

Researchers have started seriously considering the epistemic issue of whether and when we can claim an artificial intelligence (AI) has developed *machine consciousness*. Most cognitive theories of consciousness employ a *functional* characterization of the *property* of consciousness. That is, they are committed to an account of consciousness as a rule-governed *process* over *mental states*. Some cognitive scientists concerned with AI advocate an epistemically behaviorist approach to *machine consciousness*; however, such approaches taken ontologically, systematically fail to satisfy reasonable intuitions about in what consciousness ought to consist, and taken epistemically, *fail to provide sufficient evidence to individuate any internal property*, including consciousness, in non-human subjects. Therefore, in order to assess consciousness in ways that adequately account for reasonable intuitions as to its proper definition, such that we can reasonably assert the presence of *machine consciousness* in some AI, it is necessary to propose, test, and revise, *functional theories of consciousness*.

Keywords: behaviorism; functionalism; consciousness; artificial intelligence; machine consciousness

Advancements in artificial intelligence (AI) have put questions of *consciousness* into focus. In particular, authors have started seriously considering the epistemic issue of whether and when we can claim an AI has developed *machine consciousness* (Butlin et al., 2023). Of the several cognitive theories advanced which attempt to describe how consciousness might be attributable to machines, such as global workspace theory, perceptual reality monitoring and recurrent processing theory (Butlin et al., 2023; Hildt, 2023; Reggia, 2013), all rely implicitly on a functional characterization of the *property* of consciousness. Cognitive theories propose that properties such as consciousness consist of some objects (that may be interpreted as information-bearing states) undergoing a physical process describable by well-defined rules, said to proximally cause behavior. We take the characterization of behavior as mediated by rule-governed processes over some objects, and the position that behavior is proximally caused by these mediating processes to define philosophical functionalism, as it pertains to the mind.

Functionalism may be ontological or epistemic. Cognitive science is always ontologically functionalist, inasmuch as it is concerned with cognitive processes as that in which the mind is said to consist. Rejecting ontological functionalism implies either ontological behaviorism (whereby it is said that properties like consciousness can only be attributed to behavior), or positions like epiphenomalism (Robinson, 2010, whereby some properties of the mind are said to be *real* but are irrelevant to a functional characterization). However, cognitive science is not always epistemically functionalist. That is, when identifying the cognitive process a subject is undergoing, a cognitive scientist may be satisfied to infer such a process from behavioral evidence alone; thus they can be considered epistemically (i.e., methodologically) behaviorist (Moore, 1989). A cognitive scientist who requires evidence of mediating objects from sources other than behavior (such as fMRI data in humans, or the structure of AI models) can be thought of as epistemically functionalist. It is our position that, between epistemic behaviorism and epistemic functionalism, only epistemic functionalism can meaningfully attribute consciousness to nonhuman subjects, including AI.

Although the *cognitive revolution* went a long way towards justifying the functionalist study of the mind (Miller, 2003; Neisser, 1967), behaviorism has not gone extinct, even among cognitive scientists. Appeals to behaviorism may be of an ontological flavor, holding that properties like consciousness can only be understood as a behavior and cannot be meaningfully addressed as an internal property; proponents of this position such as Leslie (2015) and Moore (2009) argue consciousness is a private event, that is an internal reaction to behavior that cannot causally bear on behavior. Appeals to behaviorism may also take an epistemic flavor, claiming behavioral evidence (such as comparison to humans on modified turing test Harnad & Scherzer, 2008) is all we need to infer consciousness, even *machine consciousness*, even if consciousness is ultimately a process (Harnad & Scherzer, 2008; Palminteri & Wu, 2025).

Behaviorist arguments respecting consciousness have intuitive appeal due to their (apparent) ability to do away with the epistemic difficulty of referring to mediating objects like mental states, this is because behavior alone cannot individuate a mental state without assumptions about the nature of the behaving subject that allow inference of mental states from behavior. However, we will demonstrate how, taken ontologically, behaviorist arguments systematically fail to satisfy reasonable intuitions about in what consciousness ought to consist, and taken epistemically, they fail to provide sufficient evidence to individuate any internal property, including

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

consciousness, in non-human subjects. In what follows, we will first review how behaviorism fails to evidence consciousness, and then, we will defend functionalist ontological and epistemic commitments as adequate to the identification of *machine consciousness*.

Can Behaviorism Address Consciousness?

To determine if behaviorism can provide compelling evidence for the presence of consciousness in any subject, it is necessary to examine (1) our intuitions as to what the property of consciousness must consist of, (2) what evidence is necessary and sufficient to identify it in a subject and (3) whether a behaviorist approach can provide evidence that meets the necessary and sufficient conditions.

Block (1995) identified that discussion of consciousness is complicated because consciousness is a "hybrid-concept": it "connotes a number of different concepts and denotes a number of different phenomena" (p. 1). Properly, there is no one property of "consciousness", but several distinct properties we may intend when we refer to "consciousness". Predominantly, the various senses of consciousness researchers use are defined from an implicitly or explicitly functional perspective (Butlin et al., 2023; Reggia, 2013; Hildt, 2023). That is, they concern chiefly processes mediating behavior. Of the senses of consciousness Block described, those most thoroughly discussed in the literature are "access consciousness" (the availability of representations of one's internal state for reasoning, executive control, and action), and "monitoring consciousness" (by which one can reflect on one's own states and behaviors; Butlin et al., 2023; Hildt, 2023; Reggia, 2013).

Butlin et al. (2023) and Reggia (2013) surveyed several competing theories of consciousness. Most are functional theories, in that they characterize consciousness as a mediating process that causally bears on behavior. Each functional theory surveyed privileges some senses of consciousness in their definition of consciousness as such. They may have disagreements about the processes relevant to defining consciousness (e.g. Butlin et al., 2023 discuss at length the contention between higher order theorists and global workspace theorists), or about whether machines capable of such performance are realizable outside of a biological substrate (Aru, Larkum, & Shine, 2023); but they are all committed to an account of consciousness as a process over mental states. The relevant objects in these functionalist accounts are mental states because it makes little sense to talk about the availability of information for reasoning or self-reflection if one cannot refer to mental events that may carry such information. Referring to mental states puts functional theories of consciousness in conspicuous tension with behaviorism, wherein "mentalist" (Skinner, 1974, ch. 1) is practically a slur.

Studying consciousness as a process necessitates a theoretically-driven approach to the study of behavior (Butlin et al., 2023), and the use of introspective report in constructing theory (Block, 1995). The construction of theories of consciousness must guide the selection of competing hypotheses and experimental protocols to assess their validity, with experimental protocols often relying on participant introspective self-report. For example, Lau and Passingham (2006) wanted to see whether the global workspace account of ignition (the point at which the mind enters into an access-conscious state) is adequate to explain our intuition regarding phenomenal consciousness (often described as the subjective experience of being a certain subject in a certain mental state, cf. Nagel, 1974). Lau and Passingham show that most people can experience relative blindsight: their ability to discriminate between shapes functions independently of the ability to report having seen those shapes. Lau and Passingham therefore argue that access consciousness cannot alone account for phenomenal consciousness, as subjects could access perceptual information, without reporting a subjective experience of it. Although, to us, subjective report seems more closely related to monitoring than phenomenal consciousness, the example serves to illustrate that construction and comparison of theory is the backbone of cognitive research into consciousness. Hence, for cognitive scientists studying consciousness, their epistemology of consciousness is fundamentally theory-driven. Importantly, both the intuitions from which theory is devised and the means of report of a conscious state arise chiefly from introspection.

Conversely, the radical behaviorist position opposes reference to any mediating objects, including consciousness, in describing behavior (Skinner, 1950; Watson, 1913; Moore, 2001) and, furthermore, rejects both introspection(Watson, 1913) and hypothesis-driven testing(Skinner, 1969) as useful tools for psychology. With respect to whether consciousness or its influence can be observed from the measurement of behavior, Watson (1913) argues that "one can assume either the presence or the absence of consciousness ... without influencing in any way the mode of experimental attack upon [behavior]" (p. 161). Further, he criticizes the method of introspection, arguing that it is inherently unreliable, and that the difficulties psychologists have reproducing one another's findings are due to fundamental misunderstandings of terminology consequent to the flawed nature of introspection. Skinner (1969) questions the utility of hypotheses in psychology at all, especially those that involve mediation of behavior by mental states and processes like consciousness. According to Skinner, the effects of mental states are so small or inaccessible that, should any experiment focus on manipulating them, it "is only because the investigator has turned his attention to inaccessible events - some of them fictitious, others irrelevant" (p. 19). Skinner and Watson's position is radically behaviorist because it completely rejects even referring to mental states or processes.

The radical approach is not the only behaviorist philosophy respecting the proper objects and methodology of psychology. Like Watson, logical behaviorists are motivated by the conceptual confusion of mental terms, but, although Skinner's behaviorism was only pragmatic, the logical behaviorists view "inaccessible" mental events as metaphysically un-

acceptable. However, unlike the strict radicals, logical behaviorists like Carnap (1959) allow for the construction of theoretical terms in psychology, so long as the use of such terms in explanations of behavior are reducible to some set of propositions about observations of behavioral data. Thus, logical behaviorists rely on both "observational terms, which referred to publicly observable phenomena; and theoretical terms, which referred to logical constructs that were anchored to observation" (Moore, 2001, p. 223). Logical behaviorism is still radical because it requires all possible mental states to be available for observation: "all talk about 'mental events' is translatable into talk about overt behavior" (Putnam, 1980, p. 45). Thus, for the logical behaviorist, consciousness may not exist at all, let alone be known of from observation, since any psychological proposition must be exclusively about behavioral data to be meaningfully true. Hereafter, we will refer to both the metaphysical logical behaviorists and the pragmatic Skinnerian behaviorists as *radical* behaviorists. Thus, for the radical behaviorist, not only do theoretical constructs fail to explain behavioral events better than an interpretation of complex conditioned responses (Skinner, 1950; Watson, 1926) but they also aim to avoid the challenge of proving the existence of a construct that cannot be directly linked to observable behaviors.

In between the two positions is *methodological behaviorism*. Methodological behaviorists are ontologically functionalist, in that they believe it is acceptable to refer to mental properties mediating behavior, but epistemically behaviorist, in that mental behavior is believed to be a sufficient criterion to individuate any properties a psychologist might be interested in. Writes Moore (1989), for a methodological behaviorist, "publicly verifiable phenomena constitute the leverage by means of which one may meaningfully talk of ... mental phenomena" (p. 20). Methodological behaviorism retains "mentalism" (i.e., cognitivism) by supporting claims respecting mental phenomena with "logical inferences from more elementary phenomena that are publicly observable" (p. 21).

Although cognitive scientists are usually ontologically behaviorist, when they invoke behaviorist arguments to attribute consciousness to AI, their motivations seem to be possessed of a similar flavor to those of the radical behaviorists. They tend to suppose that, since AIs might function differently from humans, machine consciousness may not be the same property as human consciousness. Therefore, functional theories of consciousness may not meaningfully apply to AI, and we must use behavior to attribute consciousness to AI. Elamrani and Yampolskiy (2019) conduct a review of existing tests for machine consciousness, and categorize them as either *architectural* or *behavioral*. In architectural tests, candidates for machine consciousness are structurally compared with human or animal consciousness, in terms of their functional characterization. However, they require us to make ontological assumptions about the universal structure of consciousness before we can verify that an AI is conscious; thus, they can only be relied upon if we assume that machine consciousness and human or animal consciousness are similar processes. Elamrani and Yampolskiy argue that, conversely, behavioral tests might be more useful in assessing machine consciousness because they may be less restricted in relevance to only human or animal consciousness. However, as behaviors used to identify consciousness in such tests are also derived from human and animal examples, behavioral tests seem no less restricted in relevance. In fact, behavioral tests seem unable to detect the case of a conscious being that does not express itself in a way that resembles a healthy human or animal, as with a locked-in patient, where only measurements of the patient's internal state can reasonably evidence consciousness (Wu, Nicolaou, & Bogdan, 2020).

Other authors are more insistent on the superiority of behavioral testing: in AAAI's 2007 conference on consciousness in AI, Harnad and Scherzer (2008) argue that it is misguided to make reference to machine consciousness. For Harnad and Scherzer "[C]onsciousness is feeling, no more, no less" (p. 83). According to Harnad and Scherzer, interest in artificial intelligence for cognitive scientists is limited to the possible behaviors of machine intelligences, and what AI behavior can tell us about human behavior. While Harnad and Scherzer may not be strict radical behaviorists, as they allow that other mental properties might be meaningful, they recapitulate behaviorist arguments with respect to consciousness: they agree that consciousness is non-causal and there is no benefit in using theories of consciousness to explain behavior. They also agree that there are no theoretical foundations for the question of machine consciousness, writing that no "empirical discipline can even begin to explain how or why we feel. Nor is there any sign that they ever will" (p. 75). Harnad and Scherzer's approach to consciousness reads as mostly ignorant of contemporary philosophy of mind. They open with a summary of consciousness that seems only to coarsely cover (Block, 1995)'s phenomenal consciousness, problematize inference about it, then leap directly to radically behaviorist conclusions. Harnad and Scherzer exemplify the looseness of behaviorist answers to the hard epistemic problems respecting machine consciousness, as, when faced with difficulty in making sense of what they are talking about, Harnad and Scherzer unreflexively throw a poorly operationalized test at the problem, that being an extension of the Turing (1950) test, and hope a meaningful distinction between conscious-like and conscious-unlike subjects falls out.

Palminteri and Wu (2025) argue that cognitive science is only concerned with behavioral evidence and explanations of behavior, and that if we commit ourselves to further refining behavioral tests, we might reach a set of such tests that satisfy our intuitions about what it is in which consciousness consists. For Palminteri and Wu, in relating behavioral observations to mental states, we are positing the mental state as a logical construct by which we explain the behavior observed; thus cognitive science is said to be methodologically behaviorist. Palminteri and Wu accept that architectural tests may be *sufficient* to justify attribution of consciousness to an

AI, but such tests are not *necessary* criteria for consciousness attribution. Therefore, they say, some behavioral test must have the necessary and sufficient conditions for consciousness attribution. They argue that their construal of cognitive science overcomes the difficulty of making logical inferences about cognitive processes from behavior, because cognitive science is Bayesian, and Bayesian inference can sometimes infer the acceptability of a hypothesis that is not logically forced by the available evidence. However, Bayesian inference convergently produces the same results as logical inference (Hawthorne, 2024), and, until such time as it converges, Bayesian inference is only more likely to be false than a logical inference in the case of total knowledge. Bayesianism cannot make an insufficient logical reason to make an attribution of some property into a sufficient reason. Palminteri and Wu's claim that behavioral evidence is sufficient to identify consciousness seems to be contradicted by their conclusion that the "science of consciousness must embrace the behavioral methodology of cognitive science in doing away with 'necessary and sufficient conditions', and being willing to continually evolve through corroboration, falsification, and the development of new standards" (p. 7), as though the authors are subtly aware of the insufficiency of their method.

Although it is trivially true that we rely on observations of the world and, therefore, also of behavior, when making and verifying theory, without the discriminatory and simplifying power of theories, we have limited means by which to identify whether a given candidate model has consciousness based solely on its behavioral data. Firstly, behavioral tests for consciousness face the trivial challenge of the very large input-output table. A table that maps inputs to outputs can, in principle, be created such that any behavior is simulated by it arbitrarily well (Block, 2004); that is, for any such test, one can create a table that pairs all testable inputs (experimental conditions), with each output one may reasonably expect to be associated with its corresponding condition in the behavior in question. In Searle (1980)'s classic example, a large input-output table allows a man to appear to conduct a written conversation mechanically, without any language understanding. Though it refers to understanding, the structure of this type of argument matters more than its referent: for *any* property that is not ontologically reducible to behavior, but depends on a mediating mental property, it is possible to construct a function that simulates the behavior in question without possessing the required mental property. Thus, unless the mere appearance of having consciousness is an irrelevant epistemic concern, external behavior can be a necessary, but never a sufficient condition for inferring its presence.

Intuitively, it seems as though input-output tables should not be conscious (and functionally, they neither satisfy the criteria for access-conscious nor monitoring-conscious), but, if we make no ontological assumptions about what it is we are studying, behavioral evidence cannot provide an epistemic distinction between a very compelling input-output table and a conscious being. That is, if we make no claim about how our property of interest influences the world (beyond its behavioral artifacts), we can never with any satisfaction evidence that any specific property (like consciousness) is mediating a behavior, as opposed to any other. Even if we were certain that behavioral evidence could be sufficient to identify consciousness, van Rooij et al. (2024) proves that, if we cannot make strongly constraining assumptions about a given AI model, we cannot attribute to it any particular behavioral property, including consciousness. van Rooij et al. show that, in order to decide that a given model behaves according to a desired proposition defined over behavior with confidence greater than chance, we must representatively sample the behavior captured by the desired proposition. Unfortunately, they show that the size of a representative sample grows exponentially with the complexity of the behavior being modeled. Therefore, given a model that produces a set of reasonably complex behaviors, the problem of deciding whether the given model is the correct one is *intractable*, if we do not make assumptions about what processes underlie the modeled behavior and test for those properties.

Putnam (1980) pointed out that behavior is not a perfectly reliable measure of cognitive processes, and that it can be disconnected from internal experience; one can perform all behaviors associated with pain even without actually feeling it, and, likewise, it is possible for there to be people in pain who do not express any behaviors exposing that pain (though they may still give introspective report on such pain). How, then, do we determine whether or not a subject is lying to us by either falsely displaying behavioral traits associated with consciousness or falsely hiding the appropriate expressions when present. Once more, we imagine the case of locked-in syndrome. Block (2004) writes, respecting the related issue of whether behavioral evidence is sufficient to identify *intelligence*, that "human judges may be unfairly chauvinist in rejecting genuinely intelligent machines, and they may be overly liberal in accepting cleverly engineered, mindless machines" (p. 234); that is, humans are not reliable observers for determining the presence of cognitive properties, such as *consciousness*, on the basis of behavior, so long as we accept that false behavior is conceivable. Moreover, Block argue that the sense *in which* an artificial system is performing behavior is important to us. It is relevant to us whether a behavior was elicited as the result of some process of interest or as the result of interaction with a very large input-output table. We cannot accept the sufficiency of behavioral testing as described in Elamrani and Yampolskiy (2019) for, in order to be able to account for any form of consciousness, we want to be able to point to a rule-governed process taken to typify it, that is general and simple enough that it may be realizable and recognizable in artificial organisms. Behaviorist models do not seem to capture reasonable intuitions respecting when and where consciousness should be attributable to machines (Block, 2004; Putnam, 1980), and do not seem to align with cognitive science's interest in studying *cognitive processes* as opposed to just behavior. It may be hard to specify the process

of consciousness with sufficient universality to fairly capture all true cases (i.e., to be necessary and sufficient), but falling back on behavior will get us nowhere, and the consciousness literature has known it will get us nowhere for a long time.

Functional Description is Not Negotiable

As we have seen, there is always a trivial objection to the statement that some observed behavior individuates a mental property in the object of study (Block, 2004). While it is possible to ontologically commit oneself to the position that behavior is all there is to speak about, and that mental properties are reducible to behavioral properties, the philosophical work following after Turing (1950)'s famous test has shown us that such commitments produce conclusions that seem absurd to the intuition. However, cognitive scientists are, as a rule, not ontologically behaviorist. To have preoccupation with cognition as the proximal efficient cause of behavior is to be unreservedly the sort of person Skinner (1974) slagged as a "mentalist" (ch. 1). The cognitive scientist is explicitly interested in studying, not merely the relationship between stimulus and response, but *the structure of the process that mediates them, assumed to transpire in an object we might call the mind* (Boden, 2006, pp. 9-12).

Cognitive science might be *methodologically behaviorist*. They might believe that it is acceptable to refer to mental properties that are ontologically irreducible to behavior, even though behavior is the sole epistemic criterion needed to evidence any property of a subject.

The distinction between methodological and radical behaviorism is sustainable only when contemplating human minds in particular. Methodological behaviorists are committed to an ontology in which cognitive processes are meaningful to discuss only in virtue of methodological behaviorist epistemology, which permits inference about cognitive processes from observed behavior (under condition of some stimulus). When the possibility of such inference breaks down, nothing is left to support the ontological distinction between methodological and radical behaviorism, and methodological behaviorism collapses into radical behaviorism.

Inference about mental processes from behavior is possible only under certain strong assumptions about the mind. It takes the form of counterfactual expressions like: "When the subject is presented stimulus s at time t_0, the subject enters a state on some latent variable, which we will call a mental state m, at time t_1. If the subject is in mental state m at time t_1, then we observe a response with property r at time t_2, and, unless the subject is in m at t_1, it is impossible for r to be produced at t_2; therefore, m can be identified as the cause of r". Expressions of this form can only be constructed if we already know enough about the structure of the object we are describing to say that m must mediate s, r. We must also know that there is no more salient explanation than m relating the two variables over the interval t_0, t_2. However, it is always possible to imagine that s, r might be related by an input-output table that simply records behavioral properties describing an association between stimulus and response $B(s, r)$ for all possible stimuli, and retrieves the appropriate response when probed with the appropriate stimulus. We are able to make inferences about mental properties only if we have precommitted ourselves to the position that what we are studying cannot be an input-output table, or any similarly non-mental association between our variables, that we may assume it has m whenever B is observed.

Moore (1989) writes, "there is an appeal to metaphysical, a priori criteria of meaning and knowledge implicit in ... methodological behaviorism" (p. 22). More generously, the a priori criteria of methodological behaviorists might be interpreted as programmatic (Lakatos, 1978); they comprise part of the hard core of assumptions that must be taken as given (even if they do not arise directly from empirical data) and immune to falsification, in order to interpret observation to confirm theory. Every theoretician implicitly uses constraining assumptions of this type, and they are fine to have; we may hold them so long as they are useful to expanding what empirical content science can explain, and so long as they are not surpassed by a more progressive set of assumptions.

However, the assumptions appropriate to studying humans, that support methodological behaviorism, are not the assumptions appropriate to studying everything with behavior that resembles the human. The input-output table style of objection is intended for exactly the case of artificial intelligence. For instance, Bender, Gebru, McMillan-Major, and Shmitchell (2021) write, language models "do not have access to meaning" (p. 615), because they are only trained to formally recapitulate text. They can engage in the behavior of language use by mere association of inputs and outputs, without need for understanding, echoing Searle (1980). Importantly, however, Bender et al. insist that modern language models definitely lack understanding, and in this it may be objectionable; our argument only requires that they *might*. So long as it is possible that a simulated behavior merely appears to exhibit a property irreducible to behavior, behavior is not a sufficient condition to individuate that property. The input-output table style of objection expresses why, even when we can use behavior to make inferences about human mental states, the potentially arbitrary structure of computer programs forbids us from making similar inferences about artificial intelligence. Although we can generally assume that the human cognition mediating most behaviors occurs in grossly the same way, a computer program might achieve an observed behavior in *any* way, so we cannot assume they necessarily enter the same mediating states as humans do, such as states of consciousness, while engaged in that behavior. Without the ability to make inferences about mental states from behavior, methodological behaviorism collapses to radical behaviorism: there is no epistemic justification for cognitivist theoretical claims, and so the cognitivist ontology is no longer epistemically distinct from a radically behaviorist ontology. If cognitive science is concerned, not just with human minds in particular, but with cognition, with all of the processes of which a mind

may consist, no matter where they be found in nature, then cognitive science cannot be methodologically behaviorist.

So, what is to be done about the loss of methodological behaviorism? The only way forward is to examine the nature of the assumptions that made the methodological behaviorist a cognitivist. To do so, we will borrow the distinction between functional and behavioral properties found in Tomkins-Flanagan (2025, ch. 5). We have used behavioral properties above: a behavioral property describes the relationship between some stimuli and some responses of some object of study in time. Functional properties, on the other hand, describe the causal structure mediating behavior. We can imagine measuring the activity of the brain while the subject is engaged in some behavior of interest (as in the case of, e.g., fMRI), and we may observe how the prior state of the brain and stimulus come to determine behavior: by examining the causal relationship between the stimulus, the prior state, interstitial points of measure at the times leading up to the response, and the response. On first analysis, we are given a sort of functionalism similar to the one at work in systems or cognitive neuroscience. The causal relationships between points of measure across the time interval at which a causal interaction may take place (or between a point and its prior state) are described by the theorist as mathematical functions, and are integrated to calculate a prediction of its posterior state. If the posterior state can be calculated (i.e., computed), the functions proposed to mediate prior states, inputs, and posterior states are by definition computable functions. But, a causal circuit of these points of measure can be partitioned from the whole causal structure and abstracted over, forgetting the particulars of the brain that actually enacts the functions describing the circuit. The circuit can then be examined for its behavioral properties, which establishes an equivalence class of functions in time that might mediate its inputs and outputs, typified by the shortest-length function that has the computational effects given by the class. We may thusly speak of minds as abstract to brains, with a functional description that is relatively autonomous, if not independent, from the neural substrate.

Functional properties are just behavioral properties of causal circuits as they play a causal role in the overall causal structure that goes on to produce external behavior. Functional properties are, in a physical sense, just as public as any external behavior. But, as functional properties are not uniquely individuated by a subject's external behavior, we may require narrower constraint than is permitted just with reference to external behavior. Namely, that the behavior of the internal structure of a subject conspires such that they enter a mediating state of interest in the course of an external behavior, such as a state of consciousness. We can now see that, in methodological behaviorism, we assumed that the human mind must have a priori a certain causal structure linking its mental states, in order to infer cognitive process from behavior. The causal structure methodolgical behaviorism held humans to have was not totally constrained (else there would be little room for science), but it was well-defined enough that we could rule out that humans might be input-output tables. The well-definedness we rely on to make methodological behaviorism legitimate breaks down, however, when we are concerned with subjects whose minds might be quite alien to ours; for example, artificial intelligence.

Cognitive Science is Not Behaviorist

If consciousness cannot be a property of a very large input-output table (Block, 2004), then it cannot be a behavioral property, and we must move past behaviorism (radical or methodological) if we are to study minds beyond just the human. In order to move past behaviorism, we must apply constraints on what functional properties our object of study might have, so that we can meaningfully evaluate arbitrary objects of study to have those properties. We cannot expect any mental properties we want to spring spontaneously out of the data (van Rooij et al., 2024). We have to decide, precisely, before we go looking, what it is we are looking for, if we are to have any hope of finding it. There is nothing whatsoever wrong with operationalizing mental properties in this way, it is just to make a theoretical commitment whose usefulness will be tested by the productivity of scientific practice.

When concerned with mental properties like consciousness, the theoretical commitments we are speaking about are commitments on the functions and their organization that characterize consciousness. We have to have some idea of what we expect the causal structure that describes consciousness to be. If the structure we propose is too weakly constraining, then, as with input-output table objections, the structure will then fail a test of our scientific intuitions, and our definition of consciousness will need to be narrowed until it becomes sufficient to satisfy our intuitions about what is meant when we say "consciousness". If the proposed causal structure cannot be measured in humans, then it will fail the empirical test, and our intuitions will need to be adjusted. If the definition is too narrowly anthropomorphic, then it is too narrow in virtue that it fails, intuitively, to span the necessary conditions of consciousness. If the proposed definition is too narrow, it should be loosened only until such a time as it satisfies our intuition that it spans necessity, but remains sufficient to consciousness. If we cannot find ourselves satisfied that our definition is both necessary and sufficient to identify consciousness (and only consciousness), then we will once again need to adjust our intuitions until they can be satisfied. This procedure of definition, intuitive testing, redefinition, and intuitive reshaping, is perfectly ordinary, and in fact very good scientific practice. We should not be vulgar empiricists, expecting consciousness to emerge without theoretical work; we should make the effort to define our terms, look for them to be satisfied in nature, and, when they have run out of use, adjust course, and try again. It is because we are proper scientists, because we need theory, that we should aim to propose, test, and revise, functional theories of consciousness.

References

Aru, J., Larkum, M. E., & Shine, J. M. (2023, December). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences*, *46*(12), 1008–1017. doi: 10.1016/j.tins.2023.09.009

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (p. 610–623). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3442188.3445922

Block, N. (1995, June). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. doi: 10.1017/S0140525X00038188

Block, N. (2004, June). Psychologism and behaviorism. In *The turing test: Verbal behavior as the hallmark of intelligence.* The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/6928.003.0032 doi: 10.7551/mitpress/6928.003.0032

Boden, M. A. (2006). *Mind as machine*. Oxford university press.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness.* Retrieved from https://arxiv.org/abs/2308.08708

Carnap, R. (1959). *Introduction to semantics and formalization of logic*. Harvard University Press.

Elamrani, A., & Yampolskiy, R. V. (2019). Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*(5).

Harnad, S., & Scherzer, P. (2008, October). First, scale up to the robotic Turing test, then worry about feeling. *Artificial Intelligence in Medicine*, *44*(2), 83–89. doi: 10.1016/j.artmed.2008.08.008

Hawthorne, J. (2024). Inductive Logic. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2024 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2024/entries/logic-inductive/.

Hildt, E. (2023, April). The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns. *AJOB Neuroscience*, *14*(2), 58–71. doi: 10.1080/21507740.2022.2148773

Lakatos, I. (1978). *The methodology of scientific research programmes* (Vol. 1; J. Worrall & G. Currie, Eds.). Cambridge University Press.

Lau, H. C., & Passingham, R. E. (2006, December). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, *103*(49), 18763–18768. doi: 10.1073/pnas.0607716103

Leslie, J. C. (2015). Consciousness from the standpoint of behaviour analysis. *European Journal of Behavior Analysis*, *16*(2), 147–162. Retrieved from https://doi.org/10.1080/15021149.2015.1083705 doi: 10.1080/15021149.2015.1083705

Miller, G. A. (2003, March). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144. doi: 10.1016/S1364-6613(03)00029-9

Moore, J. (1989). Why methodological behaviorism is mentalistic. *Theoretical & Philosophical Psychology*. doi: 10.1037/h0091470

Moore, J. (2001, December). On Distinguishing Methodological from Radical Behaviorism. *European Journal of Behavior Analysis*, *2*(2), 221–244. doi: 10.1080/15021149.2001.11434196

Moore, J. (2009). Why the radical behaviorist conception of private events is interesting, relevant, and important. *Behaviour and Philosophy*, *37*, 21–37.

Nagel, T. (1974). What is it like to be a bat? *The philosophical review*, *83*(4), 435–450. doi: 10.2307/2183914

Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs, N.J: Prentice Hall.

Palminteri, S., & Wu, C. M. (2025, January). *A Cognitive Scientist's guide to Machine Consciousness.* PsyArXiv. doi: 10.31234/osf.io/s7ptu

Putnam, H. (1980). "Brains and behavior.". *Harvard University Press*, 24–36.

Reggia, J. A. (2013, August). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131. doi: 10.1016/j.neunet.2013.03.011

Robinson, W. S. (2010). Epiphenomenalism. *WIREs Cognitive Science*, *1*(4), 539-547. Retrieved from https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.19 doi: 10.1002/wcs.19

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. doi: 10.1017/S0140525X00005756

Skinner, B. F. (1950). ARE THEORIES OF LEARNING NECESSARY? *Psychological Review*, *57*(4), 193–216.

Skinner, B. F. (1969). *Contingencies of Reinforcement*.

Skinner, B. F. (1974). *About behaviorism*. Alfred A. Knopf, Inc.

Tomkins-Flanagan, E. (2025). *The measure of a mind: Coming to know in an integrative cognitive science*. Unpublished master's thesis, Carleton University.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460. Retrieved from http://cogprints.org/499/ (One of the most influential papers in the history of the cognitive sciences: http://cogsci.umn.edu/millennium/final.html)

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024, 9 27). Reclaiming ai as a theoretical tool for cognitive science. *Computational Brain & Behavior*. Retrieved from 10.1007/s42113-024-00217-5 doi: 10.1007/s42113-024-00217-5

Watson, J. B. (1913). PSYCHOLOGY AS THE BEHAVIORIST VIEWS IT. *Psychological Review*, *20*(2), 158–

177.

Watson, J. B. (1926, October). Behaviourism: A Psychology Based on Reflex-action. *Philosophy*, *1*(4), 454–466. doi: 10.1017/S003181910002581X

Wu, S.-J., Nicolaou, N., & Bogdan, M. (2020). Consciousness detection in a complete locked-in syndrome patient through multiscale approach analysis. *Entropy*, *22*(12). Retrieved from https://www.mdpi.com/1099-4300/22/12/1411 doi: 10.3390/e22121411