Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed

David Billy Udell
Philosophy Program
The Graduate Center, City University of New York
New York, NY 10016-4309
USA
dudell@gradcenter.cuny.edu

Eric Schwitzgebel
Department of Philosophy
University of California at Riverside
Riverside, CA 92521-0201
USA
eschwitz@ucr.edu

August 28, 2020

Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed

Abstract: Susan Schneider (2019) has proposed two new tests for consciousness in AI (artificial intelligence) systems, the AI Consciousness Test and the Chip Test. On their face, the two tests seem to have the virtue of proving satisfactory to a wide range of consciousness theorists holding divergent theoretical positions, rather than narrowly relying on the truth of any particular theory of consciousness. Unfortunately, both tests are undermined in having an 'audience problem': Those theorists with the kind of architectural worries that motivate the need for such tests should, on similar grounds, doubt that the tests establish the existence of genuine consciousness in the AI in question. Nonetheless, the proposed tests constitute progress, as they could find use by some theorists holding fitting views about consciousness and perhaps in conjunction with other tests for AI consciousness.

Word Count: approx. 9200 words

Keywords: artificial intelligence, consciousness, metaphysics, Turing Test

Susan Schneider’s Proposed Tests for AI Consciousness: Promising but Flawed

1. Introduction.

We might someday build conscious machines. But how will we know whether they are conscious? There are two ways we might find out: Either develop the correct theory of consciousness (or a theory close enough to it) and see if the machines fit the bill, or apply what we call a *neutral test* for machine consciousness. A neutral test is any procedure that can reveal whether an entity is conscious to the satisfaction of conflicting theorists of consciousness, among which the test remains neutral. Theory neutrality is, of course, a second-best prize to having the correct theory of consciousness and then simply applying that to the case of artificial intelligence (AI). But given the lack of theoretical consensus about consciousness, a neutral test satisfactory to many theorists is prize enough. In her recent book, Susan Schneider (2019) aims at such a prize.

It’s too much to hope for a *completely* neutral test valid across all possible theories of consciousness, from panpsychism (Strawson 2006; Goff 2017) to theistic substance dualism (Swinburne 2007) to corticothalamic oscillations (Crick and Koch 1990) to Higher-Order Thought theories (Rosenthal 2005). Some approaches to consciousness are going to conflict too completely with others and share too little common ground. The virtue of theory neutrality comes in degrees, relative to a range of contenders. What we can hope for is a *relatively* neutral test that works well enough across a broad range of potentially viable theories. Since the correct theory of consciousness may be developed after the first putatively conscious machine, it would be useful if in the meantime we could find a valid, relatively neutral test (see also Shevlin 2020 on ‘ecumenical heuristics’ for artificial consciousness).

The most famous of the neutral tests for consciousness is the Turing Test (Turing 1950) — originally intended as a test for 'thinking' but sometimes adapted to test for consciousness (Harnad 2003). A machine passes the Turing Test if it can verbally interact with a human judge in a way indistinguishable from how humans interact. It such a case it will be judged to be conscious. Failing the test does not, however, imply lack of consciousness. Dogs and infants fail. Thus, it is a *sufficiency test* for AI consciousness (not a necessary and sufficient criterion): When a system passes, we have (according to the test) good reason to attribute consciousness to it. Part of the test's attractiveness is its agnosticism about the internal architecture of its test subjects, and so its neutrality among many competing theories of consciousness. As Turing notes, the test is even compatible with some forms of religious dualism (p. 443; cf. Descartes'1649/1991). It might turn out to be true that any entity that passes the Turing Test has to meet certain architectural constraints, but if so, that requirement is a discoverable fact about the universe and not a stipulation built into the test.

Unfortunately, the Turing Test has some serious limitations. First, although the best current chatbots cannot pass an unconstrained Turing Test and are still easily distinguishable from human in many areas of ability when carefully probed by a skilled interrogator, some have succeeded in highly constrained versions of the Test such as the one used in the Loebner Competition (see Dennett 1998; Floridi, Taddeo and Turilli 2009; Aaronson 2014 for a critical perspective on chatbot successes), and open-domain chatbots (Adiwardana et al. 2020; Roller et al. 2020; Rosset 2020) and related machine learning models in natural language processing (Liu et al. 2019; Raffel et al. 2020; Brown et al. 2020) are improving rapidly. It is thus unclear whether a sufficiently good chatbot might someday pass a rigorous Turing Test despite lacking consciousness according to many leading theories (in which case the Test would not be an appealingly theory neutral test of consciousness). Second, although the Turing Test is neutral on the question of what interior mechanisms must be present for consciousness (though see French 1990), since whatever system outputs the right behavior passes, it is decidedly *not* neutral across theories that treat details about interior mechanisms as crucial to the presence or absence of consciousness in the sense of having 'what-it's-like'-ness or a stream of experience (for definitions, see Block 1995; Chalmers 1996; Schwitzgebel 2016). The concerns about constitution famously presented in Block's 'Chinese nation' and Searle's 'Chinese room' thought experiments, for example, are disregarded by those who would employ a Turing Test-passing standard for AI consciousness (Block 1978/2007; Searle 1980).

Susan Schneider (2019) has proposed two new tests for AI consciousness that promise to compensate for those limitations. One, the AI Consciousness Test (ACT), developed collaboratively with Edwin Turner, is similar to the Turing Test in that it focuses on verbal behavior — but verbal behavior specifically concerning the metaphysics of consciousness, and under a specific set of learning constraints that aim to prevent too easy a pass. Schneider and Turner's core idea is that if a machine, without being taught to do so, begins to speculate on philosophical questions about consciousness, such as the possible existence of a soul that continues on after bodily death, that is a sign that the machine is conscious (see also Sloman 2007; Argonov 2014). This enriches the Turing Test by shifting its focus to a topic that, under certain constraints we will explain below, the machine is expected to handle poorly unless it is actually conscious. (For other adaptations of the Turing-Test-like procedures to potentially difficult, consciousness-revealing topics, see Dennett 1991 on humor; Levesque 2011 on pronoun disambiguation; Tononi and Koch 2011 on picture gist.)

Schneider’s second new test, the Chip Test, is different from most previous proposals and has the advantage of being plausibly neutral among a wider range of theories of consciousness. The Chip Test depends on and is concerned with the existence of removable artificial ‘brain chips’ made of silicon. One temporarily suppresses one’s biological brain activity in a certain functional or physiological region, relying instead on the chip to do the cognitive work. Simultaneously, one introspects. Is the relevant type of consciousness still present? If so, one concludes that the silicon chip is capable of supporting conscious experience. This test is motivated by Searlean and Blockian concerns about the possibility that silicon (or other kinds of artificial) chips might be incapable of hosting consciousness, even if they are functionally similar at a relatively high level of description, supporting similar outward behavior. It is explicitly first-personal, relying on the tester’s own introspection to address those concerns rather than relying just on facts about outward behavior or interior structure. Like the Turing Test and Schneider and Turner’s ACT, this is a *sufficiency test* of consciousness rather than a necessity test: Passing the test is held to be sufficient evidence for a confident attribution of consciousness to the target system, while failing the test does not guarantee a lack of consciousness (there might, for instance, be a faulty connection when wiring up well-functioning chips).

AI consciousness, despite its present science-fictional air, may soon become an urgent practical issue. Within the next few decades, engineers might develop AI systems that some people, rightly or wrongly, claim have conscious experiences like ours. We will then face the question of whether such AI systems would deserve moral consideration akin to that we give to people. There is already an emerging ‘robot rights’ movement which would surely be energized by plausible claims of robot consciousness (Schwitzgebel and Garza 2015; Gunkel 2018; Ziesche and Yampolskiy 2019). So we need to think seriously in advance about how to test for consciousness among apparently conscious machines, and Schneider's new proposed tests are, in our judgment, among the most interesting relatively theory neutral tests out there (see Elamrani and Yampolskiy 2019 for a review of recent AI consciousness tests). Especially given the uncertain state of general theories of consciousness and thus of any tests that rely on particular theories, Schneider's proposals deserve a closer look.

Unfortunately, we argue, Schneider's tests are flawed — both for the same basic reason. Both tests are motivated by a certain sort of worry about the importance of interior structure: the worry that something that appears from the outside to be conscious might not genuinely be if it doesn't have the right sort of interior processes. But both tests employ empirical methods that most theorists who take such architectural worries seriously ought to be skeptical of. The tests thus have an *audience problem:* If a theorist is sufficiently skeptical about outward appearances of seeming AI consciousness to want to employ one of these tests, that theorist should also be worried that a system might pass the test without being conscious. Generally speaking, liberals about attributing AI consciousness will reasonably regard such stringent tests as unnecessary, while skeptics about AI consciousness will doubt that the tests are sufficiently stringent to demonstrate what they claim.

2. How the AI Consciousness Test Works.

Schneider and Turner's AI Consciousness Test (ACT) imagines a form of AI very unlike a human being. The AI subject is not one that looks like any human being and does not pass for human. The subject of the ACT is a sophisticated, highly intelligent AI, implemented in a supercomputer and without any sort of external, mobile body. This machine, by Schneider and Turner's hypothesis, might outperform human beings along all sorts of capability dimensions —

medical diagnosis, stock market forecasting, generating prize-winning poetry — while programmers and users legitimately wonder whether it has genuine consciousness. The ACT is designed to reveal whether these kinds of machines are conscious.

In the ACT, our one line of communication with this AI is a single chat terminal, or some other means of communicating with the system while still maintaining 'a secure, simulated environment' (Schneider 2019, p. 54). Our AI test subject lacks any physical body with which to interact with the outside world, apart from this chat terminal. It is a 'boxed' AI (Schneider 2019, pp. 53-54). The AI is assumed to be capable of linguistic communication and so it is a suitable target for interrogation.

Every human adult has an introspective familiarity with consciousness, and we report as much in our talk about our mental lives. Schneider notes that 'most adults can quickly and readily grasp concepts based on the quality of felt consciousness,' such as life after death, out of body experiences, and body swapping (2019, p. 51). It is crucial to the validity of Schneider's test that people find these concepts natural and intuitive, understanding them easily with little to no training — as suggested, for example, by the fact that even young children have no difficulty understanding movies about such topics, such as the popular body-swap film *Freaky Friday* (Schneider 2019, p. 51). It's also crucial that possessing these concepts requires having consciousness. Let's cede both points for the time being. These are fairly substantial philosophical concessions, but accepting them will allow us to focus on a more central concern about the ACT.

The ACT is a battery of natural language questions that the AI must quickly and appropriately answer. The topic of questioning is consciousness-derivative concepts like body-swapping. If the AI can readily talk our folk-psychological and folk-philosophical talk about consciousness, then it is, according to the test, conscious. The form of the argument here seems to be inference to the best explanation. We can *best explain* the AI's fluency with such concepts by assuming that it is drawing, as Schneider presumes humans do, upon an introspective familiarity with consciousness. If it lacked introspective familiarity with consciousness, the test's implicit reasoning goes, the AI would be detectably baffled, awkward, or off-key in its reflections about out-of-body experiences, substance dualism, and life after death.

To forestall an immediately obvious objection, the ACT stipulates some facts about the AI's learning context. Since Schneider speculates that the AI might be a highly capable learner, she and Turner have designed the test to guard against the possibility that it might just be telling us what we want to hear. If not appropriately constrained, the AI might succeed, for example, by simply regurgitating and reworking overheard human answers to questions on these topics (witnessed among researchers, or online, or in its training dataset). The test thus requires that inside the 'box' the AI is given no access to the outside world — a restriction not normally present in standard versions of the Turing Test. The AI's developers must 'deny the AI access to the Internet and prohibit it from gaining too much knowledge of the world, especially information about consciousness and neuroscience' (Schneider 2019, pp. 53-54). If the AI subject has no means to obtain information about consciousness other than through its own introspective familiarity, then according to Schneider, passing the ACT is sufficient to establish that the AI is conscious (see also Floridi 2005).

Another important detail regards teaching the AI enough for it to be able to interpret the questions without teaching it so much that it can fake its way through the test. If an ACT prompt question were, for example, to contain the word 'ghost,' the system would have to be provided beforehand with whatever resources or definitions are necessary for it to understand the word, and that might be enough to spoil the test, especially given the AI's stipulated intellectual sophistication. Schneider notes this worry, and suggests ACT sample questions that avoid spoiling the game with 'expressions like 'consciousness,' 'soul,' and 'mind'' (2019, p. 54). But even employing more neutral language like 'Could you survive the permanent deletion of your program? What if you learned this would occur?' (ACT Sample Question 1, p. 55) requires that the AI understand what 'survival' is, and presumably its opposite. The AI would also need a repertoire of terms through which it could appropriately express the reactions we might expect from it if it were conscious (fear or sadness, perhaps?). The AI can't be entirely prevented from learning about consciousness-adjacent aspects of the world if it's going to be a conversational partner in a test of this sort. But it's unclear how much world knowledge would be too much, in the hands of a sophisticated learning AI. Implementation of the ACT will therefore require careful aim at an epistemic sweet spot.

The issue of how much knowledge might be built into an ACT test subject could be spelled out in a few different concrete ways — Schneider's own exposition is noncommittal. Imagine, for example, that one way of developing AI general intelligence involves a process of growing the AI over time from a simple system (with little in the way of skills or knowledge) into a massive, complex system (with many skills and extensive knowledge). It would be 'unfair' to the AI to administer the ACT too early in its development — the AI would necessarily fail, just as human infants would fail, despite possibly being or later becoming conscious. For the test to have merit, it must capture the AI at the right point in its cognitive development: far enough along to reveal its developed capacities, but not so far along that, given extensive knowledge about consciousness or consciousness-adjacent topics, it can 'cheat.' Analogous considerations apply to constructed AIs. They must be fully constructed, with enough installed knowledge to be suitable for testing, but not *too much* knowledge. The AI must have sufficient general intelligence to converse fluently about topics unrelated to consciousness while remaining insulated enough from consciousness-specific information that, unless it is actually conscious, the AI will react inappropriately to questions about consciousness. It must be able to intelligently discuss a popular film about, say, swapping jobs, while conspicuously flubbing when asked to discuss the body-swapping of *Freaky Friday*. We will assume in our discussion of the ACT that this epistemic sweet spot for viable test subjects can be found.

With the ACT, Schneider and Turner have proposed a means to examine sophisticated artificial intelligence for consciousness in a way that is (1) impressively neutral about architectural details, (2) consistent with nearly complete human and AI ignorance about the physical or functional basis of consciousness, (3) allows (unlike the Turing Test) that the AI might be radically cognitively different from us in important respects that it cannot conceal, and (4) is even consistent with non-materialist views like substance dualism. In these respects it is an ambitious and intriguing test.

3. Critique of the AI Consciousness Test.

The fundamental problem with the ACT is this: The very same concerns that generate its worry invalidate its method. It begins with an architectural worry — that an AI exposed to human knowledge of consciousness might trick us into thinking it is conscious by outwardly mimicking our language about it *without* possessing the internal structures necessary for consciousness. It then attempts to address this worry with an empirical test, the ACT. But few theorists inclined to endorse the initial worry should accept the sufficiency of the empirical test.

Skeptics about AI consciousness will remain able to appeal to non-conscious mechanistic explanations to explain AI success in the ACT.

As we suggested in Section 2, the ACT appears to be justified by inference to the best explanation. If the AI were to pass the test (without cheating), the best explanation would have to be that it really did have some sort of introspective familiarity with a stream of conscious experience. The problem with this argument is that there might be a competing explanation that is equally good — an explanation solely in terms of lower-level or design-level physical or functional features disconnected from consciousness.

Consider a human being who passes the ACT. At least in principle, it seems possible to explain her passing the test by appeal to lower-level physical or functional features of her brain and cognition. For example, we might imagine that she takes the ACT under the supervision of a team of crack 22nd century neuroscientists. They might be able to explain her responses in neurophysiological or functional terms, without explicit reference to the concept of consciousness. But this would not, presumably, *compete* with an explanation in terms of consciousness, since the very processes they describe in detail (40 hertz oscillations in such-and-such an area, self-monitoring structures of such-and-such a sort, or whatever) would be intimately connected with the subject's consciousness. The high-level and low-level explanations might both be excellent and non-competing, tapping into the same general phenomena, just conceptualized in different ways.

For any AI system that passes the ACT, there will in principle be some lower-level or functional explanation that explains its passing. The question to consider is whether such an explanation might be an explanation that competes with explanation in terms of introspective familiarity with consciousness. The lower-level or functional explanation does not compete if, as presumably in the human case, it adverts to the same processes that give rise to consciousness, described in different terms. Alternatively, the lower-level or functional explanation does compete if it adverts to processes compatible with the nonexistence of consciousness.

The worry that motivates the test is what we will call an *architectural worry* about AI consciousness, concentrating on structure at this lower or functional level. It's the worry that there could be some ways of structuring a sophisticated AI such that it can make a convincing outward show of consciousness without being conscious. A convincing 'outward show' in this sense need not involve the capacity to behave exactly like a human being. It need only involve behavior that is sufficiently sophisticated that well-informed people start to reasonably suspect that the system might be conscious. A theorist with the type of architectural worry we have in mind might, for example, hold that if an outwardly sophisticated AI works by means of a single giant lookup table, such a structure would be sufficiently unlike what we see in conscious human beings and animals that we should not conclude that it is conscious. (We and Schneider bracket the metaphysical question of whether, in principle, a perfect behavioral duplicate made of different materials would necessarily be conscious.) To feel the need for the test, the doubter of AI consciousness must worry that some lower-level physical or functional explanations of sophisticated outward behavior would be explanations that compete with an explanation in terms of the AI's genuinely being conscious. Absent this worry, the grounds for doubt disappear and there's no need for a test.

Suppose now that the AI passes the ACT. We have a candidate high-level explanation: It is conscious. Call this Explanation C. There will be in principle at least one discoverable lower-level explanation: It works mechanically like thus-and-so. Call this Explanation M. Note that for inference to the best explanation to be sound, no sufficiently good competing explanation must be possible, not only now but also in the future, so Explanation M can remain hypothetical. To have confidence in the test, we must be confident that an excellent potential lower-level Explanation M would not compete with Explanation C.

We submit that most theorists who, before the test, take the architectural worry seriously and thus see the need for a test should similarly worry that there might be a good Explanation M that competes with Explanation C. For example, suppose you doubt that silicon is the right sort of material to host consciousness, despite its (let's assume) ability to host an AI that, if unboxed, makes an excellent outward show of being conscious. Or suppose you doubt that the if-then programming structure of classical AI is the right sort of structure to host consciousness, despite (let's assume) the ability of some classically structured AIs to make an excellent outward show of consciousness. In either case, your worry is that for the unboxed AI there is some lower-level or functional Explanation M for the AI's consciousness-like behavior that competes with an explanation in terms of genuine AI consciousness.

Now box this AI and train it enough that the ACT can be fairly applied, but no more than that — train it (or program it, or select it evolutionarily) right to the epistemic sweet spot we discussed in Section 2. Barring some architecturally inexplicable event, such as a miracle, it's safe to assume that there will be some in principle discoverable (if unknown to us) lower-level or functional Explanation M of the machine's passing the ACT — some blow-by-blow architectural story about how it generated the answers it did. Now ask yourself: Could Explanation M be an explanation that competes with rather than complements an explanation in terms of consciousness?

Ex hypothesi, the system is such that it could fake consciousness if unboxed. Ex hypothesi, the system achieves this by means of an underlying structure that we worry might not give rise to consciousness. Ex hypothesi, we equip the system with everything it needs to understand the questions constituting the ACT (e.g., the concepts of survival and death) and everything it needs to answer appropriately if it does happen to be conscious (e.g., the concept of sadness). By lack of miracles, the system's responses to the ACT will be some further implementation of its ordinary architectural processes. It might, for example, if given ACT Sample Question 1, do some complicated version of mechanically associating lack of survival with death and death with sadness, then answer that it would be sad to learn it would be deleted. Few AI consciousness doubters, we venture, ought to abandon their motivating architectural worry upon consideration of this situation — unless, perhaps, their motivating architectural worry is highly specific and of just the right sort. *For most AI consciousness doubters*, the underlying inference to the best explanation should not remove their doubt. This is the *audience problem*. Most theorists who have the kind of architectural worry that motivates seeking out a test like the ACT should, upon seeing that an AI passes the test, remain worried, rather than confidently concluding that the AI subject is conscious.

Of course, if we knew what kinds of structures were and were not genuinely capable of hosting consciousness, we could peek inside the AI and see if it had the right kind. But that's not a theory neutral test, and that's not the ACT.

We have argued, contra Schneider, that passing the ACT should not be regarded as sufficient to establish the presence of consciousness for most theorists who enter the test with the types of architectural doubts that motivate the existence of the test. Nonetheless, the ACT could have substantial value. For example, questions of this sort could be added to an enhanced version of the Turing Test, and perhaps they will prove especially difficult or diagnostic. Or perhaps a non-neutral test could be devised that employs the ACT, and examination of the specific internal mechanisms by which the AI answers the ACT's questions could be revealing. If, on review, the AI's answers all prove to stem from simple chat-bot-ish associative networks, that datum might be important to theorists who hold that consciousness requires more than simple chat-bot-ish associative networks. If, on the other hand, a review of the AI's mechanism for generating answers reveals something sharing plausible affinities with human introspection, theorists with the right background views might be moved toward a less skeptical position. We see considerable potential value in the test's box-and-ask approach; we only doubt that it can serve the ambitious theory-neutral diagnostic role that Schneider suggests.

4. How the Chip Test Works.

Schneider's second proposed test, the Chip Test, deals with a fundamentally different type of AI than the ACT does. The Chip Test concerns silicon-based neural prostheses and whole brain replacement, or 'brain uploading' (Bostrom 2014, pp. 35-43). The motivating worry is that replacing one's brain with silicon chips might render a person nonconscious, even if the resulting entity's outward behavior remains similar to the original person's behavior, even if it emits sounds like 'Yes, I'm still conscious,' and even if the person's family and friends are entirely convinced that nothing crucial has changed.

Schneider is concerned that silicon brain chips might be the wrong type of stuff to host consciousness but might nonetheless convincingly emulate human brain capacities (2019, p. 7). Converting one's brain to silicon — whole brain uploading — might have tremendous seeming-advantages in terms of durability, backup capacity, processing capacity, and feature management (Egan 1997; Kurzweil 2005; Chalmers 2010; Hanson 2016). But many people would reasonably regard it as a form of suicide if the resulting entities lacked consciousness entirely or lacked the type of consciousness that many regard as central to the value of living (Humphrey 2006; Kriegel 2019). Before converting our brains to silicon or any other type of artificial replacement, it's reasonable to want confirmation that the resulting system would actually have conscious experiences of the right sort. The motivating architectural worry that drives the test is the worry that accurately emulating the informational processes of the brain with enough fidelity to fool an outsider might be consistent with the absence of consciousness.

Schneider motivates this architectural worry in two ways. First, she notes the chemical differences between carbon and silicon and their significant impact on organic chemistry. Schneider suggests on these grounds that 'if chemical differences between carbon and silicon impact life itself, we should not rule out the possibility that these chemical differences could also impact whether silicon gives rise to consciousness, even if they do not hinder silicon's ability to process information in a superior manner' (2018, p. 315; see also 2019, pp. 19, 47). Second, Schneider notes that much human cognitive processing is nonconscious, including (she holds) the sophisticated functional processes underlying highly practiced skills such as driving a vehicle or converting the information in an incoming stream of light into a three-dimensional scene with objects (2019, p. 35). The specific danger of silicon brain chips, on Schneider's conception, is their potential to turn us into 'zombies' who outwardly act like conscious people but who lack conscious experience — a generalized version of replacing everything in our conscious experience with the kind of nonconscious processing that the human visual system engages in when it extracts an object's shape from a two-dimensional array of light. If we take these worries seriously, we cannot trust the seemingly introspective reports of our uploaded 'friends.' Our seeming friends might merely be nonconscious entities who outwardly behave in humanlike ways. Zombie reports of conscious experience will always be false positives.

The Chip Test addresses this worry with a specific set of guidelines for neuron-to-chip substitution, or brain uploading, that will reveal whether the technology in question can or can't support consciousness in the way that carbon-based neurons do. One implements the Chip Test by uploading slowly, one chip to its corresponding brain part at a time. When prompted, the subject introspects: Am I as conscious as before, or has some part of my phenomenal field been lost? If they are as conscious as before, they report their introspective judgment that the chip works, phenomenologically speaking, and the upload continues. If something has changed and their phenomenal field has been compromised in some way, they introspectively notice the bad result, report the problem, and halt the process. The technology in question will have been introspectively discovered to compromise consciousness, even if it otherwise functions quite well. The Chip Test fully succeeds if the brain is completely replaced by silicon chips with no introspectively detected loss of consciousness.

While Schneider holds that we shouldn't trust the seemingly introspective reports of fully siliconized 'people,' she is willing to trust people's judgments after a single brain part replacement. Plausibly, neither intelligence nor consciousness will snap out of existence after the replacement of a single, relatively small brain part and are instead the product of the combined efforts of many parts. Because theories of consciousness can take small brain regions to be of paramount importance, the chips used in the Chip Test are going to have to be smaller than what any included background theory considers to be the crucial size of the brain basis of consciousness. Phenomenological changes, Schneider assumes, would then appear incrementally, and therefore assessably and manageably. If no phenomenological changes are detected during a sufficiently gradual replacement procedure, then it's reasonable to conclude that no phenomenological changes have in fact occurred. Crucial to the validity of the Chip Test is the accuracy of introspective judgment throughout the process.

The Chip Test draws on familiar considerations from Ship-of-Theseus examples, as applied to consciousness and personal identity, in which small parts of the brain are one-by-one swapped out for functionally similar artificial equivalents until the whole brain is replaced (Moravec 1988; Searle 1992; Chalmers 1996, 2010). The Chip Test frames this gradual replacement procedure as a practical, empirical test for the consciousness of whole brain uploads, where success in the Test is taken to be a good reason to confidently believe those uploads to be conscious.

Schneider makes a few qualifications to the above story. First, it is in her view clear that silicon chips will not be microscale functional isomorphs of carbon-based brains regardless of the details of future AI engineering (Schneider and Mandik 2018). That is not what the Chip Test is testing for. Brains and chips will differ at an appropriately small, microfunctional scale. It is an open question, however, whether they can converge on a meso- or macrofunctional scale (see also Godfrey-Smith 2016).

A second qualification is that we can't infer from a lone chip failure that silicon will be forever unable to replace the relevant brain part. The problem might instead lie in the engineering. It would take a stubbornly persistent track record of these substitution failures to warrant blaming the silicon itself rather than the engineers.

Finally, the Chip Test allows that there are some contrived cases that escape it. 'Sham' brain chips might be deliberately engineered to falsely signal to the rest of the brain that they are conscious, cheating their way past the Chip Test. However, Schneider (2016) sees sham chips as a difficult and unlikely prospect. The idea here, it seems, is that if brain chips come about as the result of a good faith engineering project, we should be more confident that they support consciousness than if the chips were to come about instead as the product of deliberately malicious engineering. For the sake of this critique, we will assume that there are no deliberately engineered sham chips (see [Author's article 1] for further discussion).

5. Critique of the Chip Test.

The shortcoming of the Chip Test is this: Someone who accepts Schneider's worry about the untrustworthiness of 'introspection' after brain uploading should not be willing to rely, as the Chip Test does, on the accuracy of introspection during the Test's gradual swapping procedure. This holds across a wide range of theoretical views of introspection, including infallibilist views.

Our criticism of the Chip Test resembles our criticism of the ACT in two respects. Both critiques are conceptual: The critique is not that false positive producing edge cases are possible, but rather that there is a fatal tension internal to the test affecting it in *all* cases. The criticisms are also alike in playing each test's empirical method against its motivating worry: Most theorists who are motivated by the worry that the test is designed to address should not simultaneously accept the validity of the proposed empirical test (though the test might seem adequate to others who are *not* motivated by that worry).

The Chip Test is designed to examine and resolve the worry that consciousness might be lost if we cross over into silicon, a worry often speculated about in connection with whole brain uploading. Both the Chip Test and one-shot whole brain uploading end in the same place: a wholly siliconized brain that outwardly behaves relevantly like the original biological brain. The Chip Test differs primarily in that it insists that the process be gradual, with introspective checking along the way. For both upload procedures, let's denote the entity before brain replacement r_0 and the entity after full replacement r_n, where n is the number of steps in the gradual version of the replacement procedure. During replacement, the entity passes through stages r_1, r_2, … r_n. Both gradual replacement and one-stage whole brain uploading must end in the same place, r_n, in order for the one to be the test of the other. By stipulation, we don't trust the putative introspective judgments of entity r_n. If r_n says, 'Look, I just introspected and yes, I'm conscious just as before!' we have reason to doubt the report. This is the motivating worry of the test. In order for the Chip Test to work, there must be some epistemic advantage that the r_i entities systematically possess over r_n, such that we have reason to trust their introspective reports despite reasonably distrusting r_n's report.

Seemingly introspective reports about conscious experience may or may not be trustworthy even in the normal human case r_0 (Schwitzgebel 2011; Irvine 2013). But even if they are trustworthy in the normal human case, they might not be trustworthy in the unusual case of having pieces of one's brain swapped out. One might hold that introspective judgments are *always* trustworthy (absent a certain range of known defeaters, which we can stipulate aren't present), in other words, that unless a process takes a conscious experience as its target it is not a genuinely introspective process. This is true, for example, on 'containment' views of introspection, according to which properly formed introspective judgments contain the target conscious experiences as a part (e.g., 'I'm experiencing [this]') (Gertler 2001; Papineau 2002; Chalmers 2003; Kriegel 2009).   *Infallibilist* views of introspection of this sort contrast with what we will call *functionalist* views of introspection, on which introspection is a fallible functional process that garners information about a target mental state (Nichols & Stich 2003; Goldman 2006; Hill 2009; Schwitzgebel 2012). An AI consciousness skeptic might accept or reject an infallibilist view of introspection. Our criticism is that the Chip Test faces trouble either way (see table below).

	Skeptics about AI consciousness	Optimists about AI consciousness
Infallibilists about introspection	... *also* doubt whether r_is introspect at all.	… are already sold on r_n consciousness.
Fallibilists about introspection	... *also* doubt whether r_is reliably introspect.	… are already sold on r_n consciousness.

A Trilemma for the Chip Test: Optimists about AI consciousness have no need to test for r_n consciousness, because they are already convinced of its presence. On the other hand, AI consciousness skeptics are led to doubt either the presence or the reliability of r_i introspection (depending on their view of introspection) for the very same reason they are led to doubt r_n consciousness in the first place.

If an AI consciousness skeptic holds that genuine introspection requires and thus implies genuine consciousness, then she will want to say that a zombie r_n, despite emitting what looks from the outside like an introspective report of conscious experience, does not in fact genuinely introspect. With no genuine conscious experience for introspection to target, the report must issue, on this view, from some non-introspective process. This raises the natural question of why she should feel confident that the r_is are genuinely introspecting, instead of merely engaging in a non-introspective process similar to r_n's. After all, there is substantial architectural similarity between r_n and at least the late-stage r_is. The skeptic needs, but Schneider does not provide, some principled reason to think that entities in the r_i phases in fact introspect despite r_n's possible failure to do so — or at least good reason to believe that the r_is successfully introspect during the most crucial stages of the swapping process. Absent this, reasonable doubt about r_n introspection naturally extends out into reasonable doubt about introspection in the r_i cases as well. The infallibilist AI consciousness skeptic needs her skepticism about introspection to be assuaged for at least those critical transition points in the Test before she can accept the Chip Test as informative about r_n consciousness.

If an AI consciousness skeptic instead believes that genuine introspection does not necessarily require genuine consciousness, analogous difficulties still arise. Either r_n does not successfully introspect, merely seeming to do so, in which case the argument of the previous paragraph applies, or r_n does introspect and concludes that its consciousness has not faded or radically changed in any undesirable way. The functionalist or fallibilist AI consciousness skeptic ought not naively trust that r_n has introspected accurately. On his view, r_n might in fact be a zombie or might have a more limited or very different consciousness, despite its introspectively-based claims otherwise. Absent any reason for the fallibilist AI consciousness skeptic to trust r_n's introspective judgements, why should he trust the judgments of the r_is — especially the late-stage r_is? If r_n can mistakenly judge itself conscious, on the basis of its introspection, might those of us taking the Chip Test, or our zombie replacements, also erroneously introspect the presence of consciousness at some point in the swapping procedure? Gradualness is no assurance against error. Indeed, error is sometimes easier if we (or 'we') slowly slide into it. This concern might be mitigated if loss of consciousness is sure to occur early in the swapping process, when we are much closer to r_0 than to r_n, but we see no good reason to make that assumption. And even if we were to assume that phenomenal alterations would occur early on in the Test, it's not clear why the fallibilist AI consciousness skeptic should take those changes to be of the sort that introspection would reliably detect, rather than miss. The Chip Test awkwardly pairs skepticism about r_n's introspective judgments with unexamined confidence in the r_is' introspective judgments, and this pairing isn't theoretically stable on any view of introspection.

This objection can be made vivid with a toy case: Suppose we have an introspection module in the brain. When the module is involved in introspecting a conscious mental state, it will send query signals to other regions of the brain. Getting the right signals back from those other regions — call them regions A, B, and C, perhaps regions of the visual cortex or of the prefrontal cortex — is part of the process driving the judgment that phenomenal changes are present or absent. Now suppose that we replace region B with a silicon chip. Maybe activity in region B is magnetically suppressed while the newly installed silicon chip detects the neural signals that would normally have entered region B, does various computations, and then sends output signals to other brain regions that normally interface with region B. Among those output signals will be signals to the introspection module. When the introspection module sends its query signal to region B, what signal will it receive in return? Ex hypothesi, silicon chips are capable of emulating relevant functional processes of brain regions; that's part of the motivating worry. You can swap them in without loss of function. Given this, and given that the replacement chip is well-designed, we see no reason to think that the introspection module wouldn't or couldn't receive a signal very much like the signal it would have received from region B had region B not been magnetically suppressed. And if so, entity r_i will presumably infer that activity in region B is conscious. Maybe region B normally hosts conscious experiences of thirst. The entity might then say to itself, 'Yes I'm still feeling thirsty. I really am having that conscious experience, despite the fact that the chip is now doing the work usually done by that part of my biological brain.' This would be, as far as the entity could tell, a careful and accurate first-person introspective judgment.

An AI consciousness optimist who does *not* share the architectural worries motivating the Chip Test might be satisfied with that demonstration. But Schneider is concerned that silicon might not host consciousness even if silicon chips can emulate the macro-level information processing of the human brain. Most theorists who share in that worry should remain worried in the face of the Chip Test. They ought to worry that the chip replacing region B is not genuinely hosting consciousness, despite its feeding output to the introspection module that led the entity to conclude that consciousness remains fully present. They ought to worry, in other words, that the introspective process has gone awry. This needn't be a matter of 'sham' chips intentionally designed to fool users. It could be a straightforward engineering consequence of designing a chip to emulate the information processing of a brain region. (If, on the other hand, the brain region containing the introspection module is what is swapped out, then maybe introspection isn't occurring at all, at least in any sense of introspection that is metaphysically committed to the idea that introspection is a conscious process.)

This story relies on a cartoon model of introspection that is unlikely to closely resemble the process of introspection as it actually occurs (see [Author's articles 2]). Our criticism does not require the existence of an actual introspection module or any query process much like the above toy case. Our point is that an analogous story holds for more complex and realistic models. If silicon brain regions functionally emulate biological brain regions, there is good reason for someone with the types of worries that motivate the Chip Test to worry that swapping one for the other might either create inaccuracies of introspection or unpredictably replace the introspective process with whatever non-introspective process even zombies engage in.

The Chip Test thus, seemingly implausibly, combines distrust of the putative introspective judgments of r_n with credulousness about the putative introspective judgments of the series of r_is between r_0 and r_n. An adequate defense of the Chip Test will require careful justification of why someone skeptical about the seemingly introspective judgments of an entity that has had a full brain replacement should not be similarly skeptical about similar seemingly introspective judgments that occur throughout the gradual replacement process. Again, there seems to be an audience problem: optimists about AI consciousness will not see the test as necessary, while skeptics who see the test as necessary probably ought not find it convincing.

As with the ACT, we see value in the Chip Test despite these concerns. Even if passing the test should not be regarded as convincing evidence of consciousness to a theorist motivated by the general architectural worries that drive the test, another group of theorists who regard certain particular features of mesofunctional organization as important — e.g., IIT theorists (Oizumi, Albantakis, and Tononi 2014) — might adopt the test as one way of checking for mesofunctional fidelity in the replacement system. A theorist might worry, for example, that the one-shot whole brain upload results in a macro-level behavioral isomorph that is nonetheless crucially functionally divergent at a meso-level. That behavioral isomorph might, for example, rely heavily on feed-forward mechanisms and differ drastically from human computational architecture when it comes to internal states — a fact that might be hard to detect from outside but which might be revealed during a gradual replacement procedure. This alternative, more modest application of the Chip Test might confirm that the chiphead possesses the right meso-level functional structure. For this purpose, that is, the more demanding replacement procedure of the Chip Test *could* prove more informative than the one-shot whole brain upload. Theorists moved by the Chip Test ought to be relatively rare, however; only specific varieties of architectural worry will be such that the gradual procedure of the Chip Test is more evidentially significant than the working one-shot whole brain upload. Schneider's proposed Chip Test is thus a valuable suggestion, even if neither theory-neutral nor definitive.

6. Conclusion.

Schneider promises something that the field of consciousness studies will possibly soon need: a relatively theory-neutral practical approach to assessing the presence or absence of consciousness in AI systems that might, absent a good testing procedure, mislead us into thinking that they are genuinely conscious. Unfortunately, both of the tests she suggests are flawed. The AI Consciousness Test depends for its validity on the assumption that a boxed AI's language about consciousness will reflect the presence or absence of genuine consciousness, which stands in tension with the general worry about language/consciousness mismatch that motivates the test. The Chip Test depends for its validity on confidence in the accuracy of putative introspective reports during a gradual brain replacement procedure, but introspection cannot be relied on to play the role the test demands. For both of the proposed tests, most theorists who accept the architectural worries that motivate the test (instead of simply accepting outward signs of consciousness at face value) probably should be similarly worried about the validity of the test itself.

Nonetheless, we commend the tests as progress. Even if they are not as theory-neutral or definitive as they might at first seem, they might nevertheless find a range of applications in the right conditions, in conjunction with other tests and contingent upon the acceptance of particular theories of consciousness.

As for developing a viable theory-neutral test, we should not expect this to be easy. Theories differ enormously, with limited common ground. It may be that the AI systems for which a relatively neutral test can be agreed upon are not the controversial cases that most need such a test — though whether that turns out to be so remains to be seen. Absent a neutral test, the only way to adjudicate these unclear cases will be through the application of particular theories of consciousness, and absent a consensus around a leading candidate theory, we might face irresolvable disagreement about which AIs are conscious and which are not, with potentially enormous social and personal implications.

References

Aaronson, Scott. 2014. *My Conversation with "Eugene Goostman," the Chatbot that's All Over the News for Allegedly Passing the Turing Test.* June 9. Accessed August 17, 2020. https://www.scottaaronson.com/blog/?p=1858.

Adiwardana, Daniel, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, et al. 2020. "Towards a Human-like Open-Domain Chatbot." *arXiv preprint.* February 27. arXiv:2001.09977v3 [cs.CL].

Argonov, Victor Yu. 2014. "Experimental Methods for Unraveling the Mind-Body Problem: The Phenomenal Judgement Approach." *Journal of Mind and Behavior* (35): 51-70.

Block, Ned. 1995/2007. "On a Confusion About a Function of Consciousness." In *Consciousness, Function, and Representation,* by Ned Block. Cambridge, MA: MIT.

Block, Ned. 1978/2007. "Troubles with Functionalism." In *Consciousness, Function, and Representation,* by Ned Block. Cambridge, MA: MIT.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." *arXiv preprint.* June 5. arXiv:2005.14165v3 [cs.CL].

Chalmers, David J. 1996. *The Conscious Mind.* Oxford: Oxford University Press.

Chalmers, David J. 2003. "The Content and Epistemology of Phenomenal Belief." In *Consciousness: New Philosophical Perspectives,* edited by Quentin Smith and Aleksandar Jokic, 220-272. Oxford: Oxford University Press.

Chalmers, David J. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* (17 (9-10)): 7-65.

Crick, Francis, and Christof Koch. 1990. "Towards a Neurobiological Theory of Consciousnesss." *Seminars in the Neurosciences* (2): 263-275.

Dennett, Daniel C. 1998. *Brainchildren: Essays on Designing Minds.* Cambridge, MA: MIT.

—. 1991. *Consciousness Explained.* Boston: Little, Brown and Company.

Descartes, Renè. 1649/1991. "To More, 5 February 1649." In *The Philosophical Writings of Descartes*, translated by John Cottingham, Robert Stoothoff, Dugald Murdoch and Anthony Kenny, 360-367. Cambridge: Cambridge University Press.

Egan, Greg. 1997. *Diaspora.* Millenium.

Elamrani, Aïda, and Roman Yampolskiy. 2019. "Reviewing Tests for Machine Consciousness." *Journal of Consciousness Studies* (26 (5-6)): 35-64.

Floridi, Luciano. 2005. "Consciousness, Agents and the Knowledge Game." *Minds and Machines* (15): 415-444.

Floridi, Luciano, Mariarosaria Taddeo, and Matteo Turilli. 2009. "Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest." *Minds and Machines* (19): 145-150.

French, Robert M. 1990. "Subcognition and the Limits of the Turing Test." *Mind* XCIX (393): 53-65.

Gertler, Brie. 2001. "Introspecting Phenomenal States." *Philosophy and Phenomenological Research* (63): 305-328.

Godfrey-Smith, Peter. 2016. "Mind, Matter, and Metabolism." *The Journal of Philosophy* (10): 481-506.

Goff, Philip. 2017. *Consciousness and Fundamental Reality.* New York: Oxford.

Goldman, Alvin I. 2009. *Simulating Minds.* Oxford: Oxford University Press.

Gunkel, David J. 2018. *Robot Rights.* Cambridge, MA: MIT.

Hanson, Robin. 2016. *The Age of Em.* Oxford: Oxford University Press.

Harnad, Stevan. 2003. "Can a Machine Be Conscious? How?" *Journal of Consciousness Studies* (10 (4-5)): 67-75.

Hill, Christopher S. 2009. *Consciousness.* Cambridge: Cambridge University Press.

Humphrey, Nicholas. 2006. *Seeing Red.* Cambridge, MA: Harvard.

Irvine, Elizabeth. 2013. *Consciousness as a Scientific Concept.* Dordrecht: Springer.

Kriegel, Uriah. 2009. *Subjective Consciousness.* Oxford: Oxford University Press.

Kriegel, Uriah. 2019. "The Value of Consciousness." *Analysis* (79): 503-520.

Kurzweil, Ray. 2005. *The Singularity is Near.* New York: Penguin.

Levesque, Hector J. 2011. *The Winograd Schema Challenge.* http://commonsensereasoning.org/2011/papers/Levesque.pdf.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint.* July 26. arXiv:1907.11692v1 [cs.CL].

Moravec, Hans P. 1988. *Mind Children.* Cambridge, Massachusetts: Harvard University Press.

Nichols, Shaun, and Stephen P. Stich. 2003. *Mindreading.* Oxford: Oxford University Press.

Oizumi, Masafumi, Larrissa Albantakis, and Guilio Tononi. 2014. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." *PLOS Computational Biology* (10 (5): e1003588).

Papineau, David. 2002. *Thinking About Consciousness.* Oxford: Oxford University Press.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* (21): 1-67.

Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, et al. 2020. "Recipes for building an open-domain chatbot." *arXiv preprint.* April 30. arXiv:2004.13637v2 [cs.CL].

Rosenthal, David M. 2005. *Consciousness and Mind.* Oxford: Oxford University Press.

Rosset, Corby. 2020. *Turing-NLG: A 17-billion-parameter language model by Microsoft.* February 13. Accessed August 10, 2020. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/.

Schneider, Susan. 2018. "Artificial Intelligence, Consciousness, and Moral Status." In *The Routledge Handbook of Neuroethics*, translated by L. Syd M Johnson and Karen S. Rommelfanger, 373-375. New York: Routledge.

—. 2019. *Artificial You.* Princeton, NJ: Princeton.

—. 2016. "Susan Schneider on How to Prevent a Zombie Dictatorship." *The Splintered Mind.* June 27. http://schwitzsplinters.blogspot.com/2016/06/susanschneider-on-how-to-prevent_27.html.

Schneider, Susan, and Pete Mandik. 2018. *How Philosophy of Mind Can Shape the Future.* Vol. 6, in *Philosophy of Mind in the Twentieth and Twenty-First Centuries: The History of the Philosophy of Mind*, edited by Amy Kind, 303-319. New York: Routledge.

Schwitzgebel, Eric. 2011. *Perplexities of Consciousness.* Cambridge, MA: MIT.

Schwitzgebel, Eric. 2016. "Phenomenal Consciousness, Defined and Defended as Innocently as I Can Manage." *Journal of Consciousness Studies* (23 (11-12)): 224-235.

Schwitzgebel, Eric, and Mara Garza. 2015. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* (39): 98-119.

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* (3): 417-457.

Searle, John R. 1992. *The Rediscovery of the Mind.* Cambridge, Massachusetts: The MIT Press.

Shevlin, Henry. 2020. "General Intelligence: An Ecumenical Heuristic for Artificial Consciousness Research?" *The Journal of Artificial Intelligence and Consciousness* 7 (2). Accessed April 10, 2020.

Sloman, Aaron. 2007. "Why Some Machines May Need Qualia and How They Can Have Them: Including a Demanding New Turing Test for Robot Philosopers." *AAAI Fall Symposium.* https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aaai-consciousness.pdf.

Strawson, Galen. 2006. *Consciousness and its Place in Nature.* Edited by Anthony Freeman. Exeter: Imprint Academic.

Swinburne, Richard. 2007. "From Mental/Physical Identity to Substance Dualism." In *Persons: Human and Divine,* edited by Peter van Inwagen and Dean Zimmerman, 142-165. Oxford: Oxford University Press.

Tononi, Giulio, and Christof Koch. 2011. "Testing for Consciousness in Machines." *Scientific American: Mind*, September 1. doi:10.1038/scientificamericanmind0911-16.

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

Ziesche, Soenke, and Roman Yampolskiy. 2019. "Towards AI Welfare Science and Policies." *Big Data and Cognitive Computing* (3 (2)). doi:10.3390/bdcc3010002.