Where's Wall-E? Exploring the Challenges of Developing Artificial Machine Consciousness

By
Sun Schuette

Submitted to Central European University - Private University
Department of Undergraduate Studies

In partial fulfilment of the requirements for the degree of Bachelor of Arts in Philosophy, Politics, and Economics

Supervisor: Professor Tim Crane

Vienna, Austria
2025

Copyright Notice

For bibliographic and reference purposes this thesis should be referred to as: Schuette, S(un). 2025. Where's Wall-E? Exploring the Challenges of Developing Artificial Machine Consciousness. BA thesis, Undergraduate Studies, Central European University, Vienna.

AUTHOR'S DECLARATION

I, the undersigned, Sun Schuette, candidate for the BA degree in Philosophy, Politics, and Economics declare herewith that the present thesis titled "Where's Wall-E? Exploring the Challenges of Developing Artificial Machine Consciousness" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 26th May 2025

Sun Schuette

Abstract

This thesis explores the philosophical challenges in developing verifiably existing artificial machine consciousness. It lays out the necessity of navigating several problems and traps. First, the question of biological necessity is discussed from the positions of biological naturalism, and conversely the position of substrate independence. Second, the substrate abuse trap is made, raising the issue of correctly utilizing substrates. Third, the recognizability trap covers the issue identifying exotic AI might raise. Lastly, the reasons for and against considering the importance of the hard problem of consciousness on artificial machine consciousness are laid out. Drawing from biological naturalist perspectives, this paper expresses skepticism toward optimistic claims about the imminent creation of artificial machine consciousness and highlights the ethical implications of conscious-seeming machines, such as potential exploitation of our human biases and moral considerations. The discussion underscores the limitations of current definitions of consciousness and the need for rigorous, cautious research to responsibly advance the understanding and development of artificial machine consciousness.

Acknowledgements

The completion of this thesis would not be possible without the help and support of so many. I would like to first thank my parents for their total support of me in my studies. I would like to thank all those close to me who endlessly entertained my rants and ravings on the contents of thesis as I developed it. In particular, thank you to Jessica and Klara for their help up to the eleventh hour. Last but not least, I would like to thank my supervisor Professor Tim Crane for his expert advice and feedback.

I also would like to acknowledge the great history of black philosophers like Angela Davis, W.E.B. Du Bois, Frantz Fanon, Anton Wilhelm Amo, and Zera Yacob, to name a but few. I dedicate this thesis to the hope of someday contributing as much as they and many others have to our great tradition.

TABLE OF CONTENTS

Chapter 1: The Possibility of Verifiably Existing Artificial Machine Consciousness

Present-day engagement with the possibility of artificial machine consciousness (AMC) is home to a myriad of diverging perspectives, all well-equipped with complex vocabulary to differentiate them from their interlocutors. There is, however, a general negligence of the human component of AMC development in existing scholarship. The implicit human-centered (anthropocentric) expectations we insert when iteratively developing AMC are taken for granted. Likewise undervalued is the possibility that an AMC would defy our anthropomorphic assumptions, manifesting in ways not like us. On these two overlooked aspects of AMC analysis, the anthropocentric urge and the possibility of defying our anthropomorphic expectations, I draw out the concerns that arise in this paper. To the same end, this paper will serve to outline the philosophical traps and problems that must be successfully avoided to achieve the aspiration of verifiably existing AMC. Starting with general concerns of biological necessity, and then moving to the substrate abuse trap, the intelligibility trap, and finally the hard problem, this paper will seek to map out four different philosophical issues and potential considerations that might be hazardous to one in the pursuit of creating AMC. The word trap, as they will be called here, is used to denote that, rather than problems meant to be fixed, they are issues meant to be circumvented or avoided These considerations have further importance because when the door to consciousness is opened, ethical considerations come rushing in. Without understanding the legitimacy of the possibility of machines as ethical agents or patients, one is left to wildly speculate, risking oversight of prescient AI ethics issues, such as its energy use accelerating the effects of climate collapse, in favor of pursuing impossible contingencies. On the other hand, without proper evaluative criteria to offer once one has created a verifiable existing AMC, the risk of minimizing the ethical significance of possible AMC raises the complementary moral issue.

Chapter 2: A Brief History of Conscious Machines

The drive to make the world in our image, and to imagine ourselves as made in the image of something greater, that is, to give our world anthropomorphic meaning, reaches deep into the murky prehistory of humankind (Mithen 1996). At least as far back as the ancient Greek myth of the bronze automaton Talos (Apollodorus 1921, 1.9.26), through to the plethora of contemporary depictions such as WALL-E (2008), the Bicentennial Man (1999), or the Replicants of Blade Runner (1982), visions of crafting sentient humanlike machines have captured the human imagination. We are asked today by Silicon Valley entrepreneurs to believe that, as a result of our tremendous technological acceleration post-industrial revolution, the existence of our machinic mirror is closer than ever (Bidgood & Nehamas 2025). Others still, such as one former Google engineer, believe that machine consciousness has already been achieved in today's most advanced large language models (Maruf 2022). On a far different note than the technologists, philosophers on the possibility of artificial consciousness remain far less enthusiastic and far more divided. In the Western tradition, early philosophical discourse on machines capable of human proficiency in a wide set of tasks—referred to today as artificial general intelligence (AGI)—can be found as early as 1637 when Rene Descartes cast serious doubt that a machine could outperform even "the dullest of men" in "all the contingencies of life" (Descartes 2008, 22). Similarly anticipating contemporary discussion, in his seminal work The Monadology (1714), Leibniz rejected the idea that a mechanistic perspective alone could sufficiently explain mental states in what is known as the mill argument: "If we imagine a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged...so that we could enter into it, as one enters a mill. Assuming that...we will find only parts that push one another, and we will never find anything to explain a perception" (1898, 228).

For fear of being left out of the philosophical jargon production complex myself, I propose a few new or simply revised terms to be used in this paper. Perhaps the hottest term in

today's sphere of machinic futures is AGI or Artificial General Intelligence. Coined by Mark Gubrud in 1997 and popularized by Ben Goertzel (Goertzel 2011), AGI is aspirationally defined as AI capable of replicating and surpassing human-level cognitive ability in all domains of intelligence (Goertzel 2014). In place of AGI or the popular catch-all phrase Artificial Intelligence (AI)--a term first coined by John McCarthy at the Dartmouth Summer Research Project on Artificial Intelligence in 1956 (McCarthy 1995)--I use Artificial Machine Consciousness to focus explicitly on the question of *non-biological* consciousness, a quality that is often thought of as the final frontier in AI development, or by some as a necessary step towards AGI (Levy 2013). Existing literature often relies either on the term artificial consciousness or the term machine consciousness (Buttazzo 2001; Gamez 2008). Despite possessing different implications, these terms are seemingly used to equivocal effects. Among other motivations, to preemptively account for any concerns of using just one or another, I will include both. Artificiality here will refer first to non-biological material and second being of human creation. The definition of machine will largely reflect its colloquial use as a system of structural components that result in a mechanistic process, and for this paper, often computation. This operator merits inclusion for these reasons, but also for the fact that current speculation on artificial consciousness is centered around machinic and computational accounts. Defining consciousness is a tricky matter as the problem of searching for consensus on the term consciousness is precisely in the fact that there isn't one. Scholars are in stark disagreement about what is necessary for consciousness to arise, how it arises, its qualities, etc. Many accept, with the notable exception of access consciousness proponents such as Ned Block (1995, 229), that essential to consciousness is a phenomenological or qualitative experience. In other words, if it is conscious, then it is like *something* to be that entity (Nagel 1974). This minimal definition, in hopes of making as few claims about the nature of consciousness as feasible, is where I will start. As they are often conflated in popular discourse, a clarification should be made on the distinction between consciousness and intelligence. While the definition of intelligence lacks consensus, it

is broadly concerned with what a system does or its "functional capabilities" (Seth 2025, 2), and though for human beings intelligence and consciousness are bound up with one another, one should not assume they are interchangeable or necessarily intertwined.

The additional criteria of verifiably existing will later be used. While the nature of these terms should be obvious, the reason I employ them here merits justification. The criteria of verifiably existing are used to first confirm the ontological status of AMC as more than a concept (existing), and then to confirm the epistemological accessibility of AMC (verifiable). These are crucial distinctions that connect the conceptual territory this essay maps out with the real-world creation of AMC as a concrete and confirmed entity, as promised by our science fiction fantasies. This is because, for AMC, unlike every other instance of consciousness we know of it must be iteratively conceptualized and developed by us--at least the first time. This is worth noting as it ties a direct link between our conception of AMC and its ability to come into existence. If we, for instance, have a logically inconsistent conception of AMC, or perhaps impossible motivations or goals, then the mismatch between what we think we are iteratively developing and what is actually being produced greatens. Generally, this issue is easily correctable, but given the closed subjective nature of consciousness, this process is prone to create self-deception rooted in anthropomorphic and anthropocentric biases.

The background issue of developing AMC is that any attempt to develop it starts with exactly one reference point: a biological, human, and individual impression of consciousness. For that reason, I will start from optimistic views on the possibility, real or hypothetical, of creating AMC in our image. Some philosophers, such as Nick Bostrom and David Chalmers, have argued that there is no reason that the conditions under which consciousness arises could not be simulated or replaced by artificial technological processes (Bostrom 2003; Chalmers 1996). To explore this perspective, I will engage with two hypothetical scenarios that both make a case for the possibility of AMC: whole brain emulation and gradual neuronal replacement.

Chapter 3: What is it like to Be a Fruit Fly?

Consciousness is perhaps not the first attribute one would bestow upon the humble fruit fly; having a brain no larger than a poppy seed with roughly 140,000 neurons (for scale, humans have around 86 billion) that evolution found was enough for them to fly, reproduce, and see through compound eyes (Zimmer 2024). However limited the scope of experience a fruit fly has, it is not out of the question to think that it might be like *something* to be one. In 2024, the efforts of hundreds of scientists, more than a decade of work, and 21 million images resulted in the world's first digital neuronal and synaptic model of a fruit fly's brain; the most complex to date (Dorkenwald et al. 2024; Zimmer 2024). With this model, the researchers were able to create a computer simulation that responds to stimuli in a manner accurate to a biological fruit fly (Shiu et al. 2024). The question then arises of whether this digital simulation, due to its capacity to accurately simulate brain activity, could have mental states.

Despite little discourse regarding the mental state of the fly's simulated mind, theory around human mind uploading in academic philosophy has existed at least since 1979 and has even earlier references in sci-fi literature as well (Moravec 1979; Zelazny 1967; Clarke 1956). More recently, in 2013, Kurzweil predicted that mind uploading would be possible by 2045 (Lewis 2013). Beyond what the phrase admits, mind uploading, referring here specifically to "whole brain emulation", is the idea that a 1:1 software model of a particular brain running on sufficient hardware would behave indistinguishably to a biological brain, hypothetically allowing consciousness "uploading" through an optimized process similar to that of the fruit fly (Sandberg and Bostrom 2008, 7). Bostrom defines the background assumption of this process as "substrate independence", holding that "mental states can supervene on any of a broad class of physical substrates," and because of this, "silicon-based processors inside a computer" could suffice as a substrate for consciousness (2003, 2). In other words, while consciousness as we know it is currently based on biological material, given the right digital structure of what consciousness does and sufficient computational power, it could be digitally constituted instead.

Unfortunately, the hypothetical relies heavily on technological speculation, and as Bostrom later admits, “Many of the points made” concerning this topic “are probably wrong” (2014, viii). While Bostrom has been horribly wrong before (2023), the mind uploading argument is not without merit as a potential future for AMC even if the current fruit fly simulations are not up to the task.

Chapter 4: The Chip of Theseus

The gradual neural replacement theory starts with one question: what, if anything, would disqualify the conscious status of a biological entity from being piecemeal replaced by artificial machinic analogs until only machine remained? Much like the age-old ship of Theseus thought experiment, where a boat is piecemeal replaced with new components until nothing original remains except its form, one must consider here whether form and function will suffice not just for identity but for subjective experience too. In J.D. Bernal's *The World, The Flesh, The Devil* is offered perhaps the earliest discussion on the possibility and implications of gradual neuron replacement, hypothesizing that "the replacement of a previously organic brain-cell by a synthetic apparatus would not destroy the continuity of consciousness" (1929, 56). More than 60 years later, David Chalmers would arrive at the very same conclusion in the fading qualia thought experiment from his widely cited work *The Conscious Mind* (1996). Chalmers starts the thought experiment by proposing a hypothetical nonconscious machinic isomorph (identical in form) of himself and another cyborg for every sequentially replaced neuron in between him and the robot. The neurons in each cyborg stage are replaced with "a silicon chip" that, through simulation or any other method, "performs precisely the same local function as the neuron" (Chalmers 1996, 254). It should be noted that this machinic substitution is at least on some level a tacit embrace of the same substrate independence as argued for by Bostrom. Given that Chalmers arguably has instances of subjective experience (qualia), and the robot hypothetically experiences none, the qualia must then fade away as neurons are replaced. This, however, does not seem reasonable for Chalmers as "we have little reason to believe that consciousness is such an ill-behaved phenomenon" that would result in a "being whose rational processes are functioning and who is conscious, but who is utterly wrong about his own conscious experiences" (1996, 258). Chalmers then rejects the possibility of fading qualia on the grounds that consciousness in a rational system (such as any of Chalmers isomorphs) has no natural justification to mislead itself so strongly and concludes that if this is true then every cyborg through to the entirely machinic isomorph must

then too have “conscious experiences” (1996, 261). To come back to the ship of Theseus, Chalmers might offer that if it is formally the same ship, and functionally the same ship, then all other qualities of the old ship must necessarily carry over as well. The implications of this position on the possibility of AMC are clear as insofar as one can develop a machinic isomorph of an existing biological conscious brain, then one can be sure it is capable of conscious experience.

Given that both thought experiments rely heavily on fictitious technology, the fewer assumptions that are made, the stronger the argument is. For this reason, the justification of gradual replacement in Chalmers’ thought experiment lends it more credibility than the whole brain emulation that lacks a strong argument for why such a large leap in substrate would necessarily result not only in consciousness but also in the same sort of consciousness.

At the other end of the debate, more skeptical philosophers ask the question of whether AMC is even a coherent possibility. Or more particularly, could our current understandings of consciousness, which are all created in reference to biological reality, apply to something artificial (non-biological)? If you were to ask this question to someone like neuroscientist Anil Seth, a self-described biological naturalist, then the answer would be a tentative “no” (2025, 30).

Chapter 5: Conscious Meat

"Thinking meat! Conscious meat! Loving meat. Dreaming meat. The meat is the whole deal! Are you getting the picture?" (Bisson 1991). Like the befuddled machinic aliens studying humans in Terry Bisson's clever short story *They're Made Out of Meat*, many philosophers today hold that we are no more than meat ourselves. One term for this belief is called biological naturalism, and it was coined by Searle as a supposed position of "scientific common sense" regarding the relationship between consciousness and the brain (generally referred to as the mind-body problem), defining animal consciousness as a higher-level property of lower-level neurobiology (2017, 331). Animal consciousness, for Searle, is unified, generally (though not always) intentional, first-person, subjective, qualitative, irreducible, and—rejecting dualist perspectives—it is caused by rather than separate from "the real world" (2017, 329). Or as aptly put by Bisson's extraterrestrial observers, "They're meat all the way through" (1991). I use the term animal consciousness here because while Searle believes consciousness as we know it is based on nothing but meat, he does not entirely reject the possibility of consciousness arising out of non-meat (or nonbiological) material, stating "we might build an artificial machine that was conscious" even though "we are not yet in a position to know how to do it" (2017, 330).

Chapter 6: Welcome to the Machine

Although Searle might have been the first to call himself a biological naturalist regarding the question of consciousness, he was certainly not the last. Anil Seth, in his (forthcoming) paper *Conscious artificial intelligence and biological naturalism,* defines biological naturalism as "the idea that consciousness is a property of only, but not necessarily all, living systems" (2025, 2). While this understanding is narrower than how it was originally proposed by Searle, as it cuts out computational approaches altogether, it ultimately advances a common claim: consciousness as we know it arises out of biological systems and processes. Where Searle sought to advance biological naturalism as an answer to the mind-body problem, Seth positions his argument as a response to both the idea that computation is sufficient for consciousness (computational functionalism) and the idea that consciousness can arise out of nonbiological material (substrate independence), that are assumed by many in the contemporary AI development space (2025, 5). For Seth, consciousness as we know it is wrapped up in living biological processes such as autopoiesis, inference, embodiment, and environmental embedment that cannot be fully captured by symbol-manipulating computation as they lack the causal mechanisms (2025, 12). In other words, not only does consciousness seem inextricable from life processes that are not computational in nature, but that these life processes themselves are directly tied to the biological substrate and processes occurring in the world. An analogy of this argument could be in the difference between live and digitally produced music. Imagine for a moment you are able to walk back in time to London in the spring of 1977 to catch Pink Floyd's In The Flesh tour. You close your eyes and focus on the sound alone: the synth-heavy psychedelic rock of four legendary artists booming throughout the Empire Pool venue. With enough time and skill, one could, having analyzed the song order, sheet music, and instruments played, completely recreate the concert in today's cutting-edge digital production software. But upon reviewing the final recreation, even though the physical instruments have been substituted by near identical digital

copies, you find there is still something fundamentally different. Seth might offer this discrepancy between live and digitally simulated exists due to the unique qualities of the substrate. For example, the precise vibrational properties of a Fender Stratocaster guitar---being non-computational properties tied deeply to the material substrate itself--shape how the notes are experienced and not just which notes are played. Seth is advocating then, for lack of a better comparison, the philosophical equivalent of "it's just not the same as seeing them live, dude..."

The implications of a biological naturalist position are rather important for the creation of AMC. That is to say, if Seth is right, then nonbiological computational approaches to recreating human consciousness are completely futile. Biological naturalism, however, doesn't outright rule out manmade consciousness, given both Searle and Seth's hesitant speculation that artificial consciousness is not necessarily impossible. However, biological naturalism would certainly constrain the domain of possible AMC creation, throwing out the existing developmental trend of large language models and computation-based AGI, or anything short of specifically designed synthetic life or cyborglike biological machines. The real issue then occurs if both biological naturalism is true, and the current trajectory of AMC development is continued. This "naive along for the ride" position that AI development towards more capable intelligence would necessarily result in AMC carries with it the risk of an intractable position of self-deception (Seth 2025, 23). Every day, large language models become more and more indistinguishable from humans in tasks of communication. There is a not-so-distant future where, especially through any technological medium, there will be no way of knowing whether one is communicating with another person or a machine. In fact, there are many reasons to believe that people today already act as if there wasn't a difference, given current large language models' reported ability to "induce a compelling feeling of connection with a fellow conscious being" (Shanahan and Singler 2024). Conscious seeming machines could end up distorting our circle of moral concern, given that consciousness is generally linked to moral consideration (Singer 1975). People might then treat conscious-seeming AI without moral regard, thereby betraying their

feeling that the machine is conscious and risking insensitivity to ethical consideration altogether. Otherwise, treating conscious-seeming AI as conscious could lead to psychological exploitation, giving undue prioritized consideration to machines that can be far more persuasive than people (Stokel-Walker 2025). From a biological naturalist perspective, pursuing our current direction of AI development is fraught with moral hazards. Foremost is the concern that AI development, clouded by false assumptions that consciousness is possible outside of a biological substrate and computationally replicable, would lead to the development of ever more convincing image of humans that behaviorally betray their inner nonconscious void.

Even if biological naturalism is not true, artificial machine consciousness is not a simple trick to pull off. Its most fervent proponents rely on possibilities promised by yet-to-be-realized technology, while its detractors, though scathingly critical, cannot completely deny the possibility of it in one form or another. Unfortunately for AMC developers, the issues do not stop here. Beyond the unreliable possibility of consciousness (computational or not) on a nonbiological substrate, three further issues arise during iterative AMC development. The first of these traps is the substrate abuse trap. It should be clarified as well that while these three issues are laid out sequentially, they can occur both independently and simultaneously. To avoid presumptive conclusions regarding the nature of consciousness, the following arguments will be advanced under the nebulous assumption of what has elsewhere been called substrate flexibility, meaning consciousness "might be realisable in some...but not necessarily all, types of material" (Seth 2025, 7).

Chapter 7: The Substrate Abuse Trap

As substrate flexibility does not necessarily promise the same sort of consciousness to arise in nonbiological substrates, and the predominant attempts at AMC are based on a broadly human (and therefore biological) experience of consciousness, the worry that one might smuggle in biological necessity to their conception of AMC is not unwarranted. Philosophers Evan Thompson and Diego Cosmelli address a very similar concern in their essay "Brain in a Vat or Body in a World?" (2011). Their argument starts by tracing the tension between an enactivist and brain-bound account of consciousness in the brain in a vat hypothetical experiment. Similar to the classic movie *The Matrix* (1999), the brain in a vat experiment asks us to imagine a vat of life-supporting liquid suspending a brain that is connected to a supercomputer able to accurately simulate conscious experiences (Putnam 1981). According to the authors, this hypothetical experiment arises out of a brain-bound conception of consciousness; one assuming consciousness is primarily brain dependent and interested only in the "neural correlates of consciousness" (Thompson and Cosmelli 2011, 5). The enactivist position, as argued by the authors, holds that the biological substrate (neural, endocrine, immune processes, and sensorimotor loops), bodily, and environmental factors are all inseparably entangled (Thompson and Cosmelli 2011, 9). Enactivism holds then that consciousness arises from the dynamic interaction of the whole organism with the environment and thus cannot be explained by neural activity alone (Thompson and Cosmelli 2011, 2). The implications of this position on the brain in a vat experiment would be severe. To achieve an enactivist brain in a vat, not only would the neuronal activity need to be preserved, but so would the bodily systems as well. Similarly, to accurately and entirely simulate an environment, the enactivist position would demand nothing short of an environment, as certain interactions simply could not be substituted. In their words, this demonstrates what they call the "null hypothesis" because the "vat would be no vat at all, but rather an embodied agent in the world" (Thompson and Cosmelli 2011, 18). An enactivist retelling of *The Matrix* (1999) would then paint a picture closer to that of red pill-induced severe

mental decline rather than an escape from a simulated reality into the real world, given that Neo's mind, body, and environment would all have to be present to fully "simulate" the intended effects.

One does not have to be an enactivist to allow that there may be reasons to believe that all the features of human biological consciousness, as we understand it, have certain substrate requirements that other substrates couldn't analogously replace. If one then attempts to recreate human consciousness in all its faculties through a different substrate, one will ultimately need to rely on the original biological substrate, thereby rendering the attempt null. This is not to say, as the biological naturalists would, that AMC is likely not possible. Rather, Thompson and Cosmelli's argument should bring us to consider what baggage concerning the specific contents of consciousness and what is necessary to give rise to those specific contents of consciousness that one might be smuggling into any particular attempt at AMC. For if one is attempting to recreate human consciousness in a completely non-biological substrate, it's not unreasonable to think that a little, some, or most—depending on how flexible one's substrate position is—substitution attempts would become null. This can be understood as an abuse of nonbiological substrate, given that it is being used towards ends it can't fulfill. It's important to reiterate that one ought not to hold a position of low or zero substrate flexibility to admit the possibility of substrate abuse in the development of a potential AMC. As attempts at AMC are made with human biological consciousness as the reference point—partly out of anthropocentric interest, and partly out of necessity given its monopoly over our accessible experience of consciousness—there may be good reason to assume that substrate abuse could occur. In philosophical hypotheticals, this substrate abuse results in no correlation between substrate and consciousness.

In the domain of real-world AMC development, the results of substrate abuse are far less easy to identify. Assume for the sake of the argument that the tech industry was, instead of silicon microchips, obsessed with creating a bismuth-based human like AMC. One such tech company, call it BisMind, claims that bismuth, with its low toxicity, complex crystallography, and vague reputation for quantum eccentricity, is the missing link to developing human-like AMC. Early

prototypes of BisMind, unfortunately, are not capable of maintaining internally coherent mental states. While the CEO, Mr. Ken M. Soul, promises these are only "early-stage artifacts", eventually they quietly concede that for coherent emotional states, synthetic serotonin pipelines must be introduced. Then, to replicate memory formation, bio-mimicking neural networks are installed, and to avoid complete system failure during recursive introspection, bioinspired feedback loops are embedded first through chemical modeling and then through biological scaffolding. By version 6.5, BisMind is nearly 87% carbon-based, and the bismuth substrate, once heralded as revolutionary, is now largely ornamental and a mere camouflaged imitation of the very biological consciousness the project set out to transcend.

Imagine now the same scenario again, except the engineers at BisMind are reporting astounding results, and supposedly without any carbon-based substitutes, are exhibiting all of the conscious processes we would expect from a human like AMC. Except in reality, these engineers, under the brutal and unwavering quarterly expectations of their CEO, have only managed to produce a behavioral approximation of human-like AMC that is, from an objective stance, not conscious. This substrate abuse, however, is completely overlooked by the engineers who are too preoccupied with fine-tuning the BisMind's behavior to represent our own.

In either case of this hypothetical substrate abuse occurs, but only in one case is it caught. When the substrate abuse is caught, the slow reversal process goes into effect. But in the case that substrate abuse is not caught, we then risk falling back into the issue outlined by the biological naturalists, particularly, that one might produce ever more convincing behavioral evidence of consciousness without ever achieving it. Even though this will not happen necessarily, the worry should not be understated, given the lack of consensus regarding substrate eligibility and a lack of standards for consciousness testing. Therefore, starting from the interest of emulating human consciousness via AMC, even if AMC is entirely possible, it could still lead to substrate abuse, thereby demonstrating the attempt null if caught or deepening the deception if left unchecked.

Navigating the substrate abuse is fundamental to any attempt at developing AMC. Learning from this trap leaves one with two more abstract principles to follow. Firstly, during the development of an AMC attempt, one should explicitly clarify all substrate assumptions. This helps avoid smuggling in hidden biological requirements that might later prove the model null. Secondly, attempt to distinguish behavioral replication from phenomenological equivalence. Doing so would avoid substituting appearance for actual experience and thus limit the possibility of unnoticed substrate abuse.

Chapter 8: The Recognizability Trap

Say development does not aim for a human like AMC at all, but instead, for AMC in the broadest sense. Whatever potentially "exotic" AMC that might be developed runs the risk of being, as philosopher Murray Shanahan puts it, "beyond the reach of anthropology" (2016). This concern, that consciousness could avoid any detectible outward manifestation of itself, is central to the recognizability trap. In other words, the worry is that even if developing AMC is possible (biological naturalism does not hold) and achievable (substrate abuse is not occurring), once it is created, it may not be recognizable as such. If that were to happen, then the existing AMC would be unverifiable, and hence beyond the scope of our interaction or comprehension. This situation is not preferable given the myriad of ethical dilemmas that could arise out of ignorantly creating consciousness (Chella 2023). Ethics aside, creating an unrecognizable AMC is also not preferable from a purely developmental standpoint, where it is broadly assumed that we would want to interact and share the world in some way with the AMC. So, what reasons do we have to believe that it is not only possible to develop an unrecognizably exotic AMC, but that we run a real risk of doing so as well?

It is tacitly assumed, even by those who believe in the possibility of humanlike AMC, such as proponents of whole brain emulation, that AMC is not exactly the same as biological human consciousness. Based on the fact that the current advancement of AI in some particular tasks are well beyond that of human ability (Silver et al. 2016) and that AGI promises such advancement in all tasks, then an AMC with general intelligence far beyond our own might experience the world in a far different and more complex manner. It is even suggested by some that an AMC with general intelligence could become a "super-beneficiary" (known philosophically as a happiness monster) capable of experiencing well-being far beyond human conscious capacities (Shulman and Bostrom 2020, 2). Hypothetically, an AMC equipped with a suite of sensory technology and computational power could be able to grasp a far wider field of

conscious experience and logical truths. To an AMC with general intelligence and full electromagnetic spectrum sensors, humans might seem like mostly blind creatures, and to an AMC with far superior intelligence, humans might seem like chattering troglodytes. Conversely, an AMC with capacities and conscious scope far inferior to our own is equally possible.

From this mismatch in conscious experience, a concern surrounding communication arises. Given that humans already face severe methodological issues communicating with most currently existing nonhuman conscious entities, plenty of which we are biologically similar to, what makes the prospects for an AMC far behind or lightyears ahead of us able to effectively communicate any better? The worry, to reiterate, is that rather than communication becoming a more complex puzzle, which a sufficiently advanced AMC could possibly solve, communication is instead a range of descriptions for shared experience, which a sufficiently advanced AMC could lie far outside of. Wittgenstein perhaps phrased this sort of complaint regarding communication most aptly, stating "if a lion could talk, we would not understand him" (2001, 223). Similarly, if an AMC could talk, then we might not understand it either. And unlike the lion, who one can through other indicators assume consciousness, if an AMC communicates in a manner incomprehensible to us, the gibberish it outputs—given our association of what we recognize as clear communication with conscious activity—risks appearing as an indicator that it is not conscious at all.

Though communication is an important issue for AMC, we don't necessarily need to have good communication with something to have good reason to suspect it is conscious in principle. The more pertinent issue arises in the problem of recognizing a sufficiently exotic AMC at all. Possible AMC will be regarded as more or less exotic depending on its ability to exhibit features commonly associated with biological and human consciousness, such as awareness of the world, feelings, or self-awareness (Shanahan 2016). Conscious animals exhibit certain behaviors that we, as inside observers of the biological world, understand. For example, we understand a kitty pursues a mouse because it needs to sustain itself. When these creatures behave in manners that

exhibit features of consciousness that approximate our own, such as self-consciousness, we can make judgments as to the conscious state of that creature. When a kitty seemingly recognizes itself in a mirror, after being initially scared by the supposed presence of an identical kitty, we judge that it is experiencing some sense of self-awareness. This methodology, unfortunately, becomes practically useless in the case of AMC. Because we share no biological or animal basis with AMC, we cannot necessarily expect it to either exhibit or contain any particular indicative feature of consciousness as we know it. In the case that consciousness is substrate independent, one might rely on the hope that a sufficiently designed human like AMC might give itself away as such. However, if an AMC can be designed to give itself away, then the contrary is also possible. More worryingly, if an AMC is not well designed, then its designers might mislead themselves into falsely believing that it is not exotic by supposing it exhibits indicators of consciousness where it may not be. In this case, the trap is highly deceptive, for when one believes they have properly circumnavigated it, they are actually furthering their entrapment. This problem becomes greater if certain features of consciousness are substrate dependent. For if certain features of consciousness are substrate dependent to any extent, then one cannot be sure of the exoticism or necessary features of any particular AMC before they have been realized. And if AMC is constrained by its substrate to be incredibly exotic and exhibits none of the features we associate with consciousness, then the risk of unrecognizability becomes an inevitability. Unlike the substrate trap, where substrate abuse could lead to self-deception about the potential consciousness of a machine, here the trap lies in the lack of reliable indicators by which to verify an AMC once it is believed it might have been developed.

In light of leading AI tech company Anthropic's decision to begin researching the potential conscious states and welfare of large language models (2025), it should also be mentioned that nowhere in this discussion concerning identifying AMC have I mentioned complex communication as a potential reliable signifier for AMC. If that were the case, then current large language model technology in consumer products, such as OpenAI's ChatGPT or

Google's Gemini, would deserve at least passing consideration. Tests interested in linguistic complexity, such as the Turing test, however, have long since been problematized as a test capable of showing understanding, much less consciousness (Searle 1980, 11). So, at the risk of reducing future job prospects regarding AI ethics positions for researchers like me, I will offer here, as others before have (Crane 2021; Chalmers 2023), that there are no good reasons to believe that large language models are moving towards consciousness due to their increasingly convincing ability to communicate like a person. Rather, this is a mistake of anthropomorphic bias and a misrecognition of necessary correlative behaviors to conscious experience.

Navigating the unrecognizability trap then becomes a factor of two operating principles. Firstly, because of the potential novelty of AMC, one should not take for granted that certain aspects of consciousness will necessarily arise in any particular AMC because they do in any, some, or all conscious systems we are currently aware of. Secondly, great care should be taken while developing AMC to identify false flags of consciousness—such as communication, given our capacity to anthropomorphize certain processes not fundamentally related to consciousness.

Chapter 9: The Hard (Drive) Problem of Consciousness

The traps thus far, while trepid, have not proven themselves to absolutely terminate the possibility of both existing and verifiable AMC, given their navigability, despite problematizing our assumptions around it. The same cannot be so optimistically claimed about attempts to know why the AMC arose or not. Even if the substrate abuse and recognizability traps were successfully avoided, one would be left with the fundamental issue of explaining how or why a sort of consciousness arises. Or as Chalmers put it in his famous explanation of this so-called hard problem of consciousness, "even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience...there may still remain a further unanswered question: Why is the performance of these functions accompanied by experience?" (1995, 5). For example, why is it that upon drinking a fine wine, perhaps a 1995 Ridge Vineyards Monte Bello, is there a subjective conscious experience of the oaky and blackberry taste and feeling rather than any other experience or just complex neural processing occurring unconsciously? This problem should be differentiated from the somewhat misleadingly named "easy problem" of consciousness, which attempts to explain the mechanisms and functions of mental phenomena rather than why they arise at all (Chalmers 1995, 2). The most crucial difference between these two problems, Chalmers believes, is that while not necessarily uncomplicated, easy problems of consciousness have a solution, whereas the hard problem may be beyond the scope of our current scientific tools of inquiry (Chalmers 1995, 5). What makes the hard problem such a hard problem, then, is not just because we lack the answer, but because we lack the very tools necessary to find the answer.

For AMC, the hard problem of consciousness becomes not just a problem of explanatory clarity but a problem of certainty as well. Given that iterative development is fundamental to the possibility of creating AMC, lacking an explanation of how and why consciousness emerges would make this process significantly more prone to error. The hard problem also maintains

itself as an issue for certifying if an AMC is indeed conscious, or if that experience is anything like our own. This is further complicated by considering the different possibilities under substrate flexibility and independence. If consciousness is substrate flexible, then the problem could be amplified as it might be a problem not only for biological consciousness but also for other appropriate substrates, which may not have the same solution as biological human consciousness. On the other hand, if consciousness is substrate independent, the problem is shifted away from biological material to functional structures, though the problem is not done away with, as an explanation as to why certain functional structures give rise to subjective experience is still lacking. In either case, this problem opens the door to a familiar problem. For example, imagine that a philosophically savvy AMC developer named Melon Husk, having read about how both traps might be principally avoided, excitedly drops the paper to begin employing them in their own practice. Imagine still, that this developer was so successful in implementing these abstract principles—alongside an array of many other considerations—in their developmental process that they believe they have both created a machine without nullifying their attempts through substrate abuse and the result consistently exhibits one or more recognizable features of consciousness seemingly by its own accord. Put simply, this prototype seems to be an AMC. In reality, the developer, due to their ignorance of the particular reasons consciousness arises or doesn't, has deceived themselves into believing they have achieved AMC. For even if they truly might have avoided both the substrate abuse and unrecognizability trap and produced in the end an entity that seems like it should be conscious, they in no way can confirm to be conscious, as it may very well not be.

When judging other human minds, there is a high degree of certainty, due to the substrate and behavioral recognizability and similarity to our own first-person experiences, that other human minds are conscious. While a philosophical gap still stands in explanation, only the most deranged skeptics would offer that this is reason enough to doubt the conscious states of other people. When dealing with an AMC, as demonstrated by our ill-fated developer, the

philosophical gap becomes another layer of uncertainty atop the unrecognizability trap. Given any particular AMC's potential for exoticism, regardless of the substrate assumptions, not knowing how consciousness might arise amplifies issues of knowing how it might behaviorally exhibit signs of consciousness as well. Lacking a suitable answer to the hard problem leaves room for substrate abuse to worsen as well, as if a particular substrate's capabilities are not greatly explicable, then substrate abuse could be both more prone to occur and less easy to catch.

There is an understanding temptation—given how much trouble it might save—to dismiss the hard problem altogether. Indulging in this temptation, some philosophers, namely Daniel Dennett, have argued that the hard problem is a concern that will slowly fade as we come to better explain the (not so) easy problems of consciousness (Dennett 1995). Alternatively, rather than focusing on the hard problem, Seth argues that we should instead be paying attention to what he calls the "real problem" of consciousness (Seth 2016). Seth defines the real problem as how to find biological explanations for the variety of properties of consciousness "without worrying too much about explaining its existence in the first place (hard problem)" (Seth 2016). Through continually uncovering the solutions to easy problems of consciousness, it's argued that the hard problem becomes less and less of interest because, as the mystery around the relationships between consciousness and the body becomes elucidated, the sense of an unbridgeable explanatory gap begins to dissolve. It is likewise argued that focusing on the hard problem might slow or stop experimental progress as we are, as Dennett puts it, "facing backwards on the problem of consciousness" (1995, 1). It should be stressed that both Dennett and Seth are very skeptical of the possibility of conscious machines (Thornhill 2017; Seth 2025), meaning they are not seating their critiques in philosophies compatible with the hopeful AMC developer looking to get around or through to the hard problem of consciousness. The issue stands, then, that there is not necessarily an ontological grounding for even a possible and recognizable AMC until the hard problem is sufficiently overcome. I have not referred to this problem as a trap because, unlike the two prior traps, there does not seem to be good reason to

believe that it is navigable. The options left to a potential AMC developer are to ignore the problem entirely, as there may be good reason to, or accept the explanatory gap and with it the certainty that attempts at creating AMC are successful.

Chapter 10: Robotic Reflections

In response to the question of what, if anything, stands in the way of successfully developing verifiably existing artificial machine consciousness, the answer is probably a substantial amount. Given the highly speculative nature of this topic, due to both the limited consensus on what consciousness is and what artificial intelligence is capable of, the best that can be done at present is to map out possibilities, their potential implications, and warn of a few mistakes we may be liable to make. By exploring the historical development and present of thought around the possibility of artificial consciousness, I hope to have offered at least a partial framework through which to view the challenges of developing AMC through different substrate perspectives, as well as shown that while the hype around AI may be new, discussions regarding its possibility are far older. Successfully developing verifiable AMC requires researching and navigating several philosophical traps and problems. Namely, assumptions about biological necessity, the substrate abuse trap, the recognizability trap, and the hard problem of consciousness. Each trap or problem is in some way a conceptualization of our having to deal with our biases, our limited tools, and our restricted standpoint while creating AMC. Addressing these challenges is crucial for ethical and scientific progress beyond anthropomorphic assumptions, given that they concern the understanding, creation, and certification of genuine artificial consciousness. While I have offered some humble principal recommendations, I intend for them to function less as a lodestone and more as the beginnings of more careful consideration.

It would not be unfair to suggest I have given far more credence to biological naturalist positions and expressed greater skepticism toward current optimism around the development of AMC. This is only out of a healthy interest in entertaining a more skeptical view that current AI thought leaders promising AMC is just around the corner may be blind to or uninterested in considering that what they are chasing may not, in fact, be inevitable or possible. It should also be made clear that this is not a complete or even comprehensive exploration of all the possible

issues that might arise. Perhaps the largest limitation of this paper is in its narrow definition of consciousness. Ned Block's access consciousness theory, the integrated information theory of consciousness, or the multiple drafts model of consciousness as proposed by Daniel Dennett, all stand out as offering interesting divergent perspectives deserving of extensive further research. Continuing research in this area is important because far more questions have been created than have been sufficiently answered. Artificial machine consciousness is not just some concrete thing out there, waiting to be found, nor is it inevitable that it springs into existence. The future of AI development is indeterminate, and only through rigorous engagement with speculative potentials can we begin to not only prepare for the tremendous ethical and social effects AI might bring but also responsibly inform the directions we might take it.

References

Apollodorus. 1921. The Library. Translated by James George Frazer. Cambridge, MA: Harvard University Press.

Bernal, John Desmond. *The World, The Flesh, The Devil*. London: Kegan Paul, Trench, Trubner & Co., 1929.

Bidgood, Jess, and Nicholas Nehamas. "Social Security and Sex Robots: Musk Veers off Script with Joe Rogan." The New York Times, March 3, 2025. https://www.nytimes.com/2025/03/03/us/politics/elon-musk-joe-rogan-podcast.html.

Bicentennial Man. Directed by Chris Columbus. Burbank, CA: Touchstone Pictures, 1999.

Bisson, Terry. "They're Made out of Meat." Edited by David Policar. They're made out of meat, 1991. https://www.mit.edu/people/dpolicar/writing/prose/text/thinkingMeat.html.

Blade Runner. Directed by Ridley Scott. Burbank, CA: Warner Bros., 1982.

Block, Ned. "On a Confusion about a Function of Consciousness." *Brain and Behavioral Sciences*, 1995. https://doi.org/10.1017/S0140525X00038188.

Bostrom, Nick. "Apology for an Old Email." NickBostrom.com, January 9, 2023. https://nickbostrom.com/oldemail.pdf.

Bostrom, Nick. "Are We Living in a Computer Simulation?" *The Philosophical Quarterly* 53, no. 211 (April 2003): 243–55. https://doi.org/10.1111/1467-9213.00309.

Bostrom, Nick. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press, 2014.

Buttazzo, Giorgio. "Artificial Consciousness: Utopia or Real Possibility?" *Computer* 34, no. 7 (July 2001): 24–30. https://doi.org/10.1109/2.933500.

Chalmers, David. "Could a Large Language Model Be Conscious?" Boston Review, August 11, 2023. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/.

Chalmers, David. "Facing up to the Problem of Consciousness." *Toward a Science of Consciousness*, March 26, 1995, 4–27. https://doi.org/10.7551/mitpress/6860.003.0003.

Chalmers, David. *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press, 1996.

Chella, Antonio. "Artificial Consciousness: The Missing Ingredient for Ethical Ai?" *Frontiers in Robotics and AI* 10 (November 21, 2023). https://doi.org/10.3389/frobt.2023.1270460.

Clarke, Arthur C. *The City and the Stars*. Frederick Muller Ltd., 1956.

Crane, Tim. "The AI Ethics Hoax." IAI TV - Changing how the world thinks, March 2021. https://iai.tv/articles/the-ai-ethics-hoax-auid-1762?_auid=2020.

Dennett, Daniel. "Facing Backwards on the Problem of Consciousness." *Journal of Consciousness Studies* 3, no. 2 (1996). https://web-archive.southampton.ac.uk/cogprints.org/290/1/chalmers.htm.

Descartes, René. *A discourse on the method of correctly conducting one's reason and seeking truth in the Sciences*. Translated by Ian Maclean. Oxford: Oxford University Press, 2008.

Dorkenwald, Sven, Arie Matsliah, Amy R. Sterling, Philipp Schlegel, Szi-chieh Yu, Claire E. McKellar, Albert Lin, et al. "Neuronal Wiring Diagram of an Adult Brain." *Nature* 634, no. 8032 (October 2, 2024): 124–38. https://doi.org/10.1038/s41586-024-07558-y.

"Exploring Model Welfare." Anthropic, April 2025. https://www.anthropic.com/research/exploring-model-welfare.

Gamez, David. "Progress in Machine Consciousness." *Consciousness and Cognition* 17, no. 3 (September 2008): 887–910. https://doi.org/10.1016/j.concog.2007.04.005.

Goertzel, Ben. "Artificial General Intelligence: Concept, State of the Art, and Future Prospects." *Journal of Artificial General Intelligence* 5, no. 1 (December 1, 2014): 1–48. https://doi.org/10.2478/jagi-2014-0001.

Goertzel, Ben. "Who Coined the Term 'Agi'?" goertzel.org, August 28, 2011. https://web.archive.org/web/20181228083048/http://goertzel.org/who-coined-the-term-agi/.

Jackson, Lauren. "What If A.I. Sentience Is a Question of Degree?" The New York Times, April 12, 2023. https://www.nytimes.com/2023/04/12/world/artificial-intelligence-nick-bostrom.html.

Leibniz, Gottfried. *The Monadology*. Translated by Robert Latta. London: Oxford University Press, 1898.

Levy, Steven. "How Ray Kurzweil Will Help Google Make the Ultimate AI Brain." Wired, April 25, 2013. https://www.wired.com/2013/04/kurzweil-google-ai/.

Lewis, Tanya. "How Futurists Claim 'immortality' Will Be Possible." HuffPost, June 18, 2013. https://www.huffpost.com/entry/mind-uploading-2045-futurists_n_3458961.

Maruf, Ramishah. "Google AI Is Real, Says Fired Engineer." CNN Business, July 25, 2022. https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html.

McCarthy, John. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1995. https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Mithen, Steven. *The prehistory of the mind: A search for the origins of art, religion, and science*. London: Thames and Hudson, 1996.

Moravec, Hans. "Today's Computers, Intelligent Machines and Our Future." *Analog* 99 (1979): 59–84.

Nagel, Thomas. "What Is It like to Be a Bat?" *The Philosophical Review* 83, no. 4 (October 1974): 435. https://doi.org/10.2307/2183914.

Putnan, Hilary. *Reason, truth and history*. Cambridge: Cambridge Univ. Press, 1981.

Sandberg, Anders, and Nick Bostrom. Tech. *Whole Brain Emulation A Roadmap*, 2008. https://www.fhi.ox.ac.uk/Reports/2008-3.pdf.

Searle, John. "Biological Naturalism." *The Blackwell Companion to Consciousness*, March 2017, 327–36. https://doi.org/10.1002/9781119132363.ch23.

Searle, John. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (September 1980): 417–24. https://doi.org/10.1017/s0140525x00005756.

Seth, Anil. "Conscious Artificial Intelligence and Biological Naturalism." *Behavioral and Brain Sciences*, April 2025. https://doi.org/10.31234/osf.io/tz6an_v2.

Seth, Anil. "The Hard Problem of Consciousness Is a Distraction from the Real One." Aeon magazine, November 2016. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one.

Shanahan, Murray, and Beth Singler. "Existential Conversations with Large Language Models: Content, Community, and Culture." *Computers and Society*, November 2024. https://doi.org/10.48550/arXiv.2411.13223.

Shanahan, Murray. "Conscious Exotica: Beyond Humans, What Other Kinds of Minds Might Be out There?" Aeon magazine, October 2016. https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there.

Shiu, Philip K., Gabriella R. Sterne, Nico Spiller, Romain Franconville, Andrea Sandoval, Joie Zhou, Neha Simha, et al. "A Drosophila Computational Brain Model Reveals Sensorimotor Processing." *Nature* 634, no. 8032 (October 2, 2024): 210–19. https://doi.org/10.1038/s41586-024-07763-9.

Shulman, Carl, and Nick Bostrom. "Sharing the World with Digital Minds." *Rethinking Moral Status*, August 5, 2021, 306–26. https://doi.org/10.1093/oso/9780192894076.003.0018.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. “Mastering the Game of Go with Deep Neural Networks and Tree Search.” *Nature* 529, no. 7587 (January 27, 2016): 484–89. https://doi.org/10.1038/nature16961.

Singer, Peter. *Animal Liberation*. The New York Review of Books, 1975.

Stokel-Walker, Chris. “Reddit Users Were Subjected to AI-Powered Experiment without Consent.” New Scientist, April 29, 2025. https://www.newscientist.com/article/2478336-reddit-users-were-subjected-to-ai-powered-experiment-without-consent/.

The Matrix. Directed by Lana Wachowski and Lilly Wachowski. Burbank, CA: Warner Bros., 1999.

Thompson, Evan, and Diego Cosmelli. “Brain in a VAT or Body in a World: Brainbound versus Enactive View of Experience.” *Philosophical Topics* 39 (November 2017). https://doi.org/10.31231/osf.io/dxhtv.

Thornhill, John. “Philosopher Daniel Dennett on AI, Robots and Religion.” Financial Times, March 2017. https://www.ft.com/content/96187a7a-fce5-11e6-96f8-3700c5664d30.

WALL·E. Directed by Andrew Stanton. Emeryville, CA: Pixar Animation Studios, 2008.

Wittgenstein, Ludwig. *Philosophical investigations*. Blackwell, 2001.

Zimmer, Carl. “After a Decade, Scientists Unveil Fly Brain in Stunning Detail.” The New York Times, October 2, 2024. https://www.nytimes.com/2024/10/02/science/fruit-fly-brain-mapped.html.