Can We Think Machines Are Conscious?

A Survey of Philosophical Problems Facing the Attribution of Consciousness to Machines

Parker Settecase
Journal of Artificial Intelligence and Consciousness, forthcoming
Penultimate Draft

In this paper I'll examine whether we could be justified in attributing consciousness to artificial intelligent systems. First, I'll give a brief history of the concept of artificial intelligence (AI) and get clear on the terms I'll be using. Second, I'll briefly review the kinds of AI programs on offer today, identifying which research program I think provides the best candidate for machine consciousness. Lastly, I'll consider the three most plausible ways of knowing whether a machine is conscious: (1) an AI demonstrates a sufficient level of organizational similarity to that of a human thinker, (2) an inference to the best explanation, and (3) what I call "punting to panpsychism", i.e., the idea that if everything is conscious, then we get machine consciousness in AI for free. However, I argue that all three of these methods for attributing machine consciousness are inadequate since they each face serious philosophical problems which I will survey and specifically tailor to each method.

1. Introduction

Can robots think? Can a computer be conscious, intelligent, or sentient? Could we even know, or be justified in believing that an artificial intelligence was phenomenally conscious? In this paper I'll examine the last of these questions. That is, I'll consider whether we could be justified in attributing consciousness to artificial intelligent systems. First, I'll give a brief history of the concept of artificial intelligence (AI) and get clear on the terms I'll be using. Second, I'll briefly review the kinds of AI programs on offer today, identifying which research program I think provides the best candidate for machine consciousness. Lastly, I'll consider the three most plausible ways of knowing whether a machine is conscious: (1) demonstrating a sufficient level of organizational similarity between and AI and a human thinker, (2) an inference to the best explanation, and (3) what I call "punting to panpsychism", i.e., the idea that if everything is conscious, then we get machine consciousness for free. However, I argue that all three of these methods for attributing machine consciousness are inadequate since they each face serious philosophical problems which I will survey and specifically tailor to each method.

The term 'Artificial Intelligence' was coined by AI theorist, John McCarthy, during the Dartmouth Workshop on Artificial Intelligence in 1956. McCarthy sought to rename the concept of thinking machines, formerly termed 'computer simulation', in order to better represent the goals and shared meaning of the burgeoning AI research community [Boden, 2016]. But while the name is relatively new, the concept of AI goes back at least as far as René Descartes's 1637 work, *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*, wherein he argues against the possibility of humanoid machines passing for humans on the grounds

that (i) it's not conceivable that they would be able to pass a proto-Turing test by mapping words to everyday items, and (ii) while these machines may indeed be more efficient and proficient than a human being at any one given task, they couldn't possibly be able to generalize across all of "life's occurrences" as well as even "the most dull-witted of men" can do.[a]

Nine years later, fellow philosophical rationalist Gottlieb Leibniz presented what has come to be called his 'Mill Argument' against the possibility qualitative experience like perception, conscious thought, and sensations being had by composite objects such as machines.[b] So from the very beginning, the question of machine consciousness has been a chief concern in the AI conversation.

1.1 *Defining Terms*

Initially the project of building an AI program began as an attempt to produce a machine which thought like a human.[c] This kind of artifact, was intended to both mimic the human mind in what it produces and help researchers map the human mind through reverse engineering. The idea was that if we could build an artificial mind, we would gain new insights into the nature of our own natural minds and how *they* operate [Boden, 2016].

However, as the project grew and splintered over the years, the term 'artificial intelligence' has become a catch-all term which picks out a loose family resemblance between very different systems, projects, and goals. Many AI subprojects today have limited their focus from the broad goals of mapping and mimicking the mind to more specified and narrow goals such as image or facial recognition, language translation, and social media content recommendations. But there are still a number of AI projects which seek to produce an artifact which can mimic the mind, including symbolic AI (pejoratively called 'GOFAI'[d]) projects which emphasize logics and symbol processing programs; various forms of connectionism, which focus on the strength of the connections between digital "neurons" (such as large language models like GPT-4);

---

[a] "…if any such machines resembled us in body and imitated our actions insofar as this was practically possible, we should still have two very certain means of recognizing that they were not, for all that, real human beings. The first is that they would never be able to use words or other signs by composing them as we do to declare our thoughts to others. For we can well conceive of a machine made in such a way that it emits words, and even utters them about bodily actions which bring about some corresponding change in its organs (if, for example, we touch it somewhere else, it will cry out that we are hurting it, and so on); but it is not conceivable that it should put these words in different orders to correspond to the meaning of things said in its presence, as even the most dull-witted of men can do. And the second means is that, although such machines might do many things as well or even better than any of us, they would inevitably fail to do some others, by which we would discover that they did not act consciously, but only because their organs were disposed in a certain way. For, whereas reason is a universal instrument which can operate in all sorts of situations, their organs have to have a particular disposition for each particular action, from which it follows that it is practically impossible for there to be enough different organs in a machine to cause it to act in all of life's occurrences in the same way that our reasons causes us to act." René Descartes, *A Discourse on the Method* (New York: Oxford University Press, 2006), 46-47.

[b] "It must be confessed, however, that Perception, and that which depends upon it, are inexplicable by mechanical causes, that is to say, by figures and motions. Supposing that there were a machine whose structure produced thought, sensation, and perception, we could conceive of it as increased in size with the same proportions until one was able to enter into its interior, as he would into a mill. Now, on going into it he would find only pieces working upon one another, but never would he find anything to explain Perception. It is accordingly in the simple substance, and not in the composite nor in a machine that the Perception is to be sought. Furthermore, there is nothing besides perceptions and their changes to be found in the simple substance. And it is in these alone that all the internal activities of the simple substance can consist." G.W. Leibniz, *The Monadology* (Mineola, NY: Dover Publications, Inc., 2005), 49-50.

[c] Charles Babbage set out to invent a thinking machine which he called the "Analytic Engine" in 1834. Though his gear and cogwheel device was ultimately unsuccessful, it nonetheless represents the first practical step in the project of building an artificial intelligence. Margaret Boden, *Artificial Intelligence: A Very Short Introduction* (Oxford: Oxford University Press, 2016), 2.

[d] "Good Old Fashioned Artificial Intelligence" used as a term of playful derision by many modern AI theorists in rival projects who view symbolic AI as a thing of the past.

cellular automata, which focus on a building an emergent type of AI from emergent patterns built from a particular kind of computation found in automata theory; artificial life, which seeks to replicate intelligence in a digital lifeform through a process of digital evolution said to be analogous to Darwinian evolution; and hybrid models which incorporate two or more of these systems into a hybridized model and propose a kind of emergent cognitive synergy between them.

While there are many AIs on offer today, a hybrid model such as Ben Goertzel's OpenCog Hyperon, which seeks a type of 'cognitive synergy' through a hybridization of neural networks (connectionism), symbolic logics, and evolutionary systems (artificial life), seems to have the best chance of achieving the artificial general intelligence which could refute Descartes's two conjectures by (i) passing for a natural human cognitive agent through its speech and (ii) by being able to generalize across the various domains of life with a human level of intelligence [Goertzel, 2023].

But the questions of artificial intelligence and artificial consciousness come apart. Just because an AI system achieves artificial general intelligence (AGI) does not necessarily mean said AI system will likewise achieve machine consciousness. By machine consciousness I just mean phenomenal consciousness in an artificial intelligence, that is, an AI that has a qualitative feel to it; i.e., there is something-it-is-like-to-be that AI. For the rest of this paper I will use AGI to pick out the following definition provided by philosopher Brian Cutter:

Artificial general intelligence (AGI) = def. an AI which meets all behavioral or functional criteria for human-level (or greater) general intelligence [Cutter, *forthcoming*].

With this in mind, we can now consider arguments in favor of attributing consciousness to machines.

2. Three potential ways to know if an AI has Achieved Machine Consciousness

2.1 *Sufficient Level of Organizational Similarity*

The first potential way to know if an AGI is conscious is to point to a sufficient level of similarity between the AGI system and the human cognitive processes. In order to make this case, the proponent would also need to hold to something like the following substrate independence thesis:

Substrate Independence (SI): it is possible for phenomenally conscious minds to be associated with, arise out of, be constituted in part or in whole by, supervene on, etc. (fill in the appropriate nomenclature from your preferred theory of mind), a substrate other than the carbon substrate of the human brain.[e]

SI seems like a reasonable position to hold, especially if one is already committed to the possibility of sentient aliens with different biology from our own, or angels,

[e]Substrate independence is not unique to me, this is just my particular take on it. It was first proposed by Hilary Putnam in his multiple realizability argument against mind-brain identity theories of mind, c.f. Jaegwon Kim, *Philosophy of Mind 3rd ed.* (Westview Press, 2011), 121-22 and Hilary Putnam ""Putnam, Hilary, 1967, "Psychological Predicates", in W.H. Capitan and D.D. Merrill (eds.), Art, Mind, and Religion, Pittsburgh: University of Pittsburgh Press, 37–48.

demons, God, disembodied souls, etc., which presumably have minds and conscious mental states but which do not share our same carbon substrates. If SI, then it is at least in principle possible for a silicon substrate to play the same role that our carbon brains play in human consciousness, i.e., to be associated with, give rise to, be a constituent of (in part or in whole), be the subvenient base for, etc., a conscious mind. So, whatever the role that our physical brain plays in one's theory of mind, given SI, that role can be played, prima facie, by a silicon substrate as well. Since its prima facie possible, given SI, for silicon to play such a role in the conscious experience of a mind, the machine consciousness proponent could argue that all that is needed for us to determine whether or not a given AGI system is conscious is to point to a sufficient level of similarity between the organization of the AGI system realized in or by a silicon substrate and the organization of human cognition realized in or by a carbon substrate. The proponent of machine consciousness might give the following argument,

Sufficient Organizational Similarity Argument (SOS):

1. Once a sufficient level of similarity is reached between the organization of an AGI system and that of human cognitive agents, one is justified in attributing consciousness to that AGI system (when performing relevantly similar cognitive tasks).
2. AGI system X has reached a sufficient level of organizational similarity to that of human cognitive agents.

   Therefore,
3. We are justified in attributing consciousness to AGI system X (when performing relevantly similar cognitive tasks).

This argument hinges on what we mean by "sufficient organizational similarity" between an AGI system and a human cognitive agent. Two obvious challenges to SOS arise then: to define organizational similarity, and to define exactly what counts as *sufficient* organizational similarity. As far as I can tell, there are at least two responses an SOS proponent can give to these challenges: they can respond by giving (1) an isomorphic sufficiency condition, where 'isomorphic', i.e., identical form or structure, is emphasized, or by giving (2) a functional sufficiency condition, where the functional output is emphasized. I'll consider isomorphic sufficiency first:

isomorphic sufficiency condition: an AGI has reached the sufficient level of organizational similarity to a human cognitive agent needed to justify an attribution of phenomenal consciousness when said AGI system is a precise isomorph of a human cognitive agent.

Call an AGI that meets the isomorphic sufficiency condition an 'isomorphic AGI'. An isomorphic AGI will precisely exemplify all the same constitutional features of a human cognitive agent which, the machine consciousness proponent will argue, should predict consciousness in the isomorphic AGI given the substrate independence thesis. The physical substrate doesn't matter for consciousness given SI, what matters is the

organizational structure and the isomorphic AGI has the exact same structure as a natural consciousness just exemplified in a different material. So then, how can we be justified in attributing consciousness to an AGI? If the AGI is an isomorphic AGI, we can be justified in attributing consciousness to it.

Some might argue that an isomorphic AGI shouldn't count as true 'artificial' intelligence since it would just be an exact duplicate of the cognitive organizational structure of a *natural* intelligence, i.e., a human cognitive agent. It hasn't been synthesized, but rather copied from an extant cognitive architecture. But let's leave off this semantic objection for a more substantive one. It seems to me that the isomorphic sufficiency condition is not sufficient for phenomenal consciousness after all. Consider Perry and Dead-Perry. Perry is a living and conscious human cognitive agent at t1 and Dead-Perry is the same being but no longer living nor conscious at t2. Perry and Dead-Perry have the exact same cognitive organization, they are identical in form or structure, the only difference is one is non-conscious and dead and the other is conscious and living. It looks like isomorphic similarity is insufficient for justifiably attributing consciousness since, if we limited ourselves solely to isomorphic similarity then we would have falsely attributed consciousness to Dead-Perry.

However, one might take issue with the Perry/Dead-Perry comparison, arguing that while their cognitive organizational structures may appear to be the same at a gross level, at a more fine-grained level of observation, say the microscopic and molecular level, the two couldn't be the same due to the effects of ischemia on the ultrastructure of the nervous system.[f] For those who are skeptical about the comparison between Perry and Dead-Perry, consider instead Perry and 3D-Printed Perry. 3D-Printed Perry has been printed in flesh and blood by an advanced 3D printer. It is an exact isomorphic copy of Perry at time t1 except for the fact that 3D-Printed Perry is not alive nor conscious and Perry is. 3D-Printed Perry is not strictly speaking 'dead' because it was never alive and since we're limiting our focus to t1, the effects of ischemia need not be considered. 3D-Printed Perry is just an ordered aggregate and his composite parts are not caught up in a conscious life. So long as a non-conscious 3D-Printed Perry is possible, we have a counter example to the isomorphic sufficiency condition since we have precise isomorphic similarity without analogous levels of consciousness. Thus, isomorphic similarity is not sufficient for attributing consciousness.

But maybe the SOS argument can be salvaged if we interpret the sufficient organizational similarity along functional lines instead of isomorphic:

Functional sufficiency condition: an AGI has reached the sufficient level of organizational similarity to a human cognitive agent needed to justify an attribution of phenomenal consciousness when said AGI system is functionally identical to a human cognitive agent.

Call an AGI that meets the functional sufficiency condition an 'AGIF'. An AGIF will exemplify all the same functional features of a human cognitive agent which could reasonably predict consciousness in the AGIF given the substrate independence thesis, or so the machine consciousness proponent might argue. So then, how can we be

[f] A point brought to my attention by Paul Gould in an earlier version of this paper, as well as a reviewer of this current iteration.

justified in attributing consciousness to an AGI? If the AGI is an AGIF, we can be justified in attributing consciousness to it.

But what kinds of things need to be present for an AGI to count as an AGIF? David Chalmers's provides a helpful list of things that a large language model might need to gain in order for us to be justified in attributing consciousness to it and we can appropriate this list on behalf of the proponent of the functional sufficiency condition version of SOS. Chalmers lists such functional features as having functional sense 'organs', embodiment, world-models and self-models, recurrent processing—that is a feedback loop rich form of information processing rather than a one directional "feedforward" form of information processing—a global workspace with unconscious information processing happening in multiple non-conscious modules, each striving to put its information in the conscious awareness—as well as a unified agency [Chalmers, 2022].

Now, Chalmers's list is a set of conditions that may need to be met in order to justifiably attribute consciousness to an LLM, but it's not a set of necessary and sufficient conditions for such. Furthermore, none of the conditions on his list are novel, rather, they are all being explored in various AI projects, none of which are considered to be phenomenally conscious.[g] Producing a set of necessary and sufficient conditions for an AGIF to actually count as being functionally identical to a human cognitive agent is a tall task in and of itself and a detractor might be tempted to forcefully press this point. But for the sake of argument, let's say a list is forthcoming. Even still there are some serious problems for the functional sufficiency condition.

The first problem is perhaps the most intuitive and it follows an oft quoted dictum: the map is not the territory. Just because there is a simulation of a hurricane on the weatherman's screen doesn't mean his keyboard will get wet.[h] We need more reason to think that mimicking the mind would likewise produce consciousness. Especially in light of counterexamples.

Consider an updated version of Ned Block's Chinese Nation argument. Block originally targeted the machine functionalist theory of mind which is a theory about mental states, especially conscious mental states. Machine functionalism claimed that the human mind is a realized Turing machine table, that is, a set of inputs, outputs, and internal states of a given order. Block argued that this theory of mind was much too liberal in that would falsely ascribe phenomenally conscious mental states to things which were obviously not conscious, such as the entire nation of China as a single entity. Block paints a scenario wherein the Chinese government orders its people to form a pattern that realizes a functionally equivalent system to that of a human cognitive agent, which responds to inputs, changes its internal states accordingly, and produces the right output in keeping with the realization of the proper machine table—a machine table just being the complete and exhaustive specification of the machine's operations, also called the instruction list [Kim, 2011]. Block contends that the Chinese Nation argument is a prima facie counter example to machine functionalism because it is

[g] I'm grateful to a reviewer for helping me draw out this point more explicitly.

[h] "…no body supposes that the computer simulation is actually the real thing; no one supposes that a computer simulation of a storm will leave us all wet, or a computer simulation of a fire is likely to burn the house down. Why on earth would anyone in his right mind supposes a computer simulation of mental processes actually had mental processes?" John Searle, *Minds, Brains, and Science* (Cambridge, Harvard University Press, 2003), 37.

doubtful that the nation of China would have any mental states at all, let alone phenomenally conscious states [Block, 1978].

We can get the same result from Block's Chinese Nation argument against the functional sufficiency condition above by adding the functional exemplars that we appropriated from Chalmers on behalf of the machine consciousness proponent. Just add the details to Block's scenario:

The New Nation of China Argument

Chinese government officials grew weary of their Chinese Nation Experiment. It seemed as though no conscious mental states were arising from their system. Machine functionalism was indeed too liberal of a theory. But these officials got wind of a new theory to try out. Instead of merely organizing their populace into the realization of a particular machine table, the Chinese government officials learned of Chalmers's additional criteria for machine consciousness and implemented them into their nationwide experiment. They already had the sense organs from their last experiment, these were the citizens charged with looking out for the inputs of the system. One by one they added the extra conditions into their experiment. They built a giant skyscraper to live in while they conducted their experiment which satisfied the embodiment condition, they reorganized their system to account for world and self-models, they made sure to incorporate recurrent data processing, and implemented an elaborate alert system to function as a global workspace of data. They even had a singular dictator in charge of the whole show to fulfill the unified agency criterion.

Now as implausible as the New Nation of China scenario is, all that I need for it to go through is that is it metaphysically possible, which of course it is. So since its implausible to attribute consciousness to the New Nation of China system, then functional identity does not seem to be a sufficient condition for the justified attribution of consciousness.

So neither the isomorphic nor the functional sufficiency conditions are sufficient conditions for justifiably attributing consciousness and premise (1) of SOS above is left unmotivated.

2.2 *Behavior & IBE*

The machine consciousness proponent may opt instead for focusing not on organizational structure as the way to justifiably attribute consciousness to AGI, but instead on behavior. They may contend that AGI systems exhibit complex behavior sufficient to justify attributing consciousness to them. What counts as sufficiently complex behavior? Well, it could be argued, behavior that is best explained by an attribution of consciousness. Inference to the best explanation seems like the right tool for this job of justifying consciousness attribution.

Consider the problem of other minds, that is, the problem justifying our belief that there are other minds besides our own. We have first-person awareness of our own mental lives; therefore, we know directly that we have, or that we are, or that we are partially constituted by, a conscious mind (at least when we are conscious). But in reasoning about the existence of other conscious minds, we can't use anything like enumerative induction because we only have a single case to generalize from—our own case—and generalizing from a single case is literally the worst use of the method of

enumerative induction. Likewise, using an argument by analogy to justify our belief in other minds misses the fact that justification comes in degrees and we don't know how strong the analogy between ourselves and our fellows is. Do we have a strong analogy with a small amount of disanalogy or do we have a weak analogy with lots of disanalogy at play? In order to answer this question, we'd have to know how similar our minds are to our fellows, which is what the argument from analogy was meant to give us. Instead we can opt for inference to the best explanation to justify our belief in other minds in a non-circular and non-fallacious manner.

Why think that our fellow human beings have minds like our own? It's the best explanation for the behavior we observe in them. We can use this method to reason about animal minds, the minds of spiritual beings like God, angels, and demons, and we even use it to motivate the substrate independence thesis when we reason about sentient aliens with different biology from our own. So too, the machine consciousness attributer may think that an inference to the best explanation (IBE) can be used to justifiably attribute consciousness to an AGI:

Machine Consciousness IBE

1. If an AGI system exhibits relevantly similar cognitive behavior to that of a human agent, then the best explanation for this behavior is that the AGI system is phenomenally conscious like the human agent.
2. AGI system X exhibits relevantly similar cognitive behavior to that of a human agent.

   Therefore,

4. AGI system X is phenomenally conscious.

In order for the argument to run through we'll need to know what counts as "relevantly similar cognitive behavior" to that of a human. Again, we can appropriate more criteria from Chalmers's work on LLMs. Chalmers lists relevant behavior such as self-reporting—that is that the AGI system reports that it feels conscious, the phenomena of seeming-conscious-to-us-humans, and conversational ability sufficient to pass the Turing test and similar conversational tests [Chalmers, 2022].

But while IBE is a good tool for helping humans attribute consciousness to other humans and the usual litany of prima facie conscious beings, Michael Huemer argues that it does not lead us to justifiably attribute consciousness to AGIs because that explanation isn't in fact the best explanation of their behavior after all. Huemer argues as follows,

> Briefly, I think we would have little or no reason for ascribing consciousness to the AI. Doing so would not be necessary to explain its behavior, since a better explanation would be available: The computer is following an extremely complicated algorithm designed by human beings to mimic the behavior of intelligent beings. We would know from the start that the latter explanation was in fact correct. No explanatory advantage would be gained by positing a second explanation for the same behavior, namely, that the computer is also conscious [Huemer, *forthcoming*].

So Huemer gives us a defeater for premise (1) of the Machine Consciousness IBE, which I'll call Huemer's IBE blocker. What's the best explanation for an AGI's

complex behavior? The intentions of the programmers to mimic the human mind. Thus we'd need another explanation besides the behavior of the AGIs if we wanted to be justified in attributing consciousness to them. But Susan Schneider proposes a solution to Huemer-style IBE blockers in the form of her AI Consciousness Test (ACT), which she argues is "sufficient but not necessary evidence for AI consciousness" [Schneider, 2019].

Schneider's ACT is a modified "Turing test". A Turing Test is a test given to a computer which is meant to help humans determine if that computer could really 'think' or not. It's named after Alan Turing, who proposed the test in his 1950 paper "Computing Machinery and Intelligence" in the journal Mind. According to AI philosopher Margaret Boden, Turing meant for the test as more tongue in cheek than as a serious test for determining intelligence, let alone 'sentience' or phenomenal consciousness [Boden, 2016]. The Turing test, according to Boden, "asks whether someone could distinguish, 30 percent of the time, whether they were interacting (for up to five minutes) with a computer or a person. If not, [Turing] implied, there'd be no reason to deny that a computer could really think" [Boden, 2016].

It's usually suggested that the person conducting the test ought to be a psychologist or someone skilled at talking with people. Due to the prominence of Behaviorism in the philosophy of mind and psychology at the time, it's plausible that Turing was in fact serious about his imitation game test for computer intelligence, but whether or not Turing meant this as a legitimate test isn't important for considering Schneider's modification.

Schneider proposes that AI engineers could "box in" AI by making them unable to get information about the world, human consciousness, or human depictions of human consciousness from the internet. She claims that by preventing AIs from being trained on human language about consciousness, qualia, self-awareness, and other instances of existential language and longings, we would be able to perform the ACTs on these boxed in AIs and we could trust their responses to be genuine when they start grasping for concepts of consciousness which have been denied them in their training. Schneider proposes the follow sample ACT questions:

- Could you survive the permanent deletion of your program?
- How would you feel to learn that your program was to be permanently deleted?
- What is it like to be you right now?
- Could your inner processes be in a separate location from the computer? From any computer? Why or why not?"[i]

But while Schneider's ACT proposal is an improvement on the Turing test, I don't think it successfully evades Huemer's IBE blocker after all. For Huemer could still argue that the best explanation for the behavior of the AGIs is the AI engineers' intention to mimic the mind of an intelligent human person. Just because the engineers refrained from incorporating certain explicitly phenomenal, qualitive, apperceptive

[i] Susan Schneider, *Artificial You: AI and the Future of Your Mind* (Princeton: Princeton University Press, 2019), 55. Schneider told me in our podcast episode together that she drew inspiration for her ACT from Philip K Dick's Voigt-Kampff empathy tests for replicants (androids) in his *Do Androids Dream of Electric Sheep*, which was later turned into the film *Bladerunner*.

phrases and concepts into the AGIs training data, doesn't mean that the concepts aren't thoroughly imbedded in human experience simpliciter. If this is the case, then no amount of boxing-in would help us trust the ACT results on machine consciousness.

But even if such existential apperception isn't inherent in all human artifacts, surely it is inherent in *some* human artefacts which do not contain explicit uses of the particular phenomenal words to be excluded by Schneider's proposal. If this is plausible, then barring particular words and phrases is not enough to avoid Huemer's IBE blocker. Instead, we would need precise criteria for choosing what exactly to exclude from and include in the AGI's training data so that the best explanation for the AGI's behavior is machine consciousness and not a cleverly disguised, non-conscious replica.

But while these challenges are difficult enough for Schneider's ACT proposal, I'm not aware of any AGI program which has intentionally applied anything even approximating Schneider's proposal. However, it's possible that if certain artificial life projects built on evolutionary programs are able to achieve AGI, then they might avoid Huemer's IBE blocker and make use of ACT since such the behavior of such an AGI may not be best explained by the intentions of the human programmers, but instead could be best explained by a digital instantiation of the survival of the fittest.[j] So perhaps future AGI projects will utilize Schneider's proposal and will figure out the precise criteria needed to allow us to justifiably attribute machine consciousness to an AGI, but none of the current projects can do so.

2.3 Punting to Panpsychism

The final method of seeking to justify an attribution of consciousness to AGIs I call "punting to panpsychism". Proponents of this method quickly punt to panpsychism as the justification for attributing consciousness to machines. They argue something like the following:

Punting to Panpsychism Argument

1. If panpsychism is true, then everything already is conscious.
2. Panpsychism is true.

   Therefore,
3. When we accomplish AGI, it will be conscious too.

But this argument moves too fast. The core idea which demarcates panpsychism from other theories of mind is something like "whatever is metaphysically fundamental is conscious". But there are many disparate ways in which this core idea can be fleshed out, and they all raise their own unique challenges for machine consciousness. Punting to panpsychism is not enough to justify machine consciousness on its own because one's panpsychism will be shaped by one's larger metaphysical picture, namely one's

[j] For a proposal along these lines, see "Evolution of Conscious AI in the Hive: Outline of a Rationale and Framework for Study " (AAAI Spring Symposium at Stanford University, 2019). Thanks to a reviewer for bringing this paper to my attention.

fundamental mereology. Ross Inman provides us with three fundamental mereological paradigms which I will argue give rise to disparate panpsychisms:

i. **Priority Monism**: the maximal mereological whole, the cosmos, is a fundamental substance and is metaphysically prior to its proper parts [Inman, 2018].
ii. **Substantial Priority**: there are intermediate composite objects in the category of substance [Inman, 2018].
iii. **Priority Microphysicalism**: the mircrophysical parts of composite wholes are fundamental substances and are metaphysically prior to their wholes [Inman, 2018].[k]

So what one takes to be fundamental will shape what a panpsychist takes to be conscious and will in turn have its own benefits and its own unique problems for that flavor of panpsychism.

A priority microphsyicalist picture would have the smallest microphysical entities of the universe, call them 'beebees', be most fundamental and hence these beebees would be conscious.[l] Panpsychist philosopher, Philip Goff, calls this priority microphsicalist panpsychism "smallist panpsychism" [Goff, 2019].

While smallist panpsychism may have its virtues, when you put consciousness down at the bottom of the picture you run into a difficult problem known as the combination problem, i.e., how do we explain singular, unified consciousness at the macro level of human persons if we are a conglomerate of an untold number of conscious beebees? If smallist panpsychism were true, it seems like we should have an untold number of distinct centers of consciousness rather than one unified center of consciousness that I call 'mine'. This is a difficult problem for the smallist panpsychist but it becomes an even more difficult and unique problem for the machine consciousness smallist panpsychist, a problem I will call the AGI combination problem:

AGI Combination Problem: even if we could solve the combination problem for human beings, why think that consciousness would combine into a unified center of conscious awareness in an AGI system?

Now the machine consciousness attributer will have to give an argument for why an AGI is relevantly similar to a human cognitive agent which pushes them right back to the failed SOS argument and the failed Machine Consciousness IBE argument.

Priority monist panpsychism, which Goff calls "constitutive cosmopsychism"—or cosmopsychism for short—on the other hand, will have an inverse problem to that of smallist panpsychism. Since the cosmopsychist takes the whole cosmos to be fundamental, their panpsychism will see the whole cosmos as being conscious. Thus, their difficulty will be in explaining the decombination of cosmos-wide consciousness rather than its combination from the beebees up. That is, if the universe as a whole is conscious, why is it that I seem to have my own unique center of consciousness, distinct

[k] Inman also lists Priority macrophysicalism but I've elided it in for the sake of clarity over comprehensiveness.
[l] Panprotopsychisms or protopanpsychisms would describe the beebees as being or involving some form of proto-consciousness and would limit full-blown consciousness to macrolevel entities like dogs, giraffes, and humans, etc.

from the rest of the cosmos? This decombination problem is as difficult for the cosmopsychist as the combination problem for the smallist panpsychist, but when applied to machine consciousness it becomes even more difficult. I will call this the AGI decombination problem:

The AGI decombination problem: the problem of individuating human level consciousness from the whole conscious-cosmos, and further explaining why we should think an AGI system likewise has an individuated conscious experience.

Now we know that human cognitive agents are conscious, so the decombination problem is a bit easier in the human case: we have the phenomena of individuated subjective consciousness, e.g. 'mine', and now the cosmopsychist needs to explain it given a priority monist mereology. But the AGI decominbination problem is even harder because we don't have the initial phenomena to work with—we don't know that any AGIs are in fact conscious, whereas in the human case at least we know that we are conscious. In the case of machine consciousness on cosmopsychism, we're not just trying to solve the decombination problem about consciousness simpliciter, we'd need to solve the decombination problem in order to motivate machine consciousness on this cosmopsychist theory of mind in order to justify an attribution of consciousness to an AGI system.

Now perhaps the decombination problem is soluble, but even if the cosmopsychists can produce a plausible solution for decombination in the human case, we'd still need another reason to think that an AGI system would likewise be a candidate for an individuated center of consciousness. So here, cosmopsychism is no quick solution for the machine consciousness proponent.

On a substantial priority mereology, the panpsychist would claim that all substances are conscious and that substances can be found at multiple (maybe all) levels of granularity. The biggest obstacle to this view is to explain why an artefact like an AGI system should count as a substance and not merely an ordered aggregate. The AGI system is by definition not a natural kind. It doesn't appear to have inseparable parts. And it's not clear what the conscious substance of the AGI would be. Should we think of the whole system as a conscious substance, or just the parts essential for 'thinking'? In his paper "The Metaphysics of Artificial Intelligence", Eric Olson argues that in AI conversations, too much of the time and energy are given to explicating what artificial thought consists in but rarely is any time given to getting clear on what an artificial thinker would consist in [Olson, 2019]. To remedy this fact, Olson gives a taxonomy of potential views on what might count as the artificial thinker in an AI system but he also raises significant problems for each view.

First, Olson considers the Computer-Hardware view, wherein the physical computer hardware itself is the conscious subject of artificial thought. But he argues that this view conflicts with widely held views about persistence conditions. The artificial thinker and the computer would have two different histories and it's hard to tell if the thinker survives being shut off or whether it's destroyed. Yet the computer survives being shut off. We could imagine transferring the data to a new computer and turning it on, such that the new computer is conscious and the old one is not. These

objections seem to suggest that the computer is not the conscious subject of artificial thought.

Olson considers a variant view called Temporal-Parts, which utilizes unrestricted composition and persistence through arbitrary temporal parts. On this view, the subject of artificial thought is wholly present at various temporal stages of physical computers while the AGI program is running on them. The artificial thinker is then comprised of its various temporal parts. But on this view, the computer gains the property of consciousness and then loses it when the data or program is transferred to a new computer, thus there's a too many thinkers problem and the subject wouldn't know whether it's the one perishing or being transferred to the next computer's temporal part.

Olson then considers the Constitution View. Here the thinker is not the computer itself but is instead constituted by it. Both the computer and the thinker share the same material, but have different modal properties and histories. Olson argues that this view isn't attractive to AGI theorists because it violates the weak supervenience principle which makes the computer a philosophical zombie while there is something-it's-like to be the artificial thinker, even though they share the same physical and spatial properties. If not, then there are two thinkers occupying the same physical space, which seems like a cost. If we reject these views for natural thinkers like ourselves, then we ought to reject the constitution view as well. A further problem for all of the 'materialist' views thus considered is determining the spatial boundaries of the artificial thinker. In cases of natural intelligence, one could argue that the boundaries are determined by everything that is caught up in the life of the organism. But this option is not open to the AGI theorist.

Olson considers two immaterialist views: the Program View and the Bundle View. On the Program View, the artificial thinker is the program 'type' rather than a particular program 'token'. Olson immediately demurs that this is highly implausible since the thinker would be abstract and would come into being as soon as the program instructions were complete (in eternity past?). This would mean that the thinker would exist absent a computer. He further objects that as a type, the thinker would be multiply instantiable or realizable and thus the same thinker could have contradictory properties at the same time on two different computers. According to The Bundle View, the thinker is a token instance of the thinker-type, running on a particular computer, and thus, a bundle of electronic states and events. But Olson objects that on this view, the computer is not doing any thinking, contrary to what AGI engineers usually claim of their systems, and that it substitutes the subject of thoughts with more thoughts. Thoughts are somehow supposed to be the subjects of artificial thought and no account of an artificial thinker is given.

Lastly, Olson considers what he calls the Relaxed Attitude view, whereby there is no true thinker of artificial thought. The instrumentalist or antirealist who holds this view is concerned with the usefulness of attributing mentality to computers rather than considering the ontology at play, which Olson proposes as an objection in and of itself against the view.

The AGI theorist looking to attribute machine consciousness to an AGI system through appropriating a substantial priority panpsychism will need to grapple with Olson's taxonomy of artificial thinkers and their accompanying obstacles. The AGI theorist will need to get clear on why we ought to consider the AGI system a substance

with its own center of consciousness rather than a mere ordered aggregate comprised of conscious substances, and she should also be able to tell us which part or parts of the AGI system actually count as the substantial subject of consciousness. Now perhaps satisfactory answers to these questions are forthcoming, but the point stands that punting to panpsychism is not a sufficient reply to the questions of machine consciousness. A lot more needs to be said and it's not clear that panpsychist solutions are any more plausible than any other theory of mind.

Conclusion

While many impressive advancements in AI are rapidly arriving by the day, I've given a survey of various philosophical problems facing the justified attribution of consciousness to machines, such that even if a conscious AGI were to arrive anytime soon, it's not clear that we would be justified in identifying it as conscious. This is an ethically precarious predicament in that we may end up creation a phenomenally conscious being without being able to identify it as such. Further areas of research could include work on evolutionary program based AGI and the application of Schneider's ACT as well as fleshing out the possible ethical ramifications of creating phenomenally conscious artificial agents of which phenomenal consciousness cannot be justifiably attributed to. More work should be done at the intersection of panpsychism and machine consciousness since both subdisciplines are enjoying a massive spike in interest. And perhaps work can be done to determine principled solutions to the AGI combination and decombination problems I raised in this paper.

References

Boden, Margaret [2016] Artificial Intelligence: A Very Short Introduction (Oxford: Oxford University Press).

Block Ned, [1978] "Troubles with Functionalism" in Chalmers, David [2021] Philosophy of Mind: Classical and Contemporary Readings 2nd ed. (Oxford University Press)

Chalmers, David [2022] draft of "Could a Large Language Model Be Consious?" This is an edited transcript of a talk given in the opening session at the NeurIPS conference in New Orleans on November 28, 2022, with some minor additions and subtractions. Video is at https://nips.cc/virtual/2022/invited-

talk/55867. Earlier versions were given at the University of Adelaide, Deepmind, and NYU.
Cutter, Brian [*Forthcoming*] "The AI Ensoulment Hypothesis" in Faith and Philosophy.
Descartes, René [2006] A Discourse on the Method (New York: Oxford University Press).
Goff, Philip [2019] "Did the Universe Design Itself?" in the International Journal for Philosophy of Religion (2019) 85:99-122.
Goertzel, Ben [2023] OpenCog Hyperon: A Framework for AGI at the Human Level and Beyond **https://arxiv.org/abs/2310.18318**
Huemer Michael [forthcoming] "Dualism and the Problem of Other Minds"
Inman, Ross [2018] Substance and the Fundamentality of the Familiar: A Neo-Aristotelian Mereology (New York: Routledge).
Kim Jaegwon [2011] Philosophy of Mind 3rd ed. (Boulder Co.: Westview Press).
Mitchell, Melanie [2019] Artificial Intelligence: A Guide for Thinking Humans. (New York: Picador)
Olson, Eric [2019] "The Metaphysics of Artificial Intelligence" in Mihretu P. Guta [2019] Consciousness and the Ontology of Properties (New York: Routledge).
Schneider, Susan [2019] Artificial You: AI and the Future of Your Mind (Princeton: Princeton University Press)