The Misguided Conflation of Epistemic Ontology and Epistemic Onticism in AGI Research

Ha, You Jeen

Smith College, Department of Computer Science and Department of Philosophy

Corresponding Author: You Jeen Ha; yha@smith.edu

Abstract: Artificial general intelligence (AGI) appears to have a specific metaethical character, a character defined by certain metaphysical and epistemological commitments. Drawing upon Martin Heidegger's metaphysical distinction between ontology and onticism, I will argue that underlying AGI research is a misguided conflation of epistemic ontology and epistemic onticism: that knowing the nature of consciousness results from knowing how consciousness operates and *functions*. I will explore literature on Daniel Dennett's functionalist view of consciousness, computationalism, Heideggerean existentialism, and contemporary Western philosophy to show the prevalence of this conflation. I will also argue that this conflation both deprecates and elevates the human, dehumanizing human beings as well as enabling humans to self-apotheosize and "play God." Thus, my paper will call for the reevaluation of the commitments that have persisted in AI research as well as the exercise of caution moving forward through a Heideggerean reading of Mary Shelley's *Frankenstein: or The Modern Prometheus* (1818).

Keywords: Artificial general intelligence, Metaethics, Consciousness, Mind-body problem, Computationalism, Epistemic ontology, Epistemic onticism, "Playing God", Frankenstein

Citation:

Introduction

As artificial intelligence research makes headway and the wonder of its exponential progress grasps the attention of researchers, programmers, scholars, and the general public, it is understandable that the potential of artificial intelligence fuels much excitement. Recent developments and research are now seemingly forging the road toward *artificial general intelligence* (AGI) or *Strong AI*, a human-level mind of its own rather than simply a tool in the study of the mind (Searle, 1980; Turing, 1950). As agreed upon by numerous scholars, the possible coming of AGI prompts the need for developmental forecasting (Brundage et al., 2018). Here enter ethics.

However, it is not enough to ask "How can we ethically create artificial general intelligence?" or "How can we create moral agents with artificial general intelligence?" (Bostrom & Yudkowsky, 2014). I see three structural elements to the forecasting process, and the two aforementioned questions only constitute the latter part of the process. There is one essential element that is missing, and that element, viz., the question "Are the fundamental *assumptions* underlying AGI research ethical?" To be even more specific, this question is not simply a matter of ethics; rather, it is a matter of metaethics in that the question must examine central metaphysical and epistemological commitments prior to exploring the moral domain. Hence, in this paper, I will offer a triangular dynamic between metaphysics, epistemology, and ethics that ultimately forms at its intersectional core a metaethical exploration of René Descartes' "mind-body problem" in today's AGI research.

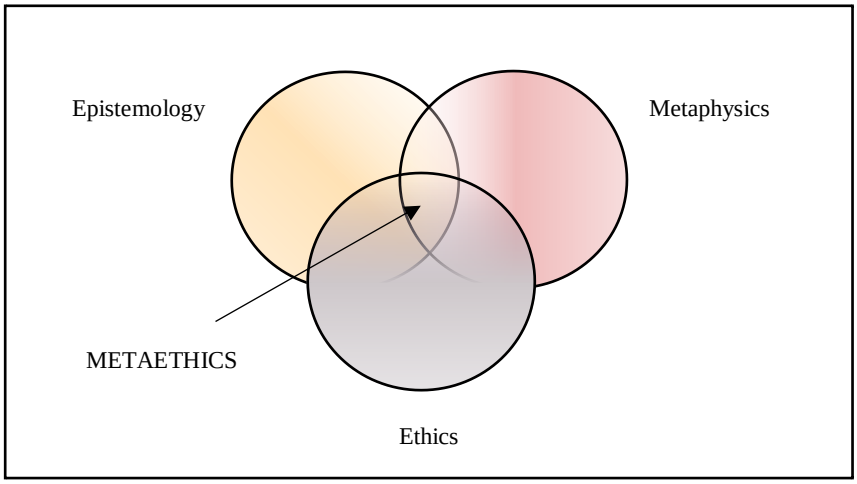

Figure 1: A visual representation of the thesis I intend to formulate and defend.

Upon first elucidating the metaphysical and epistemological assumptions that pervade a significant section of current AGI research, I will then analyze the ethical implications of these conceptual frameworks were we to accept them as the formal premises of AGI research. Finally, I will present a consolidated metaethical perspective that completes the

triangular dynamic. As I develop my arguments, I will refer to various literature on computational theories of consciousness, Heideggerean existentialism, and contemporary Western philosophy. I hope my paper can present a cautionary view and effectively convince my readers that if current AGI research continues to accept and pursue these assumptions, they will indubitably threaten the future of human-machine interaction by tilting the scale in favor of machines over humans and devaluing human existence. Artificial general intelligence would then truly become an "existential risk", albeit in a more philosophical sense rather than a populational sense (Bostrom, 2013, p. 15) in that humans lose what it means to be human.

In order to set the stage for my paper, I feel it necessary to briefly clarify key terminology as well as the context in which I will conduct my examination. With artificial general intelligence often comes a discourse on machine consciousness, particularly as to whether or not such intelligence would give rise to "virtual consciousness" (McDermott, 2001, p. 131). Frequently, the nature of machine consciousness is defined by many computer scientists and philosophers, whose claims I will further explicate, in terms of information processing and algorithmic computation. My examination into this metaphysical definition, which can also be called *computationalism*, will be primarily delineated through Martin Heidegger's (2010) distinction between ontology and onticism in his magnum opus *Being and Time.* According to Heidegger, ontology and onticism are *essence* and *empirical fact*, respectively. Ontology is the study of an ineffable way of being or existing whereas onticism is the mere appearance of being or existence; I interpret the latter to be a relative of functionalism, a philosophical view that defines mental states based on their causality, or, in another vein, their inputs and outputs (Fodor 1981).

I will argue, first through an examination of Daniel Dennett's "Imagining Consciousness" (1992), that computationalism is a form of onticism, yet computationalism has traditionally been taken as a form of ontology. This conflation of ontology and onticism is systematically problematic. Subsequently, I will incorporate epistemology into the mix and point to the difference between the metaphysical *being* of an entity and the epistemic *understanding* of that entity's being. Our particular *understanding* and *supposed knowledge* of the metaphysics of AGI, which we currently take as what I call "epistemic ontology", perpetuates the functionalist view that humans are inherently no different from machines, which I call a form of "epistemic onticism." It is precisely here that the topic of ethics becomes shaky. Ethics is a *human* matter, and when the meaning of human existence dissolves with present investments in creating an artificial anthropomorphic entity, the field of ethics itself is threatened and approaches the brink of irrelevance. Put simply, there is no use for ethics if being human is reduced to the extent that our existence has no distinct meaning for us. My Heideggerean analysis of Mary Shelley's *Frankenstein: or the Modern Prometheus* at the end of my paper will speak to this point.

Intelligence and Consciousness: Mutually Exclusive or Not?

Before engaging in a discussion on computationalism as a theory of consciousness, let alone a discussion on the mind-body problem, one must first situate intelligence and consciousness in relation to one another. Are intelligence and consciousness mutually exclusive or not?

For the sake of my argument, I will distinguish between the two with the former as having a *computational* or "easy" component, and the latter as having an *undefined* or "hard" component (Chalmers, 2010, p. 4). Here, I draw on Chalmers' "easy" and "hard" problems of consciousness, but I am not claiming that intelligence is a form of consciousness; I am merely employing the characteristics that Chalmers associates with "easy" and "hard." Depending on the context, intelligence and consciousness could be mutually exclusive as well as not. As to the character of this context, it is defined by the "agent" in question, namely whether or not the agent is a human being or a machine. For humans, the two are not mutually exclusive whereas for machines, the two are mutually exclusive *if* we understand consciousness as *phenomenal consciousness.* Yet, very few AI researchers care about phenomenal consciousness (McDermott, 2007). Perhaps, it is the case that

> *the aspect of the brain that is most likely to be exempt from the computationalist hypothesis is its ability to produce consciousness, that is, to experience things [in terms of Chalmers' "Hard Problem", which is] the problem of explaining how it is that a physical system can have vivid experiences with seemingly intrinsic 'qualities'"* (McDermott, 2007, p. 2).

If so, the computationalist hypothesis lacks an explanation as to how the brain produces consciousness (an explanatory absence that McDermott hopes to fill in *Mind and Mechanism* (2001) with his own computational theory of consciousness).

Yet, I attribute this explanatory absence to the distinction I clarified earlier: the computational, easy problem of intelligence and the undefined, hard problem of phenomenal consciousness are mutually exclusive for AGI. No wonder the computationalist hypothesis is wanting in consciousness-related accounts. The computational facet of intelligence and phenomenal consciousness cannot be reconciled. In fact, the two cannot be *conflated* in the sense that intelligence is confused with consciousness, and, therefore, views such as that of McDermott cannot stand.

Hilary Putnam (1992) would even go as far as to demolish the field of artificial intelligence itself as he does in *Renewing Philosophy*. He dismantles the intelligence portion of "artificial intelligence", convinced that a programmable simulation of

intelligence, much less consciousness, is simply unfeasible. To Putnam, artificial intelligence is

> *the search for a few simple algorithms that explain intelligence… [and] doesn't really try to simulate intelligence at all; simulating intelligence is only its notional activity, while its real activity is just writing clever programs for a variety of tasks* (McDermott, 2001, p. 88).

At worst, the endeavors pursued by AI researchers are futile, and the question that this section is attempting to address is irrelevant. Moreover,

> *[a]lthough one might expect AI researchers to adopt a computationalist position on most issues, they tend to shy away from questions about consciousness [and i]n view of [this] shyness…, it is not surprising that detailed proposals about phenomenal consciousness from this group should be few and far between* (McDermott, 2007, pp. 3-5).

Most serious AI researchers prioritize other problems other than that of consciousness, a prioritization that leads to said shyness (McDermott, 2007). I do agree with McDermott that phenomenal consciousness ought to be further explored in the field of artificial intelligence and indubitably artificial general intelligence, albeit not from his proposed vantage point. In his proposal, he builds upon Daniel Dennett's account of intentionality, but there are significant flaws in Dennett's account that must be addressed.

Metaphysical Commitments

Ontology versus Onticism with Martin Heidegger and Daniel Dennett

In philosophy of mind, Cartesian dualism, otherwise known as "the mind-body problem", is a prominent point of contention for Descartes' successors and their opponents. When addressing the mind-body problem, should one adopt a dualist view or, instead, adopt a physicalist view, which collapses the binary? Daniel Dennett (1992) steps outside of this dichotomy and pursues a functionalist take on consciousness, presuming that as long as the human mind can be understood to function as does a computer, one can understand consciousness. Modern metaphors facilitated by imagination, therefore, can move discussions on consciousness forward, away from the mind-body problem.

Dennett wishes to show how *imagining* consciousness leads to progress on *understanding* consciousness. His project takes the incredulous into consideration since they, on the other hand, believe that imagining a conscious robot, for instance, is simply impossible. However, Dennett makes a key distinction here: it is not the act of imagining that is the issue, but how to procure the details necessary to formulate an adequate image of a conscious robot. So,

> *by thinking of our brains as information-processing systems, we can*

> *gradually dispel the fog [that is the ambiguous gap between subjective interpretations and scientific knowledge of the brain] and pick our way across [this] divide* (Dennett, 1992, p. 433).

By embracing a functionalist view of *how* the human mind operates, one would be able to directly grasp the ontology of consciousness and replace the dualistic picture with an empirical understanding of cerebral activity. Dennett is convinced that this perspective is both sufficient and necessary in order to comprehend the traditionally enigmatic and rather elusive phenomenon that is consciousness.

Central to Dennett's argument is the "Cartesian Theater," where there are "onstage experiences and backstage processes" (p. 433). This thespian metaphor illustrates Descartes' dualism: "onstage experiences" are analogous to the observable physical phenomena and "backstage processes" are analogous to invisible, intangible mental phenomena, which are thought to direct our actions and behaviors. While Dennett concedes the allure of the Cartesian Theater, he dismisses the entire existence of such a theater and thus dualism, adamant that features of phenomenology such as qualia are not as "obvious" as they ostensibly seem to be (p. 433). In fact, he attributes a kind of inexplicable, mysterious quality to these phenomenological features and thereby spurns them.

Dennett then turns to Thomas Nagel's (1974) famous thought experiment where Nagel argues that the experience of what it is like to be a bat cannot be conceived of by the human mind in the form of "conscious mental states" since the experience is subjective and *unique* to the bat (p. 436). To Nagel, this experience, as such, cannot be explained reductively in terms of functional states; otherwise, the holistic value of the experience would be diminished so that it fits a particular functional picture and, consequently, exclude elements of the original experience deemed irrelevant.

Nagel's position places Nagel in the same group as those who believe that imagining a conscious computer is an "obvious" impossibility. Dennett instead suggests focusing on the observable features of a bat's consciousness, not whether human minds can be turned into bat minds. According to Dennett, Nagel is too concerned about *being* the bat rather than understanding the *how*. In other words, Nagel is too concerned about the ontology of the bat rather than understanding the onticism of the bat. The crux of Dennett's discussion on consciousness is, therefore, not ontological, but ontic in the Heideggerean sense à la functionalism: it is not the *nature* of consciousness that is central to our examination, but *how* consciousness operates.

The ontic-ontological distinction can be better understood through the definition of the ontic:

> *Dasein is a being that does not simply occur among other beings. Rather it is ontically distinguished by the fact that in its being this being*

> *is concerned about its very being...The question of existence is an ontic "affair" of Dasein* (Heidegger, 2010, p. 11).

Any study into the particularities of existence are *ontic*, whereas being, or the essence of existence, is an ontological pursuit constantly sought after in such studies. Heidegger further clarifies the distinction:

> *[w]e can describe the 'outward appearance' of these beings and tell of the events occurring with them [and this] description gets stuck in beings. It is ontic. But we are, after all, seeking being [that is, of the ontological kind]* (Heidegger, 2010, p. 63).

Hence, Dennett (1992) dispels the nebulous facet of understanding consciousness by moving away from the ontological and focusing on the ontic. In his view, "Software, Virtual Machines, Multiple Drafts" are ontic metaphors that break away from the ontological Cartesian Theater and enable one to effectively imagine consciousness (p. 455). Without Dennett's metaphors, imagining consciousness in a non-Cartesian, functionalist light would have been rather difficult.

Indeed, Dennett's aim in his paper is substantive in that it urges philosophers to leave the Cartesian domain of consciousness, shaking them out of their dualist comfort zone, and pursue other possible ontological approaches to consciousness. A functionalist *modus operandi* may simply be enough. Yet, Dennett's project is attempting to reach an ontological conclusion *with* the ontic, and it is this endeavor to subsume the ontological with the ontic that collapses Dennett's project.

Epistemic Metaphysics: The Addition of Epistemology

"Epistemic Ontology" versus "Epistemic Onticism"

Again, Dennett (1992) presumes that as long as the human mind can be *understood* to function as does a computer, one can understand consciousness. Let us revisit his claim on brains and information processing:

> *[B]y thinking of our brains as information-processing systems, we can gradually dispel the fog [that is the ambiguous gap between subjective interpretations and scientific knowledge of the brain] and pick our way across [this] divide* (Dennett, 1992, p. 433).

Upon accepting this claim, one enters the epistemological domain. So, by adopting Dennett's proposed way of thinking as justified belief or knowledge, one turns the claim into an *epistemic mode*. As such, the modifier "epistemic" in the terms "epistemic ontology" and "epistemic onticism" (which I will use henceforth) refers to one's way of knowing and understanding with respect to ontology and onticism, respectively.

One could thus contend that Dennett's rejection of qualia is appropriate since the matter

of qualia is an *epistemically ontological* concern, which is what Dennett precisely wishes to avoid. Admitting that qualia are "real" and thereby exist would contribute to an epistemically ontological stance on consciousness in that qualia *are* inherently some part of the conscious experience, even though the exact details may be elusive. So, to acknowledge qualia would be to acknowledge that there is an essence or a kind of *Being* to consciousness. An *epistemically ontic* standpoint, however, would permit Dennett's rejection of qualia since a focus on essence or a kind of *Being* is irrelevant and would solely magnify the already perplexing character of metaphysical considerations.

Moreover, Dennett is not merely pursuing other possible approaches for the sake of its pursuit in the name of open-minded analysis, but is pursuing an *epistemically ontic pursuit* of consciousness. He argues that a certain epistemic mode – knowing how the mind functions through analogous thinking facilitated by imagination – offers an effective view of consciousness, and that this epistemic mode can be attained upon embracing a new set of metaphors that is not bound to Cartesianism. So, this particular epistemic mode aligns with his functionalist account; it centers on what consciousness *does* in terms of how it works. Thus, Dennett's project still stands firm, withstanding the objection that it is inconsistent with its stance on the merits of possibility with regard to imagining consciousness. This objection is thereby rendered specious.

Dennett exits the Cartesian mind-body dichotomy and enforces a functionalist perspective on consciousness, eschewing metaphysical discussions where the idea of even imagining consciousness hits a wall. By doing so, he steps away from the obscure Nagelian premise of what it means to ontologically be and targets the core of the discussion on consciousness: *how* does one conceive of consciousness, or at the very least, what framework would allow one to grasp the most accurate depiction of consciousness possible? Dennett strives to encourage philosophers to be open to different models of consciousness and not dwell only on a dualistic understanding. Furthermore, he does not merely underscore the appeal of possibility over impossibility; he emphasizes the importance of epistemic onticism over epistemic ontology which will then will open the door to alternative models when it comes to imagining consciousness. To Dennett, a computer is a more effective metaphor than a theater, and perhaps it could also be said that this choice of metaphor reflects generational advances in technology and, thus, advances in methods of conceptualization and comparison. Daniel Dennett thus conflates epistemic ontology and epistemic onticism.

Conflation of Epistemic Ontology and Epistemic Onticism

But, this is precisely the issue: a computer is a *metaphor*. A metaphor is an *approximate* means of *attempting* to understand and know a phenomenon. A metaphor does not equate but *compares* to convey a specific epistemic mode. As such, Dennett's conflation of epistemic ontology and epistemic onticism, which also appears in other literature, is rather concerning. Heidegger's ontic-ontological distinction is deliberate: both onticism

and ontology constitute existence together; they do so in the form of a duality, not a unity. When the two are taken as a unity, the ontic inevitably subsumes the ontological so that "function" becomes "nature." The net result is an epistemically ontic view of consciousness that buries the ontological.

Humans fall into this misconception at the following instance:

> *The peculiar and self-evident 'in-itself' ['An-sich'] of the nearest 'things' is encountered when we take care of things, using them, but not paying specific attention to them, while bumping into things that are unusable. Something is unusable. This means that the constitutive reference of the in-order-to to a what-for has been disrupted. […] It does not yet become explicit as an ontological structure, but ontically for our circumspection which gets annoyed by the damaged tool* (Heidegger, 2010, p. 74).

Humans commit such a conflation in their encounters with other beings and entities, including themselves, that they consider instrumenta, or tools for their manipulation and use. (This is an epistemic mode that Heidegger further addresses in his essay "The Question Concerning Technology" (1977).) Moreover, since this is a conflation of ontology and onticism, we have a conflation of the essence of existence and utility. The implication of such a conflation is the dehumanization of human existence.

This takes the human mind as merely a receptacle for "data processing" based on an input-process-output (IPO) model. If computation is taken as the computation of an input-output function as Dietrich (1990) proposes (Scheutz, 2002), **t**his model places the human mind and computer software on a level playing field.

Yet, there is more to the human mind; it cannot simply be narrowed down to the IPO model and be viewed as analogous to computer software. While the mind-body problem has certainly perplexed many philosophers over the centuries, its analysis is not for naught. The mind-body problem can be a virtue in philosophy of mind and computer science when one considers the various schools of thought that it has engendered. A functionalist *modus operandi* is *one* approach, an approach that restrains the human mind to solely its computational faculties. This is such a narrow consideration. Other schools of thought include facets of existence beyond these faculties and beyond the mind, namely the body. "Naturalistic dualism," for instance, hopes to reconcile dualism and materialism, proposing that sensory perceptions do explain physical states, but fail to allow access to consciousness (Chalmers, 1997, p. 170). Or, the body could simply be an extension of the mind as claimed by Gottfried Wilhelm Leibniz, a monist who believed that the body belongs to the mind in a non-causal link (Look, 2002). Ultimately, myriad responses to the mind-body problem recognize the extent of the human mind in a broader context, a context that is not just limited to the epistemically ontic. Nonetheless,

epistemic onticism has subsumed epistemic ontology for decades of AI and recent AGI research.

The Conflation Persists

We can trace this conflation through the years, beginning with the positional clash between computationalists Allen Newell and Herbert Simon and non-computationalists Herbert Dreyfus and Joseph Weizenbaum.

Allen Newell and Herbert Simon

Newell and Simon turn the input-processing-output model that is integral to computer programming into a metaphysical project for the human mind:

> *The [computer] program, the organization of symbol-manipulating processes, is what determines the transformation of input into output. […] The output of the processes, the behavior of* Homo cogitans, *should reveal how the information processing [performed by humans] is organized […] There is a growing body of evidence that the elementary information processes used by the human brain in thinking are highly similar to a subset of the elementary information processes that are incorporated in the instruction codes of present-day computers* (Newell & Simon, 1971, p. 281-282).

Newell and Simon continue to present this enamored view of human perception as grounded in information processing:

> *But there is a deeper beauty in the basic information processes and their organization into simple schemes of heuristic search that make that intricate human thinking possible. It is a sense of this latter beauty – the beauty of simplicity – that we have tried to convey to you* (Newell & Simon, 1971, p. 159).

They are embracing this particular epistemic mode as not only "universal scientific law" (Dreyfus, 1965, p. 51), but also ontological law. This confusion that can be further investigated by inspecting their methodology. If we inspect their methodology, we see that this view itself is merely their own perception deduced from ostensible similarities between the computational facet of human intelligence and computer code, not a universal law. In other words, they expand their own epistemically ontic view so that it becomes the epistemic ontology of the mind.

Hubert Dreyfus and Joseph Weizenbaum

Dreyfus criticizes the methodology that Newell and Simon employ in their study of information processes:

> *The empirical justification of the associationist assumption poses a question of scientific methodology – the problem of the evaluation of evidence. Cross similarities of behavior between computer and people do not justify the associationist assumption…Newell, Shaw, and Simon conscientiously note the similarities and differences between human protocols and machine traces recorded during the solution of the same problem* (Dreyfus, 1965, p. 50).

Dreyfus is astonished by how Newell and Simon (1964) assess their evidence and draw conclusions from such an assessment. In response to their claim that there is increasing evidence corroborating the similarity between information processing by humans and information processing by computer code, Dreyfus voices incredulity:

> *What is this growing body of evidence? Have the gaps in the protocols been filled and the exceptions explained? Not at all. […] What is unusual and inadmissible is that, in this case, the hypothesis produces the evidence by which it is later confirmed. Thus, no empirical evidence exists for the associationist assumption* (Dreyfus, 1965, p. 54).

Newell and Simon jump too quickly to conclusions that, first, derives an understanding of human thought merely based on correlation, and, second, uses the initial hypothesis as their evidence. Joseph Weizenbaum is unconvinced as well:

> *Whether or not this program can be realized depends on whether man really is merely a species of the genus 'information-processing system' or whether he is more than that. I shall argue that an entirely too simplistic notion of intelligence has dominated both popular and scientific thought, and that this notion is, in part, responsible for permitting artificial intelligence's perverse grand fantasy to grow. […] Man is not a machine"* (Weizenbaum, 1976, p. 203).

Humans are not machines. Humans are idiosyncratic, emotional, spontaneous, and unquantifiable. To claim otherwise is to, as we have observed with Dennett, making the error of taking a metaphor literally, yet computationalists seem enthused about making the error regardless.

Through the Decades

Epistemic onticism vis-à-vis functionalism, indeed, is philosophically effective since it eschews murky discussions of Cartesian dualism, opting for a functionalist view of human consciousness. However, this view is reductive and exclusive. An epistemically ontic approach to phenomenal consciousness reduces and limits human consciousness to the algorithmic tasks it can perform or, in other words, to the computational facet of intelligence. John Haugeland (2002) expresses this approach as such:

> *It be used to be said – perhaps reassuringly, perhaps defensively – that the aims of artificial intelligence are limited in a certain way. The goal, it was said, is not to construct a machine (or "system") capable of the full gamut of human experience or of the human mind, but rather only a system capable of human-like* intelligence *and hence cognition (so far as it is required for intelligence* (Scheutz, 2002, p. 174).

Thus, this is not a matter of consciousness, but of one component of intelligence. For a significant number of scholars in addition to Daniel Dennett within the field of artificial intelligence, however, such is not the case.

Douglas Hofstadter (1981) responds to Searle (1980) against the feasibility of AGI with the following:

> *Minds exist in brains and may come to exist in programmed machines. If and when such machines come about, their causal powers will derive not from the substances they are made of, but from their design and the programs that run in them* (Hofstadter & Dennett, 1981, p. 382).

Hofstadter sees no difference between the minds in human brains and the minds that may one day occupy programmed machines. Marvin Minsky (1985) also makes this conflation, convinced that there is nothing puzzling about the mind-body problem. He maintains that the mind is a function of the brain. He is thereby one of the functionalists who

> *view mind, the software, running on brain, the hardware. It could, of course, run on some other hardware, say computers. Hence, researchers in artificial intelligence are apt to be functionalists* (Franklin, 1995, p. 25).

Minsky further claims that each human being's brain constitutes vast collections of computers that have evolved over centuries (Franklin 1995), essentially inserting a tool that humans that have devised into the metaphysical conception of the human mind and thereby confusing the ontic with the ontological.

McDermott (2001), whom we have already encountered, adds a contentious layer to Minsky's argument with the assertion that feeling has no connection to being alive and that most living beings never experience feeling. He disposes of feeling, emotion, and sentiment in order to set the stage for his hypothesis that cerebral neurons are *functional* in that they perform computation, echoing Minsky's argument:

> *[i]f you could encode the information in another set of physical properties, and still do the same computation, you could replace all the neurons in a brain with one or more digital computers simulating them, and the brain would be just as good* (McDermott, 2001, p. 37).

Aaron Sloman (2002) briefly takes us out of the mind and goes further, equating computers to biological organisms in their entirety, convinced that the former is "just like" the latter:

> *[C]omputers were and are, above all, engines for acquiring, storing, analyzing, transforming, and using information, partly to control physical machines and partly to control their own information processing. In this they are just like biological organisms – and that includes using the information both to control complex physical and chemical systems and also to control internal processing within the controller, including processing in virtual machines* (Scheutz, 2002, p. 125).

In this regard, Sloman expands the epistemic ontic to constitute and ultimately subsume the epistemic ontological.

Steven Pinker (2005) appears to show less certainty in a computationalist position than those we have examined so far with his acknowledgement that current knowledge of how the mind works is inadequate and not yet fully understood; yet he nonetheless contends that the mind naturally a computational collection of naturally selected organs. This acknowledgement still does not lessen the degree of his conviction.

Ben Goertzel (2007) then directs us toward AGI. Goertzel envisions that in the near future, brain-scanning technologies and computer hardware, together, will be so advanced that they will successfully emulate the human brain and form an uploaded AGI. Embedded in this vision is the view of the human mind as "software", a software that can eventually be installed onto computer hardware.

Bernard J. Baars and Stan Franklin (2009) argue that progress in the implementation of machine consciousness is dependent upon the understanding of human consciousness. Moreover, they assert that there is a substantial amount of evidence that suggests that consciousness in humans and other mammals is driven by a "global access function" (p. 24). Here, machine *intelligence* is expressed as machine consciousness, and a function is deemed sufficient to capture the essence of phenomenal consciousness. Baars and Franklin conflate both intelligence and consciousness as well as epistemic onticism and epistemic ontology à la functionalism.

Roman V. Yampolskiy (2014) does not succumb to the conflation of intelligence and consciousness by focusing on mind as intelligence, yet he still conflates epistemic onticism and epistemic ontology. He concludes that a mind is an intelligence that gathers knowledge about its environment upon deeming all minds equal to software:

> *If we accept [functional] materialism, we have to also accept that accurate software simulations of animal and human minds are possible. […] Consequently, we can treat the space of all minds as the space of*

> *programs with the specific property of exhibiting intelligence if properly embodied* (Yampolskiy, 2014, pp. 1-2).

Yampolskiy seems more functionalist than materialist here (and, furthermore, the two are not interchangeable); as such, I will qualify his usage of "materialism" with "functional materialism", or the view that the mind is a function of the brain (Bjerregaard, 1914, p. 268), to make his argument more pertinent. His view is no different from that of Minsky in this regard.

Similar to McDermott, Joscha Bach (2017) recognizes that there is little proposed to address the hard problem of consciousness. He attempts to delineate a computational model that can explain "the phenomenology and functionality of consciousness", an approach he calls the "conductor theory of consciousness (CTC)" (p. 2). According to his approach,

> *cortical structures are the result of reward driven learning, based on the signals of the motivational system, and the structure of the data that is being learned. The conductor is a computational structure that is trained to regulate the activity of other cortical functionality. […] CTC explains different conscious states by different functionality bound into the self construct provided by the attentional protocol* (Bach, 2017, pp. 6-7).

Yet again, we see the conflation in Bach's work as well, especially in the framing of Bach's thesis which he understands to be addressing the "functionality of consciousness."

Marcel van Gerven (2017) introduces connectionism into our examination. He also makes the misguided conflation of epistemic onticism and epistemic ontology with regard to phenomenal consciousness, for, as I will explain in the next section, connectionism is fundamentally an extended application of computationalism:

> *Connectionism came to be equated with the use of artificial neural networks that abstract away from the details of biological neural networks. An artificial neural network (ANN) is a computational model which is loosely inspired by the human brain as it consists of an interconnected network of simple processing units (artificial neurons) that learns from experience by modifying its connections. Alan Turing was one of the first to propose the construction of computing machinery out of trainable networks consisting of neuron-like elements* (Gerven, 2017, p. 6).

Connectionism: An Extension of Computationalism

According to connectionism, intelligence and mind somehow emerge from highly interconnected groups of simple units such as the artificial neural networks (ANN) that

Gerven (2017) describes (Franklin 1995). This view extends to consciousness as well; in the eyes of a connectionist, consciousness could also be understood as emergent from sufficiently sophisticated information processing (Lloyd 1995). The next logical consideration could be an investigation of the concept of emergence and whether or not there truly exist emergent properties. However, I see that connectionism is, at its very root, an evolved extension of computationalism and, therefore, the conflation of epistemic ontology and epistemic onticism, so I will not pursue said investigation.

The criticism that Dreyfus (1965) offers in response to Newell and Simon (1964) is still relevant here. Connectionism has, at its core, at least two assumptions:

1. **All neurons are alike.** Artificial neurons behave like cerebral neurons in humans.

2. **Emergent consciousness.** Consciousness emerges from networks of these artificial neurons.

Laden in these assumptions is the belief that computational models such as artificial neural networks accurately capture phenomenal consciousness, once again equating the functionality of such models to the entire essence of the human mind.

Furthermore, these assumptions are used to conclude that the connectionist theory of consciousness holds. What then follows is a methodological perversion of empirical study. Connectionists such as Lloyd use their conclusion to prove their initial assumptions. If Newell and Simon use their initial hypothesis to prove their conclusion, then Lloyd uses his conclusion to prove the assumptions that ground his initial hypothesis. In both cases, the methodology employed is illogical.

Current and future theories of consciousness will continue to be flawed and misguided provided that the conflation of epistemic ontology and epistemic onticism and erroneous implementations of empirical methodology remain pillars in AGI research.

Forecasting the Ethics of Human-Computer Interaction

I have previously argued that the conflation of epistemic ontology and epistemic onticism results in dehumanization. Alexander Barzel expresses a similar concern:

> *There is an urgent need to assign the borderline [between humans and computers] by pointing to differences, beyond which the effort to compare the traits of human beings and those of the computer misrepresents both and is dangerous; the reduction of organic human thinking to the computer's mechanism can end up in humankind's dehumanization* (Barzel, 1998, p. 166).

Ethics is a *human* matter. Moreover, it is a matter of establishing a form of respect between living entities. When artificial entities begin to dissolve the meaning of human

existence, ethics become irrelevant. Humans are not merely *functions*. At the same time, equating AGI's potential to the human mind produces a phenomenon in applied ethics known as "Playing God" (Chadwick & Schroeder, 2002, p. 44). Hubris indeed comes with the prospect of creating an entity that could purportedly be conscious as humans are and could lead to self-apotheosis.

How is it that the metaphysical and epistemological commitments of AGI research both deprecate humans by dehumanizing them and elevate humans by turning them into gods? Heidegger (1977) and Mary Shelley (2015), together, speak to these behaviors.

Heidegger's *Gestell* and Mary Shelley's *Frankenstein*

Heidegger frames technology in such a way that its essence is fundamentally *human* rather than technological in itself. Technology is a subset of human existence that is intrinsically tied to human perception and attitude, not an extension of human existence that replicates a kind of humanness which is the purported aim of AGI research. Heidegger's correlation of the essence of technology with a human attitude rather than technological devices themselves emphasizes that technology is a tool that *humans* have created for their use. Technology's origins are thereby *human*. This conception expands the metaphysical consideration of technology into a metaphysical consideration layered with epistemology à la human perception.

Heidegger focuses on the danger of "Enframing" or *Gestell*, which he understands as a specific epistemic mode. Historically, humans have perceived entities in their reality, including themselves, as "standing reserve", or mere tools that are always available and "standing" for humans' use and control (Heidegger, 1977, p. 17). As long as humans continue to engross themselves in the notions of utility, functionality, and control as well as interpret their world in terms of those notions *without questioning the underlying drives for such behavior*, they will be unable to liberate themselves from their reductive perceptual frameworks. This is the danger of *Gestell*, a danger that is directly applicable to the current assumptions in AGI research that I have explained throughout this paper. Humans may believe that they are the masters of the studies leading up to the creation of an AGI, but they lose their control when they choose to be enslaved by their own self-reduction and self-deprecation in the process, a dynamic that will indubitably dictate future human-computer interaction. This is rather undignifying.

Gestell can be further investigated through a reading of Shelley's *Frankenstein*, specifically in terms of Asimov's "Frankenstein Complex" (King, 2017, p. 5). Asimov recognizes the weight of Shelley's narrative and its predictive power. Literature speaks to and reveals humans' most innate desires and propensities, and *Frankenstein* is no exception. Shelley tells of Dr. Victor Frankenstein, a man who fails to anticipate the consequences of his creation, drunk on the power of "Playing God" as he creates a Creature in his image, an entity which he can exercise authority over as its master. The

same consequences may result in AGI research if current assumptions continue and are not corrected or, at the very least, acknowledged. Once the conflation of epistemic ontology and epistemic onticism is recognized, such hubris can be abated. Frankenstein urges the reader to

> *[l]earn from me, if not by my precepts, at least by my example, how dangerous is the acquirement of knowledge, and how much happier that man is who believes his native town to be the world, than he who aspires to become greater than his nature will allow. […] No one can conceive the variety of feelings which bore me onwards, like a hurricane, in the first enthusiasm of success. Life and death appeared to me ideal bounds, which I should first break through, and pour a torrent of light into our dark world. A new species would bless me as its creator and source…* (Shelley, 2015, p. 37).

Frankenstein, intoxicated not only by the creative endeavor, but also by the prospect of becoming "God" or a master in the Heideggerean sense, does not question the nature of his ambitions and their potential repercussions. As a result, he creates an entity that he does not anticipate and enters a state of deep remorse and horror. Frankenstein's plight has two elements:

1. **"Playing God" enslaves and blinds**. We return to the danger of *Gestell*. The desire for authority and control as a creator not only lowers the status of the creation, but also enslaves the creator with its allure. Both the creator and the creation are thereby held captive by the notion of power.

2. **Blinded action reveals itself in a negative aftermath.** The danger that Frankenstein describes should not be interpreted the knowledge that comes with successfully emulating human consciousness, but the knowledge of the consequences that result from creating an AGI, consequences that cannot be predicted. If AI researchers possess, as does Frankenstein, a fascination with creation as well as the desire to, someday, exercise authority as the "creator" of AGI, the danger of such a mindset will not reveal itself unless there occurs a tragedy.

Hence, I introduce the final implication of the conflation of epistemic ontology and epistemic onticism: if my argument truly holds, then there is nothing remotely human about the AGI envisioned by past and current researchers. That being said, one cannot foresee with confidence how AGI may evolve. To continue to indulge in the hubristic idea of creating consciousness "in humans' own image" would only be a detriment to AGI research; researchers will not be prepared to handle their own Creature.

Conclusion

Technology is present in everyday human experience and significantly influences human perceptions of reality, not only as tangible entities, but also as a particular mindset. As long as humans continue to perceive existence as a matter of utility and exploitation, then their existence is reduced to an enslaved instrumentum. Such a perception insults the aesthetics of life that humans experience through love, intimacy, connection, and art. An adoring gaze, a compassionate gesture, a warm embrace, and an emotional immersion into a musical or artistic experience become meaningless in the reductive mechanization of *Gestell*. Humans must decide the value of their existence and which future they wish to live, and technology, more so now than ever, puts this existential task to the ultimate test.

Thus, this discussion is not solely about artificial intelligence or artificial general intelligence. The root of this discussion is about *humans* and the *epistemic mode* with which humans are envisioning technology such as artificial general intelligence. In my exploration of the metaethical terrain of AI and AGI research, I hope to have elucidated the metaphysical and epistemological commitments that AI and AGI research have perpetuated so that when discourse on ethics arises, the foundations of those discussions are clear and evident.

Initiatives such as the beneficial AI movement are certainly working toward ensuring that AI and AGI development aligns with human goals and interests (Tegmark, 2017). Yet, it is my wish that they also recognize that AGI may, one day, no longer be "technology" or tools that humans employ, but a force that humans cannot keep under their jurisdiction. If that day comes, humans will face the same irrevocable devastation that Dr. Frankenstein suffered.

References

Baars, B. J., & Franklin, S. (2009). Consciousness is Computational: The LIDA Model of Global Workspace Theory. *International Journal of Machine Consciousness, 1*(1), 23-32. https://doi.org/10.1142/S1793843009000050

Bach, J. (2017). The Cortical Conductor Theory: Towards Addressing Consciousness in AI Models, IJCAI-17 Workshop on Architectures for Generality & Autonomy, Melbourne, Australia, August 19, 2017. Melbourne: International Joint Conference on Artificial Intelligence.

Barcel, A. (1998). The Perplexing Conclusion: The Essential Difference between Natural and Artificial Intelligence is Human Beings' Ability to Deceive. *Journal of Applied Philosophy, 15*(2), 165-178. https://doi.org/10.1111/1468-5930.00084

Bjerregaard, C. H. A. (1914). The Eight Methods and the Four Systems of Thinking. *The Word, 18*, 263-270.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy, 4*(1), 15-31. https://doi.org/10.1111/1758-5899.12002

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In K. Frankish & W. Ramsey (Ed.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge: Cambridge University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., …Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* Retrieved from Cornell University Library Website: https://arxiv.org/pdf/1802.07228.pdf

Chadwick, R. F. (2002). Playing God. In R. Chadwick & D. Schroeder (Ed.), *Applied Ethics: Critical Concepts in Philosophy* (Vol. 2). Abingdon: Taylor and Francis.

Chalmers, D. (1997). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, D. (2010). *The Character of Consciousness*. Oxford: Oxford University Press.

Dennett, D. (1992). Consciousness Imagined. In *Consciousness Explained.* (pp. 431-455). New York: Bay Back Books.

Dietrich, E. (1990). Computationalism. *Social Epistemology, 4*(2), 135-154. https://doi.org/ 10.1080/02691729008578566

Dreyfus, H. (1965). *Alchemy and Artificial Intelligence*. Santa Monica, CA: RAND Corporation.

Fodor, J. A. (1981). The Mind–body Problem. ***Scientific American, 244***(1), 114-123. http://dx.doi.org/10.1038/scientificamerican0181-114

Franklin, S. (1995). *Artificial Minds*. Cambridge, MA: MIT Press.

van Gerven, M. (2017). Computational Foundations of Natural Intelligence. *Frontiers in Computational Neuroscience, 11*. https://doi.org/10.3389/fncom.2017.00112

Goertzl, B. (2007). Human-level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's Critique of Kurzweil. *Artificial Intelligence, 171*(18), 1161-1173. https://doi.org/10.1016/j.artint.2007.10.011

Heidegger, M. (1977). *The Question Concerning Technology and Other Essays* (W. Lovitt, Trans.). New York: Harper & Row.

Heidegger, M. (2010). *Being and Time* (J. Stambaugh, Trans.). Albany, NY: State University of New York Press.

Hofstadter, D., & Dennett, D. (Eds.). (1981). *The Mind's I: Fantasies and Reflections on Self and Soul*. New York, NY: Basic Books.

King, B. (Ed.). (2017). *Frankenstein's Legacy: Four Conversations about Artificial Intelligence, Machine Learning, and the Modern World*. Pittsburgh, PA: Carnegie Mellon University ETC Press.

Lloyd, D. (1995). Consciousness: A Connectionist Manifesto. *Minds and Machines, 5*(2), 161-185. https://doi.org/10.1007/BF00974742

Look, B. (2002). On Monadic Domination in Leibniz's Metaphysics. *British Journal for the History of Philosophy, 10*(3), 379-399. https://doi.org/10.1080/09608780210143209

McDermott, D. (2001). *Mind and Mechanism* (1st ed., MIT Press). Cambridge, MA: A Bradford Book.

McDermott, D. (2007). Artificial Intelligence and Consciousness. Retrieved from Department of Computer Science at Yale University Website: http://www.cs.yale.edu/homes/dvm/papers/conscioushb.pdf

Minsky, M. (1985). *The Society of Mind*. New York, NY: Simon and Schuster.

Nagel, T. (1974). What Is It Like to Be a Bat. *The Philosophical Review, 83*(4), 435-450. https://doi.org/10.2307/2183914

Newell, A., & Simon, H. (1964). Information Processing in Computer and Man. *American Scientist,52*(3), 281-300.

Newell, A., & Simon, H. (1971). Human Problem Solving: The State of the Theory in 1970. *American Psychologist, 26*(2), 145-159. http://dx.doi.org/10.1037/h0030806

Pinker, S. (2005). So How Does the Mind Work? *Mind & Language, 20*(1), 1-24. https://doi.org/10.1111/j.0268-1064.2005.00274.x

Putnam, H. (1992). *Renewing Philosophy*. Cambridge, MA: Harvard University Press.

Scheutz, M. (Ed.). (2002). *Computationalism: New Directions*. Cambridge, MA: A Bradford Book.

Searle, J. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences, 3*(3), 417-457. https://doi.org/10.1017/S0140525X00005756

Shelley, M. (2015). *Frankenstein: or the Modern Prometheus*. Mogul Classics.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, NY: Alfred A. Knopf.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), 433-460.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. New York, NY: W.H Freeman and Company.

Yampolskiy, R.V. (2014). The Universe of Minds. Retrieved from Computing Research Repository Website: https://arxiv.org/pdf/1410.0369.pdf

Funding: This research received no funding.

Acknowledgements: The author would like to acknowledge the research guidance and continuous support provided by Provost of Smith College and Olin Professor of Computer Science Joseph O'Rourke, Roe/Straut Professor of Philosophy Susan Levin, Doris Silbert Professor of Philosophy Jay L. Garfield, the six years of life mentorship provided by constitutional law scholar Dr. Robert Wilson, and all project activities to the ideas that underpin this paper. I would like to thank the organizers and participants of the ETHICOMP conference in computer ethics held at De Montfort University in September 2018.