THE AI ENSOULMENT HYPOTHESIS

Brian Cutter

(Forthcoming in *Faith and Philosophy*)

According to the AI ensoulment hypothesis, some future AI systems will be endowed with immaterial souls. I argue that we should have at least a middling credence in the AI ensoulment hypothesis, conditional on our eventual creation of AGI and the truth of substance dualism in the human case. I offer two arguments. The first relies on an analogy between aliens and AI. The second rests on the conjecture that ensoulment occurs whenever a physical system is "fit to possess" a soul, where very roughly this amounts to being physically structured in such a way that the system can meaningfully cooperate with the operations of the soul.

1. Introduction

Suppose that substance dualism is true for human beings. Associated with each living human body is a soul—an immaterial entity that is somehow the basis or fundamental bearer of certain of our mental capacities, such as the capacity for consciousness, thought, and rational agency. Under this assumption, should we expect future AI systems to possess souls as well? Let's call the hypothesis that some future AI systems will have souls *the AI ensoulment hypothesis*. I shall argue that we should take the AI ensoulment hypothesis seriously. If we have souls, we shouldn't be confident that future computers won't.

To simplify the question, let's set aside difficulties about forecasting the behavioral and functional capacities of future AI systems. We'll assume that eventually we will create an "artificial general intelligence" (AGI), an AI that meets all behavioral or functional criteria for human-level (or greater) general intelligence. (By stipulation, the concept of AGI used here is a

purely behavioral/functional concept, so it's not part of the concept of an AGI that it is phenomenally conscious or has any genuine mental life at all.) We'll assume that these machines are able to perform more-or-less any behavioral task as well as or better than a typical human, and we'll also assume their internal functional organization relevantly resembles that of a human brain, though perhaps only very abstractly. I am concerned with whether AGIs will have souls if we ever manage to create them. I'll leave it to others to forecast whether and when we will actually do so. My thesis is that we should have at least a middling credence in the AI ensoulment hypothesis, conditional upon our eventual creation of AGI and the truth of substance dualism in the human case. ("At least a middling credence" means a level of confidence that isn't very low. It's somewhat artificial to give precise numbers, but I would stand by a precisification of ≥ 0.25 or so.)

I offer two related arguments for this thesis, the *alien-analogy argument*, and the *fitting-recipient argument*. The basic idea of the alien-analogy argument is simple. If we met seemingly intelligent aliens that resemble us abstractly in behavioral and functional respects but differ from us in material composition, it would be unreasonable to dismiss the hypothesis that they have souls, provided we do. And if we shouldn't dismiss the alien ensoulment hypothesis, neither should we dismiss the AI ensoulment hypothesis, for the cases are closely analogous. The fitting-recipient argument rests on the idea that, if substance dualism is true, it is a reasonable conjecture that ensoulment occurs whenever a physical system is "fit to possess" a soul, where very roughly this amounts to being physically structured in such a way that the system can meaningfully cooperate with the operations of the soul. It is then argued that being fit to possess a soul depends only on a system's behavioral capacities and functional organization. Since the human body is fit to possess a soul, a functionally human-like AI would be as well.

Since the central thesis is conditional on substance dualism and our eventual creation of AGI, this thesis would be uninteresting if we could confidently dismiss either assumption. But I think it would be unreasonable to rule out either. The difficulty of finding a place for consciousness within the physical world is plausibly a reason to accept some kind of dualism, and several considerations favor substance dualism over mere property dualism. For example, there is the attractive idea that fundamental properties (like consciousness, if some form of dualism is true) should attach to fundamental entities, and a simple soul is a better candidate for a fundamental entity than a brain or body composed of trillions of particles.[1] Substance dualism also nicely accommodates our intuitions about the determinacy of personal identity over time,[2] resolves the "mental problem of the many,"[3] and may provide a better account of the unity of consciousness[4] and simplify the fundamental psychophysical laws.[5] As for the assumption that we will create AGI, this receives some support from the startling pace of progress in AI over recent decades and the predictions of AI researchers, which suggest an arrival date for AGI within the next century.[6]

---

[1] Chalmers, *The Character of Consciousness*, 139n36.

[2] Swinburne, *The Evolution of the Soul*, ch. 9; Parfit, *Reasons and Persons*, ch. 11; Huemer, *Knowledge, Reality, and Value*, ch. 12.

[3] Unger, *All the Power in the World*.

[4] Hasker, *The Emergent Self*.

[5] Collins, "A scientific case for the soul."

[6] Grace et al., "When will AI Exceed Human Performance?" is the most comprehensive survey of AI researchers on these topics. According to the average respondent, there is a 50% probability that unaided machines will be able to perform every task better and more cheaply than humans by 2061. However, survey responses were somewhat inconsistent across questions, and some responses suggest later estimates for the arrival of AGI.

A conclusion to the effect that the AI ensoulment hypothesis is flat-out true, or very likely, would perhaps be more interesting than my comparatively modest "middling credence" thesis. However, even the modest conclusion could have hugely significant normative implications. If we are moderately confident that an AI has a soul relevantly like ours, we should be moderately confident that it has significant moral status. This creates a powerful "moral risk" argument for giving normative weight to the AI's potential interests.[7] In any event, I doubt we are justified in accepting the stronger conclusions. While I will only argue that we should have at *least* a middling credence, my own view is we should also have at *most* a middling credence in AI ensoulment. I suspect the investigation below will partially justify the latter claim by revealing how many difficult and unsettled questions there are about the material conditions for ensoulment. But I'll also suggest a way in which we may one day be in a position to get empirical evidence that modestly bears on the question of AI ensoulment. We may not be fated to a permanent state of middling agnosticism.

After some preliminary clarifications (§2), I defend the two arguments in §§3–4. I conclude with a brief discussion of some practical and theoretical ramifications of the AI ensoulment hypothesis (§5).

2. Preliminaries on Souls and Ensoulment Conditions

As a preliminary matter, we need to clarify three concepts: *soul*, *ensoulment*, and *nomologically sufficient conditions for ensoulment*.

Soul: By a "soul," I mean a concrete immaterial entity that is, in some manner, the basis of certain mental capacities, such as consciousness, thought, or rational agency, and which is

[7] Compare moral risk arguments for the wrongness of abortion if there is at least a moderate chance the fetus is a person (Beckwith, *Defending Life*, 150-2).

united (in a sense clarified below) to a physical system. In calling the soul "immaterial," I mean, roughly, that it isn't made of physical stuff. It isn't composed of atoms, molecules, or anything else dealt with in microphysics. I say the soul is "the basis of" certain mental capacities in order to maintain neutrality about the exact relationship between our souls and our mental properties, including whether the soul is the fundamental bearer of properties like thinking, feeling, and choosing. On one view, it is the soul that thinks, feels, and chooses. On another view, it is something else that thinks, feels, and chooses—perhaps a substance composed of body and soul—but this "something else" does so at least partly in virtue of the operations of the soul (much as a person breathes in virtue of the activity of the lungs, though it would be incorrect to say that the lungs breathe). Officially I take no stand on this issue, though for brevity I will occasionally attribute mental properties to the soul. (In each case it should be clear how such attributions could be rephrased, more longwindedly, in metaphysically neutral terms.) Nor do I take a stand on the related question of whether you are your soul or your soul is just a proper part of you, alongside your body.

The latter question is sometimes thought to be a central point of disagreement between a broadly Cartesian and a broadly Aristotelian conception of the soul. There are other points of disagreement about which I shall also remain neutral, including: whether the soul is responsible for non-mental vital functions like nutrition, growth, and metabolism, whether non-human animals have souls, and whether souls are substances in any sense of "substance" that means more than "concrete particular." Relatedly, for the purposes of this paper, "substance dualism" should be taken to imply that the soul is a non-physical substance in the sense of being a concrete

particular distinct from any purely physical thing, event, or process, but does not imply that the soul is a substance in any more loaded sense of "substance."[8]

Ensoulment: It is standardly assumed that the union of the human body and soul consists at least partly, and perhaps entirely, in their direct causal relations to one another. First, we have "bottom-up" (body-to-soul) causal influence. For example, my sensory experiences are caused by states of my body, such as states of the sensory systems in my brain. While physical objects outside my body can influence my soul (my laptop is among the causes of my current sensory experience, for example) they do so only via their influence on my body. Second, many substance dualists also accept "top-down" (soul-to-body) causal influence. For example, my intention to raise my hand is causally responsible for my hand rising. While my soul might causally influence things outside my body (my pen, for example), it does so only via its influence on my body. The interplay of bottom-up and top-down influences might often be complex and subtle, with each side making minute adjustments in a seamless give-and-take, a dance with no lead. I'll assume that a physical system (like a physical computer) is "ensouled" (or "has a soul," or is "united to a soul") only if there are direct causal connections of this kind between the physical system and a soul, such that the states or activities of the physical system directly and systematically affect the soul, or *vice versa* (inclusive "or"). Much more could be said about the notion of ensoulment, but these minimal remarks will suffice for our purposes.[9]

---

[8] See Feser, "Aquinas on the Human Soul" and Moreland, "In Defense of a Thomistic-Like Dualism" for recent discussion and defense of a broadly Aristotelian conception of the soul. See Swinburne, "Cartesian Substance Dualism" for a defense of a broadly Cartesian account.

[9] For more detailed accounts of ensoulment along these lines, see Foster, *The Immaterial Self*, 261-6, and Swinburne, *The Evolution of the Soul*, ch. 8.

Nomologically sufficient conditions for ensoulment: Ensoulment presumably isn't a random or haphazard affair. If substance dualism is true, we can expect there to be reliable, counterfactual-supporting regularities to the effect that, if a physical system meets condition C, a soul is united to it. For one potential example, let C = the physical profile of a human embryo (or later-stage human fetus if ensoulment occurs long after conception). I will say that a condition C that figures in such a regularity is a *nomologically sufficient condition* for ensoulment. The AI ensoulment hypothesis roughly amounts to the hypothesis that the matter constituting some future AI systems will satisfy some nomologically sufficient condition for ensoulment. For a functionally human-like AI, this would be the case if something like the following conjecture (here stated vaguely) is correct:

> *Functional sufficiency*: Having a human-like functional organization (i.e., a functional organization that suitably resembles that of a typical human body/brain) is a nomologically sufficient condition for ensoulment.

The arguments in §3 and §4 can be read as supporting something like functional sufficiency. I leave open exactly which functional similarities to the human body/brain are relevant in order for functional sufficiency to kick in, but potential candidates may include: having a global-workspace architecture,[10] world-modeling and self-modeling capacities[11], "re-entrant" signaling

[10] Baars, *A Cognitive Theory of Consciousness.*

[11] Rosenthal, "Consciousness and Mind", Hofstadter, *I am a Strange Loop*.

in perceptual processing,[12] complex linguistic behavior, an internal system of representations that exhibits productivity and systematicity,[13] and goal-directed behavior.[14]

I use the language of "nomological" sufficiency because the relevant regularities are assumed to be law-like in their reliability and counterfactual robustness, though I remain neutral on whether they strictly qualify as laws of nature. This question may turn on the relationship between laws of nature and causal powers, as well as theological questions about the generation of souls. For example, if natural laws are descriptions of the causal powers of natural entities, then these regularities won't qualify as laws of nature on the anti-emergentist view that natural entities never cause a soul to come into existence, but only serve as the occasion for God to create and infuse a soul.[15] But even if souls are always specially created and infused by God, and not causally emergent from matter, we can still speak of "nomologically sufficient conditions for ensoulment" in the intended sense. There need only be counterfactually robust regularities such

[12] Lamme, "Toward a True Neural Stance on Consciousness."

[13] Fodor, *Psychosemantics*, appendix.

[14] A potential concern for functional sufficiency is that it seems to imply that the people of China would collectively be united to a soul in Ned Block's "China brain" scenario, in which the people of China emulate the functions of a human brain through radio communication (Block, "Troubles with functionalism"). I'm not convinced that this is an unacceptable result. But in any event, the arguments below are compatible with modified versions of functional sufficiency that avoid this result. For example, one can add a further condition that the proper parts of the system are not themselves ensouled. This would be analogous to the exclusion condition in Putnam's classic defense of functionalism, which ensures that conscious systems are not decomposable into conscious parts (Putnam, "Psychological Predicates.").

[15] For a recent defense of anti-emergentist substance dualism, see Rickabaugh, "Against Emergent Dualism." Defenders of emergentist substance dualism include Hasker, *The Emergent Self*; Zimmerman, "From Experience to Experiencer," and Popper and Eccles, *The Self and its Brain*.

that, when a physical system satisfies certain conditions, God can be counted on to specially create and infuse a soul (allowing for isolated exceptions, as with miraculous exceptions to other laws, and allowing that God always has the sovereign power to refrain from creating a soul in any given case). Again, if we think it's a safe bet that the next human embryo/fetus will receive a soul, we implicitly accept that there are such regularities.

Relatedly, the AI ensoulment hypothesis does not imply that we will create the souls of future AIs. If future machines have souls, our part will be to create the material conditions for ensoulment. This leaves open whether we would be correctly described as creator or cause of the soul. It may be helpful to consider the analogous question in the human case. Any substance dualist will say that when matter is arranged in certain ways—for example, in a human-embryo-like way—a soul is joined to that matter (at least typically). Now, we can do things—in the bedroom or the IVF lab—that cause matter to be so arranged. We bring about the material conditions for ensoulment. Some may wish to say that we thereby indirectly cause the soul to exist, but this further claim is not mandatory. One could instead accept the anti-emergentist view described above. Alan Turing, though not himself a substance dualist, recommends such a view to the theist substance dualist, writing, "In attempting to construct [ensouled] machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates."[16] Alternatively, one could say that the soul is pre-existent and the embryonic material conditions are just the cause or occasion the soul's *union* with matter, not its existence. The key point is that whatever story is correct in the human case, a corresponding story should hold for AI souls. We (human programmers and technicians) bring about certain

[16] Turing, "Computing Machinery and Intelligence," 443.

physical conditions, which either cause or occasion the coming-to-be of the AI's soul (if souls aren't pre-existent), and which either cause or occasion the soul's union with the matter of the computer.

A substance dualism that endorses functional sufficiency can be usefully compared to David Chalmers' functionalist property dualism.[17] Chalmers defends the *principle of organizational invariance*, a law of nature according to which systems with the same causal organization are alike in phenomenal respects. A corollary of this principle is that any system with the same functional organization as a typical (conscious) human being would be conscious. For Chalmers, consciousness is ontologically distinct from any functional property, but some functional properties are nomologically sufficient for consciousness. Likewise, according to the view suggested here, ensoulment—the state of being united with an immaterial soul—is ontologically distinct from any functional property of a physical system, but some functional properties may nonetheless be nomologically sufficient for ensoulment. (However, it is worth noting that functional sufficiency is weaker than the substance dualist analogue of organizational invariance, which says that any two functionally equivalent systems are alike with respect to ensoulment. I remain neutral on this stronger principle.)

3. The Alien-Analogy Argument

Imagine we meet aliens that behave in seemingly intelligent ways. They use language, build advanced machines, and create art. They have a complex civilization, with sophisticated governments and economic systems. Suppose we also learn that their internal functional organization resembles ours to some extent, despite differences in detail. They have nervous systems that take in, process, and store information in ways that loosely and abstractly resemble

[17] Chalmers, *The Conscious Mind*.

human neural processing. However, their material makeup is very different from ours. Perhaps they are made of some kind of non-carbon-based green goo. I suspect most would agree that we should not dismiss the hypothesis that they have souls, provided we do. This is the first premise of the alien-analogy argument:

> **A1.** If we met aliens that differ from us in their material constitution, but resemble us in high-level functional respects and exhibit seemingly intelligent behavior, then we should have at least a middling credence that they have souls (assuming we do).

I suspect many would accept a stronger claim—that we should have a high credence in alien ensoulment. But the weaker claim will suffice for our purposes. Since they resemble us in the kind of behavior and functioning associated with ensoulment in the human case—the behavior and functioning we take as evidence for consciousness, thought, and rational agency in other humans—there would be a moderately strong argument from analogy that the aliens resemble us with respect to ensoulment. On the other hand, their material makeup is very different from ours, which perhaps weakens the argument from analogy to some extent. We might therefore be more confident that our human neighbor has a soul than that the aliens do. But a mere difference in material substrate surely doesn't weaken the argument from analogy to a degree that would justify outright dismissal of the alien ensoulment hypothesis. After all, it's not obvious why a carbon-based substrate should be a nomological requirement on ensoulment. It *might* be, but it would be unreasonable to assume it must be.

I doubt the first premise will meet much resistance. Most of the debate will center on the second:

A2. If (A1) is true, we should also have a middling credence that a functionally human-like AI—an AI system whose functional organization abstractly resembles that of a human body/brain—would have a soul (again, assuming we do).

The motivation for (A2) is just that the alien case and the AI case seem very closely analogous, at least *prima facie*. In both cases, we have something that abstractly resembles us in mind-relevant functional/behavioral respects, but differs from us in its low-level material constitution. If low-level material differences aren't a good reason to reject alien ensoulment, they aren't a good reason to reject AI ensoulment. (Indeed, if we consider the special case of an AI that is a perfect functional isomorph of a human, such as a detailed simulation of a human brain, we may have a stronger argument from analogy for AI ensoulment than for alien ensoulment, provided the aliens only imperfectly resemble us in behavioral/functional respects.)

Those who wish to reject (A2) must point to some *relevant difference* between the AI and the alien that would justify different attitudes toward the two ensoulment hypotheses. Specifically, one would need to identify some property F such that we can reasonably be confident that (i) F is nomologically necessary for ensoulment (and thus *we* have F), (ii) the AI lacks F, and (iii) the aliens have F. Any F that fails to meet condition (i) wouldn't be a *relevant* difference, and any that fails to meet (ii) and (iii) wouldn't amount to a difference at all. These are severe constraints. To meet them, F must be a property that distinguishes humans and AIs (on pain of violating (i) or (ii)) but *not* a property that distinguishes humans and aliens (on pain of violating (i) or (iii)). So F probably can't be anything specific to do with low-level material constitution (for then it would divide humans and aliens), nor could F be anything to do with functional organization or behavioral capacities (for then it wouldn't divide humans and functionally human-like AIs).

This doesn't leave the opponent of (A2) much to work with, but there are still some potential differences worth considering. First, there may be *biological* differences: the AI and the alien might differ with respect to whether they are alive, for example. Second, there are *historical* differences: the AI and the alien arose through different sorts of processes, one artificial, the other natural, one through intentional design, the other (we may assume) through Darwinian evolution. Third, there may be *causal* differences. Perhaps physical causal closure holds for the physical computer that constitutes the AI, but not for the body of the alien.[18]

3.1 Biological Differences: Life and Underived Teleology

Perhaps the relevant difference is *life*. It might be said that (i) the alien is alive while the AI is not, and (ii) life is a nomologically necessary condition on ensoulment. I am skeptical of both parts of the life response. First, it's not obvious that future AI systems won't qualify as living things. The life response assumes that aliens are alive, so the operative conception of life must be relatively abstract and substrate-neutral. It can't be a requirement on life that it involve specific molecular processes that are unique to terrestrial biology. We can't require DNA molecules, carbon-based matter, ATP-based metabolism, or anything of the sort. The markers of life must be higher level functions, such as reproduction, metabolism, homeostasis, growth, and so forth. But future AI may exhibit many of these functions. They may reproduce by making copies of their code, perhaps with small "mutations" of certain parameters. They may have homeostatic

[18] The alien-analogy argument has some obvious ties to multiple-realizability arguments against type-identity theory (Putnam, "Psychological Predicates"), which often invoke aliens who partially resemble humans in mental respects while differing greatly in their material constitution. Where traditional multiple-realizability arguments suggest that having a human-like physical-chemical makeup is not necessary for (say) feeling pain, the argument here suggests that having a human-like physical-chemical makeup may not be (nomologically) necessary for ensoulment.

control systems, as many current AI systems do. They may exhibit a kind of metabolism, taking in and using energy from their environment. They may exhibit a kind of growth, as they increase memory and processing capacity by expanding onto new hardware.[19] Of course the similarities to organic life will be fairly abstract. Below a certain level of description, the processes look quite different. But the same may be true of the relationship between human and alien biology.

There are different views about how these high-level functions are related to life. A reductive view would say that these functions *constitute* life. To be alive *just is* to exhibit sufficiently many of the relevant functions to sufficiently high degree. A more metaphysically inflationary view might say that these functions are mere symptoms of life. Perhaps life is a primitive property, or perhaps it consists in the possession of a non-physical *élan vital*, or perhaps it involves the possession of a soul. On either the reductive or non-reductive view, we can't dismiss the idea that future AIs will be alive. We've seen that future AIs may exhibit many of the relevant functions, so if the non-reductive view is correct, our functionally human-like AI exhibits many of the characteristic symptoms of life. In that case, it would be unreasonable to be highly confident that it's not alive. If the reductive view is correct, then it may well qualify as alive, since it exhibits many of the functions that constitute life. Or perhaps, if a reductive view is correct, it is simply a verbal question whether future AI will be alive, much as it is arguably verbal question whether a virus is alive. It will turn on more-or-less arbitrary choices about thresholds and precisifications. Exactly which functions go on the life-constituting list, and exactly how many need to be satisfied? Is it a majority, or a weighted majority, and with what weights? To what extent must the functions be performed, and how do those functions get precisified? For example, what ways of adding matter qualify as "growth"? What ways of

[19] Bostrom, *Superintelligence*, ch. 8 envisions a scenario of this kind.

producing something similar to oneself count as "reproduction"? And so on. Different ways of resolving this indeterminacy will yield different precise properties. Probably some precisifications will include future AIs and others won't. But if it can be a mere verbal question whether something is alive, then life would seem to be an inappropriate condition on ensoulment, for it's presumably not a verbal question whether a given physical system is ensouled.[20]

It's possible, of course, that some particular precisification of "alive"—$life_{217}$, say—is a requirement on ensoulment, and that this precisification is one that excludes future AI. But the *mere possibility* that some AI-excluding precisification of "life" is a requirement on ensoulment does not threaten my argument. To threaten my argument, we would need to be justified in having a high credence that some such requirement actually holds. It's hard to see how this could be justified.

This brings us to the second major problem with the life response. Even if we assume that future AIs aren't alive in the relevant sense (but aliens are), this is only a good reason to be confident in the rejection of AI ensoulment if we have strong reason to suppose that only living things can possess a soul. But as far as I can see, we don't have reason to suppose this—at least if "life" is understood in such a way that exhibiting the functions and behavior associated with mentality is not sufficient for life, which is a presupposition of the life response.

One might argue: the only things we know to have souls are humans (and perhaps other animals), all of which are living things, so that's some reason to accept a life-requirement on ensoulment. But this style of argument would equally undermine the alien ensoulment

[20] Tye makes a similar point, that if consciousness is never vague, we shouldn't accept views according to which consciousness depends on vague physical conditions, such as having approximately 40MHz neural oscillation (Tye, *Vagueness and the Evolution of Consciousness,* ch. 1).

hypothesis, whose plausibility has already been defended, since one could say the same about *any* property that humans have and the aliens lack (e.g., being carbon-based). An Aristotelian might say: a soul is simply a principle of life, so trivially only a living thing can have a soul. I have two responses: First, I doubt anyone is justified in having a very high credence in Aristotle's metaphysics of biology. Second, the dispute here may to some extent be verbal. If "soul" for you is analytically tied to the concept of life, then substitute another word (perhaps "spirit"). I am concerned with whether future AI systems will be associated with immaterial entities of a kind that serve as the basis of mental capacities like consciousness, thought, and so on. If one insists that it's improper to call such an entity a soul if the matter associated with it is non-living, I reply that I am not concerned with what we call it. If one insists, alternatively, that any matter joined to this kind of entity would *ipso facto* count as "living," then I reply (reiterating a point above): in that case, one is not justified in denying that future AI systems will be alive.

A related response to (A2), again with a somewhat Aristotelian flavor, says that the relevant difference between the AI and the alien is teleological. The AI has a kind of teleology, but it is merely "derived teleology." Its purposes are derivative from the intentions of its designers, or so one might think. The aliens, on the other hand, might exhibit underived intentionality. Like living things on earth, perhaps the aliens (or their parts or faculties) have functions or purposes that don't simply derive from designer intentions. And perhaps only things with underived teleology have souls.

As with life, we can distinguish reductive and non-reductive accounts of underived teleology. According to one popular reductive account, underived teleology is grounded in

evolutionary history.[21] Very roughly, evolutionary accounts hold that the function of (say) an organ is the activity that causally explains the existence and continuation of that organ within the relevant population via the mechanism of natural selection. This is a historical condition, and in the next section I'll suggest that historical conditions probably aren't relevant to ensoulment. But in any case, an AI could meet this condition. Some AI systems are created via evolutionary algorithms, whereby a selection function identifies the top performers in a population of programs, which are then copied with small adjustments to its parameters, and on and on, applying the selection function to successive generations of programs.[22] The same point holds for accounts that ground teleology not in backward-looking facts about the evolutionary past, but in forward-looking facts about what a thing does or would do. If x's having the underived function to φ can be rooted in the fact that x actually φs, or would φ under a sufficient range of counterfactual circumstances, there would appear to be no obstacle to an AI exhibiting underived teleology in this sense.

Alternatively, one could accept a non-reductive account, according to which teleology isn't grounded in designer intentions, evolutionary history, or in actual or counterfactual behavior. Rather, to have the function of φ-ing is just to have a primitive directedness toward φ-ing. However, while some notion of teleology surely applies to living things (or to their parts or faculties), it is far from obvious that living things exhibit this metaphysically inflationary kind of teleology. But supposing they do, why should we be confident that future AI systems won't have intrinsic teleology? Indeed, I suspect the most plausible account of intrinsic teleology involves a

[21] Millikan, *Language, Thought, and other Biological Categories*; Neander, "Functions as Selected Effects."

[22] See Chalmers, "The Singularity: A Philosophical Analysis" for discussion of the possibility of achieving human-level AI through some such evolutionary process.

kind of soul-based directedness, where (very roughly) for a material system to have intrinsic teleology is for it to be united to a soul that directs it toward such-and-such end.[23] In that case, denying intrinsic teleology to the AI would require independent grounds for denying the AI ensoulment hypothesis.

The temptation to deny that an AI could have underived teleology may be due to the fact that, like most artifacts, an AI may have derived teleology rooted in the intentions of its designers. Perhaps there is a tacit "exclusion assumption" that derived teleology somehow excludes underived teleology, so that a thing can't have underived functions if it has designer-imposed functions. But this exclusion assumption is incorrect. Future scientists may create animals *in vitro* with various intentions, but the scientists' creative intentions would not preclude the animals from exhibiting the kind of underived teleology found in naturally produced animals of the same kind. For example, if scientists create a dog *in vitro* with the sole intention that its ears flop about in a cute manner, this would not prevent its ears from having the underived function of enabling hearing.

Another closely related response is that the relevant difference between the alien and the AI is that only the former is a *substance* in a heavyweight neo-Aristotelian sense.[24] According to this response, the alien, like a terrestrial organism, is a primitive unity, a whole prior to its parts, a being whose activity and development is directed by an internal species-specific principle that determines the kind of being that it is.[25] The AI, in contrast, is merely a structured aggregate,

---

[23] Pruss defends such an account (Pruss, *Norms, Natures, and God*).

[24] Thanks to an anonymous referee for encouraging me to consider this response.

[25] See Inman, *Substance and the Fundamentality of the Familiar*, Oderberg, *Real Essentialism*, and Moreland and Rae, *Body and Soul* for accounts of substance along these lines.

with parts prior to the whole, something that doesn't belong to any natural kind and isn't directed by any internal principle, but is assembled and structured by external causes (its human makers). And perhaps only heavyweight substances have souls. In other words, perhaps a structured collection of material parts (like those that make up your body, an alien's body, or a computer) is united to a soul only if those material parts are incorporated into a heavyweight substance.

As before, I doubt whether anyone is justified in having a high credence in the metaphysical assumptions behind this response. But if we accept the background metaphysical scheme (including the existence of heavyweight substances, the distinction between heavyweight substances and mere aggregates, the assumption that terrestrial organisms are heavyweight substances, and the claim that only heavyweight substances have souls), it's unclear why we should deny that future AIs will be heavyweight substances. We need to consider the epistemology of heavyweight substancehood. Given a structured collection of material parts, what would be evidence that those parts are incorporated into a heavyweight substance? For example, when we encounter our hypothetical aliens, what observations would convince us that the matter of their bodies is incorporated into a heavyweight substance? Presumably the evidence would include the fact that they seem to behave intelligently, the fact that their parts seem to work in concert towards goals, the fact that the alien body coherently develops its capacities over time in response to environmental input, as well as the other high-level functional features associated with life mentioned above. We've seen that future AIs may exhibit many such features. These may only be fallible markers of substancehood, but if and when they are present in an AI system, we will have presumptive evidence that it is a heavyweight substance. Now, this presumption might be defeated by further considerations. (The most interesting potential defeater is that, arguably, we can explain why the AI exhibits the relevant features by appeal to purely

physical or mechanistic causes, without positing any heavyweight "internal principle" directing the system's development (but not so for the alien?). This will be addressed in §3.3 on "causal differences.") But given that future AIs may exhibit several plausible empirical markers of heavyweight substancehood, we cannot simply assume without argument that future AIs won't be heavyweight substances.

I conclude this section by noting a shortcoming common to several responses considered above. Many of these responses claim that a material system is associated with a soul only if it has a certain metaphysical status, such as being a heavyweight substance, exhibiting primitive teleology, or enjoying non-reductive life. Even if we set aside doubts about whether *we* have this metaphysical status, difficult epistemological questions remain concerning which other things have this status. What observable features would be evidence that a given material system has or doesn't have this status? Not only do these questions seem just as difficult as the corresponding question about ensoulment with which we started, but their answers seem to largely overlap. Independently plausible empirical markers of ensoulment (such as the abstract patterns of behavior, functioning, and development common to humans and our hypothetical aliens and future AIs) are also plausible empirical markers of heavyweight substancehood, primitive teleology, and non-reductive life (and vice versa). For this reason, shifting the debate to whether future AIs will have the relevant metaphysical status does very little to help us answer our original question of whether future AIs will have souls. The metaphysical assumptions behind these responses, if true, may shed light on the metaphysics of ensoulment, but they don't seem to help much with the epistemology of ensoulment.

3.2 Historical Differences

The next response is the historical response, according to which the relevant difference between aliens and AIs lies in their causal history. There are many historical differences between the AI and the alien. For example, AIs are intentionally created by people—they are artifacts—while the aliens arose through Darwinian evolution, or so we can assume. One might think that only non-artifacts are eligible for ensoulment. But it's hard to see why this difference, or any purely historical difference, would be relevant to ensoulment. To take a variation on the example above, suppose scientists are someday able to create something physically just like a human embryo "from scratch," assembling it molecule by molecule in a lab. It might then be made to gestate in an artificial or natural womb, resulting in a fully developed human body. Very plausibly, this lab embryo would receive a soul. (Substitute later-stage human fetus if you accept delayed ensoulment.) If so, being an artifact doesn't make something ineligible for ensoulment, since in this case the embryo is an artifact, something intentionally produced by people.

Some neo-Aristotelians might respond that, because the ensouled lab embryo is a heavyweight substance, it is not properly described as an artifact, despite having an artifact-like etiology. According to this response, only "aggregates" belong to the category of artifact.[26] (At most, we can say that, immediately prior to ensoulment, the embryo-like physical configuration produced by the scientists was an artifact. But upon ensoulment, the matter that previously constituted the artifact became incorporated into a newly generated heavyweight substance—a human embryo—which does not count as an artifact.) But this move would undermine the artifact response in a different way. For if one insists that something with an artifact-like etiology doesn't count as an artifact when it's a heavyweight substance, then (for the reasons given in

[26] Cf. Oderberg, *Real Essentialism*, 166–70

§3.1) we can no longer assume that future AIs will be artifacts. Like the lab embryo, they might merely have an artifact-like etiology.[27]

Our verdicts about lab-made embryos lend some support to the idea that ensoulment is history-insensitive, in the following sense:

> **History insensitivity**: Whether some matter is associated with a soul at *t* only depends on what's going on with that matter at *t*.

History insensitivity might be further supported by our judgments about related scenarios, like "swamp man"—a perfect physical replica of a human being produced by a chance assemblage of atoms thrown up by lightning striking a swamp.[28] To my mind, it would be unreasonable to be very confident that swamp man doesn't have a soul, provided we do.

Despite the initial plausibility of history-insensitivity, there are views that reject an analogous history-insensitivity claim for consciousness. According to these views, whether an individual is conscious (and the specific character of his conscious experiences) hinges on facts about the distant past. This commitment is found in the externalist representationalism of Tye, Dretske, Lycan, and others.[29] According to externalist representationalism, consciousness requires contentful internal representations, and internal representations acquire their contents in virtue of historical facts, such as the correlations between brain states and external properties throughout an organism's evolutionary history. Now, the views of Tye, Dretske, and Lycan are forms of reductive materialism, so they conflict with our substance dualist assumptions. But one could take the extrinsic, history-sensitive properties that these theorists *identify* with

---

[27] Thanks to an anonymous referee for encouragement to consider this response.

[28] Davidson, "Knowing one's own Mind."

[29] Tye, *Ten Problems of Consciousness*; Dretske, *Naturalizing the Mind*; Lycan, *Consciousness and Experience*.

consciousness, and instead posit that they are the *direct causal basis* of consciousness.[30] Within a substance dualist framework where consciousness requires ensoulment, we might treat the relevant historical properties as nomologically necessary for ensoulment.

The resulting history-sensitive form of substance dualism is coherent, but is not especially attractive. First, history-sensitive accounts of consciousness are unnatural bedfellows with substance dualism. The former are usually motivated by a desire to uphold a reductive, naturalistic conception of the mind (given that the most plausible reductive theories of content are externalist and history-sensitive).[31] The substance dualist doesn't share this desire. Second, it is very odd to suppose that the direct causal basis of consciousness is a fact that constitutively depends on the distant past. This doesn't seem to be the way causation works elsewhere in nature. If a particle is caused to swerve, for example, the proximate cause only includes conditions at or immediately prior to the time of the swerve. The immediate cause isn't some fact that constitutively involves occurrences decades or centuries before the swerve. Third, history-sensitive substance dualism would inherit all the standard problems with the more familiar (reductive) history-sensitive views of consciousness. For example, one common objection to these views is that they yield the counterintuitive verdict that swamp man is a zombie.[32] Another objection, which has been vigorously pressed by Adam Pautz, is that historical-externalist views seem to conflict with empirical discoveries in neuroscience and psychophysics, such as the fact that the resemblance structure of phenomenal color space doesn't match the resemblance

[30] For discussion of dualist views along these lines, see Dalbey and Saad, "Internal Constraints for Phenomenal Externalists," and Cutter, *Sensory Experience and the Sensible Qualities*, 172.

[31] See, e.g., Lycan, "The Case for Phenomenal Externalism," 21.

[32] But see Tye, *Ten Problems of Consciousness*, ch. 5, for attempt to avoid this result

structure of the external reflectance properties that are identified with phenomenal colors by historical-externalist views.[33] For these reasons, I conclude that adopting a history-sensitive substance dualism to respond to (A2) is not a promising response to the alien-analogy argument.

3.3 Causal Differences

A final response to (A2) is the causal-closure response, according to which the relevant difference between aliens and AIs lies in whether they satisfy physical causal closure. A background assumption of the causal-closure response is that interactionist dualism is true for humans. In other words, the soul has a non-redundant causal influence on physical processes in the human body, at least within the brain. Hence, within a human brain, events sometimes occur that cannot be fully causally explained by purely physical events, such as immediately prior neural activation levels or microphysical processes. As an illustrative device, we can imagine a Laplacian demon who is given a complete physical description of the world at a time, together with the laws govern the behavior of physical entities when no immaterial influences are at work. On the basis of this knowledge, the demon has an expectation about how the matter in your brain will behave, assuming it isn't subject to non-physical influences. If interactionism is true, the demon is sometimes surprised. The physical events in your brain sometimes violate his expectations. In other words, the neurons or the atoms in your brain occasionally behave differently from what one would anticipate on the basis of physical laws and physical prior conditions alone.

If interactionism is true, then physical causal closure (hereafter, just "causal closure") fails for human bodies. If so, it would be natural to suspect that causal closure fails for other ensouled physical systems. Perhaps this is the relevant difference between aliens and AIs: causal

[33] Pautz, *Perception*, ch. 4.

closure fails for alien bodies, but not for future AI systems. The demon would be surprised by some physical occurrences within the alien's body, but not by any physical occurrences within future computers.

Some dualists reject the claim that causal closure is violated even in the human body, favoring epiphenomenalism or widespread overdetermination.[34] The causal-closure response is unavailable to them. But here I'll assume, for the sake of argument, that causal closure fails for any ensouled physical system.

I grant that, if we discovered that physical causal closure holds for future AI but not the aliens, this difference would be evidence for a difference in their ensoulment status. Conversely, if we discovered that causal closure holds for the alien bodies but not for the AI, this would be evidence for a difference in the opposite direction. But all of this is irrelevant to the current argument. If we think that aliens and AIs differ with respect to causal closure, this belief surely rests on a prior belief that they differ with respect to ensoulment (or some related status, like having irreducible mentality). This prior belief would require independent justification.

In our hypothetical alien encounter, we can assume we don't have any *direct* evidence for violations of causal closure within alien bodies—that is, reasons to accept causal-closure violations that don't rest on a prior belief that aliens have souls or irreducible mentality—just as we don't have direct evidence for violations of causal closure in human bodies. (We haven't identified specific anomalous events in the nervous system where our neural activity clearly begins deviating from what we would expect if only physical causes were operative. If one

[34] See Campbell, *Body and Mind* and Jackson, "Epiphenomenal Qualia" for a defense of epiphenomenalism. See Lowe, *Subjects of Experience* and Mills, "Interactionism and Overdetermination" for a defense of overdeterminationism.

believes that causal closure is violated in human bodies, this belief likely rests on a prior belief that humans have irreducible mental states, together with further assumptions to rule out epiphenomenalism and overdetermination. The belief that humans have souls or irreducible mentality is not based on a prior discovery that physics is inadequate to explain certain bodily processes.[35]) If we come to accept that aliens have the kind of irreducible mentality that goes with ensoulment, it may be reasonable to assume that their irreducible mental states aren't epiphenomenal and don't overdetermine their effects alongside redundant physical causes. From here, we can conclude that there must be causal closure violations within alien bodies. But this conclusion would rest on a prior belief that aliens have irreducible mentality of the kind associated with ensoulment, a belief that would have to be justified on other grounds, just as in the human case.

Similarly, we don't currently have any direct evidence concerning whether causal closure holds for future AI, for we haven't even observed these machines. Of course, some will hold that we have inductive justification for the belief that *all* physical systems satisfy causal closure. But we are assuming, with the objector, that there is an exception for ensouled physical systems. Under this assumption, it's hard to see how one could be justified in holding that future AIs will satisfy causal closure unless one has independent reasons to reject AI ensoulment (in which case the argument should focus on those independent considerations). It might be said that any behaviorally intelligent AI system can be expected to satisfy causal closure, since if there were external, non-physical influences on its physical operations, we would expect it to glitch or crash. After all, digital computers are delicate instruments. Many ways of fiddling with their physical components, even in small ways, can cause the whole system to grind to a halt. (It is

[35] Cf. Chalmers, *The Conscious Mind*, 101–2, 109.

sometimes noted that biological brains differ from standard digital computers in this respect, which is why the former but not the latter exhibit "graceful degradation.") So, if a physical computer produces seemingly intelligent behavior instead of crashing, we can safely assume it *isn't* subject to external non-physical influence, and so probably isn't ensouled.

This argument is unconvincing. It may be that an external influence that randomly fiddles with the physical processes in a computer would very likely cause it to crash. But random fiddling is not the kind of influence we should expect on the AI ensoulment hypothesis. In the human case, we can safely assume that (given interactionism) our souls don't randomly fiddle with the physical processes in our brains, but shape and structure those processes in the direction of increased order and rational coherence, resulting in the kinds of behaviors that we take to manifest intelligence. On the AI ensoulment hypothesis, we should expect a similar type of influence on the AI hardware, not random fiddling that leads to crashes and glitches.

It is an interesting question whether we should expect to get evidence in the future about causal closure in advanced AI systems. We might try running the same AI program on two computers that start in the same physical state, feed them identical inputs, and observe whether the outputs differ. If the computers satisfy causal closure, we would expect the same outputs. But this wouldn't be decisive evidence for causal closure because we wouldn't necessarily expect different outputs if causal closure were false. Perhaps the input causes the computer to enter physical state P1, which causes ("bottom-up") soul state M, which then causes ("top-down") physical state P2, which then leads to the output. If there is no indeterminism in these causal transitions, we would expect the same outputs for the two systems, even if the computers are united with interactionist souls. But it could be argued that a degree of indeterminism is to be expected on the ensoulment hypothesis, perhaps because we should expect that ensoulment

confers incompatibilist free will. In that case, we would expect AIs that start in physically identical states and receive the same inputs at least occasionally to produce different outputs. If we sometimes observe different outputs, this would be evidence for ensoulment. If we always observe the same outputs, this would be evidence against ensoulment. In this way, and perhaps in other ways as well, we might be in a position to get evidence about AI ensoulment in the future. The strength of this evidence would depend largely on the degree to which we should expect the relevant kind of indeterminism conditional upon ensoulment, something I will not try to adjudicate here.[36]

4. The Fitting-Recipient Argument

The next argument rests on the conjecture that a physical system is united to a soul whenever that physical system is "fit" to possess a soul. Very roughly, for a physical system to be fit to possess a soul is for it to be structured in such a way that it can complement or meaningfully cooperate with the operations of the soul. Turing (1950: 443) invokes a similar idea in the course of arguing that a theistic substance dualist should accept the possibility of machine ensoulment, suggesting that a brain or machine may need a certain degree of complexity and sophistication in order to "minister to the needs of [the] soul." To a first pass, we can formulate the fitting-recipient argument as follows:

[36] There is a potential epistemic asymmetry here worth noting. Suppose one is certain that the systems will produce the same outputs given non-ensoulment, but only (say) 50% confident that they will produce different outputs given ensoulment. Then observing different outputs would be decisive evidence for ensoulment, but observing the same outputs would only be modest evidence against ensoulment. So, for example, if one's prior for AI ensoulment is 0.5, then given the likelihoods above, updating on the observation of same outputs would lower this probability to 1/3, while updating on the opposite observation would raise this probability to 1.

F1. Being fit to possess a soul is a nomologically sufficient condition for a physical system to be united to a soul.

F2. A functionally human-like machine (i.e., a machine that relevantly resembles a human body/brain in its functional organization and behavioral capacities) would be fit to possess a soul.

C. Therefore, a functionally human-like machine would be united to a soul.

Whether a body is fit to receive a soul is partly a matter of what the body is like and partly a matter of what the soul is like. In particular, fittingness will be a function of the basic powers or operations of the soul. My concern here is whether a machine might be fit to receive a soul with powers relevantly like ours. I'll assume that our souls have certain basic powers or capacities, including a capacity for sensory experience, judgment, practical and theoretical reasoning, and choice. From here, the notion of being a fit to possess a soul with such powers is best illustrated by examples of the ways in which the states and powers of the human body complement the basic powers or operations of the soul.

Inputs to sensory consciousness: First, the body has a rich set of internal states (in particular, the neural representations in the perceptual systems of the brain) that reliably correlate with external states of the environment. These physical states can serve as the direct causal basis of the soul's states of sensory consciousness. This enables our states of sensory consciousness to provide a rich and detailed model of our environment, on the basis of which we can form many rational and accurate beliefs about our environment. Most other physical systems, like rocks or chairs, do not have this feature. There is no set of internal states of a rock that could serve as inputs to sensory experience that would enable a soul to richly and reliably model the environment. (The internal states of the rock have *some* connection with the ambient

environment. For example, their temperature roughly tracks the ambient temperature. But the connections are very thin, and any sensory model causally rooted in these states would be extremely impoverished.)

Behavioral repertoire: The body has a rich behavioral repertoire that can express the complexities and subtleties of our mental life. As our outward behavior has its proximate cause in the brain, our behavior can express the subtleties of our mental life because our brains have enough structure to reflect or encode these subtleties. J.P. Moreland makes a similar point in his explanation of why God created bodies.

> Bodies provide power for action in the created world. Further, the more complicated an animal's consciousness is, the more complex and finely tuned the body would need to be to be responsive to the fine-graded mental states in causal interaction with it. Consider a form of consciousness with a complexity sufficient to engage in a variety of quite specific actions associated with precise nuances in thought, believe, emotion, desire, and so forth. On this view, if such a consciousness were causally connected to a material object without the physical complexity needed to register in the physical world the appropriate mental complexity, that mental complexity would be wasted.[37]

Again, typical physical systems, like rocks or chairs, do not have a similarly rich behavioral repertoire. If a soul were joined to a rock, then even if the internal operations of the soul were complex and subtle, its inner life could not be adequately reflected in the behavior of the rock. The rock simply doesn't have a rich enough behavioral repertoire or a rich enough internal organization to proximately guide behavior.

Neural representations and reasoning: Third, the brain is structured in a manner that may facilitate the soul's powers of reasoning, providing a rich store of neurally encoded memories that the soul may draw on, as well as short-lived neural representations in working memory that may facilitate reasoning, serving as a kind of "neural scratch pad." A dualist is likely to hold that

[37] Moreland, "The Argument from Consciousness," 335.

no sequence of physical events in the brain *is* a process of reasoning. But some such sequences may aid or complement the process of reasoning, just as a sequence of marks on a scratch pad may aid one's reasoning without constituting one's reasoning.[38] Again, a typical physical system, like a rock or a chair, presumably doesn't have internal states that can facilitate reasoning in this way. Hence, a rock is less suited to cooperate with the soul's powers of reasoning than the human body/brain.

In these and many other ways, the human body seems to be structured in a way that allows it to facilitate or cooperate with the operations of the soul. In this sense, it is "fit to possess" a soul. Moreover, it is natural to suppose that the explanation for why a living human body possesses a soul, while rocks and human corpses (for example) do not, is that a living human body is *fit* to possess a soul in this sense, while rocks and corpses are not. Indeed, I suspect that our intuitive judgments about ensoulment in actual and counterfactual cases are implicitly guided by a fittingness condition. For example, we find it natural to suppose that intelligently behaving aliens have souls (conditional on substance dualism in the human case), and this is plausibly because their bodies are structured in a way that would complement the operations of a soul. On the other hand, while it may be conceivable that souls are joined to things like rocks or chairs or corpses, we do not take such hypotheses seriously, and this is likely because they don't have a physical organization that could allow them to meaningfully cooperate with the operations of a soul.[39] These considerations lend support to F1, the conjecture that being

[38] *Pace* Clark and Chalmers, "The Extended Mind."

[39] We've seen that fitness to possess a soul is relative to the powers of the soul. It could be that there are many kinds of souls, with different sets of basic powers, in which case a physical system that isn't fit to receive a soul relevantly like ours is fit to receive a different (perhaps more basic) type of soul. For example, one could hold that a human

fit to possess a soul is a nomologically sufficient condition for a physical system to be united to a soul.

Now, fittingness in the sense described above seems to be a matter of degree. A system can have, to varying degrees, a structure that allows it to cooperate with the operations of the soul. (And surely among human beings there is variation on this score, for example due to brain damage or developmental immaturity.) Where should we set the threshold? For our purposes, we can set the threshold high, at the level of a typical (living, adult) human body. Since F1 only offers a sufficient condition for ensoulment, a high-threshold reading of F1 does not entail that systems falling short of the threshold lack souls.[40]

Let's turn to F2: a functionally human-like machine (i.e., a machine that relevantly resembles a human body/brain in its functional organization and behavioral capacities) would be fit to possess a soul. The motivation for F2 is that being fit to possess a soul is entirely a matter of the functional organization and behavioral capacities of a physical system. Note that all of the examples above illustrating the human body's fitness to possess a soul involved only the behavioral capacities and abstract functional properties of the human body. For example, we mentioned the fact that the brain has a rich set of internal states that can serve as inputs for

---

body is fit to receive soul with rational powers (inter alia), while an animal body is only fit to receive a lower-grade soul without rational powers.

[40] Might a lower threshold interpretation of fittingness provide a necessary *and* sufficient condition for receiving a soul with basic powers relevantly like ours? Perhaps, but there are some difficulties with this idea. A very low-threshold might have implausible consequences bordering on "pan-ensoulment," while a moderate to high threshold might be overly restrictive, potentially excluding human infants and the severely mentally disabled. Some of these problems might be avoided with a disjunctive ensoulment condition: the physical system either meets, or has the natural potential to develop into something that meets, such-and-such fittingness threshold.

sensory consciousness, allowing sensory consciousness to richly model the environment. This feature of the brain is substrate-independent. A computer could have the same feature, with the relevant states having a non-biological material realization. The same goes for the other examples. Since fitness seems to depend only on behavioral capacities and functional organization, and since the human body is fit to possess a soul, it follows that a functionally human-like machine would be fit as well.

Given F1 and F2, a functionally human-like AI would have a soul. While I think F1 and F2 are both plausible, I don't think we can be certain of either. F1 is a reasonable hypothesis, but like most interesting hypotheses about the nomologically sufficient conditions for ensoulment, it is speculative. There may also be grounds for doubting F2. Perhaps the most significant concern is that digital computers are effectively deterministic (deterministic "for all practical purposes," as van Inwagen puts it[41]). Even if there is physical indeterminism at microscopic scales, computers are designed in such a way that their macro-level behavior is insensitive to micro-scale indeterminism. In contrast, some have suggested that physical processes in the brain might be sensitive to quantum indeterminism, with the brain amplifying microscopic indeterminism into macroscopic indeterminism at the scale of observable behavior.[42] One could argue that a physical system that is effectively deterministic, in the manner of a digital computer, is thereby unfit to be united to a soul with powers of free choice. The thought would be that free choice requires a degree of indeterminism, so if a system's behavior is going to reflect its choices, there must be indeterminism at the level of macroscopic behavior. One could argue that, if a soul is

[41] van Inwagen, *An Essay on Free Will*, 198.

[42] Swinburne, *The Evolution of the Soul*, appendix D; van Inwagen, *An Essay on Free Will*, 199, Kane, *The Significance of Free Will*, ch. 7.

joined to a physical system structured so as to be physically deterministic at macro-scales, the soul's agency can only induce indeterminism in the system's behavior by somehow doing violence to its physical structure.

It is hard to evaluate this objection without a clearer understanding of what the substance dualist should say about the relationship between physical indeterminism and free will—a vexed issue that I cannot attempt to sort out in this paper. But without resolving this difficult question, two remarks may somewhat blunt the force of this concern. First, even if one thinks that an effectively deterministic physical system is not fit to receive a soul with incompatiblist freedom, such a system might still be fit to receive a soul with powers of rational agency that don't involve incompatibilist freedom (e.g., the kind of agency endorsed by compatibilists). Second, it's possible that some future computers or artificial agents will not be effectively deterministic in the manner of current digital computers.[43]

While there are modest grounds for doubting F1 and F2, the aim here is only to argue that we ought to have at least middling credence that a functionally human-like AI would have a soul. It is sufficient for our purposes if the conjunction of F1 and F2 are moderately plausible, warranting at least a middling credence. The considerations above in support of F1 and F2 are, I think, adequate to support this modest claim.

5. Conclusion

I've argued that we should take seriously the hypothesis that future AIs will be endowed with souls, conditional on substance dualism and our eventual creation of AGI. If we take the AI ensoulment hypothesis seriously, several important questions arise. I conclude by mentioning just a few.

[43] Schwitzgebel and Garza, "A Defense of the Rights of Artificial Intelligences," 104.

First, there would be moral questions: if an AI has a soul, then plausibly it has a significant degree of moral status. Should we regard an ensouled AI as morally on a par with humans? One might argue that we should give greater moral consideration to our fellow humans since we are members of the same species, just as we arguably have greater obligation to family members than to non-family. But one could equally argue that we have *greater* obligation to AIs on the grounds that we are their creators, just as we have greater obligation to our children on account of bringing them into existence.[44] Within a broadly Kantian ethical framework, there would be interesting questions about what it could mean to respect an AI's autonomy when we are the ones designing its goals and motivations.[45] Is it permissible to create an ensouled AI that will be used as an instrument for human purposes? Is it better or worse if we design its motivations so that it *likes* being used as an instrument for human purposes?

Second, there may be difficult questions about personal identity. The personal identity literature is full of strange scenarios involving teletransporters, brain hemisphere transplants, amoeba men, and the like. These scenarios present theoretical difficulties, but few practical difficulties in the human domain. In practice, we don't have to worry about whether to hold a teletransporter copy of a human liable for debts incurred prior to teletransportation, or which of two fission products should be held responsible for pre-fission crimes, or whether to give voting rights to each of a thousand recently made copies of a single human. But with digital agents, scenarios like these could be commonplace, as the data that constitutes an AI is transferred onto new hardware, or copied and run on multiple systems, or altered or edited in countless ways. If future AI systems are soulless—purely physical—then questions about what happens to the

[44] Schwitzgebel and Garza, "A Defense of the Rights of Artificial Intelligences," 108–10

[45] Schneider, *Artificial You*.

original AI in many such scenarios are liable to strike us as "empty questions," to borrow Parfit's phrase.[46] They would be akin to asking whether we have the "same watch" after disassembling and reassembling the parts, or the "same copy of *The Cat in the Hat*" after erasing and rewriting the words on the original pages. But as Parfit emphasized, such questions about personal identity would be substantive and normatively significant if the subject in question has a soul, for it is always a substantive question whether a given soul persists through a given change.

Third, there would be metaphysical questions about the kinds or grades of souls that future AIs may possess. For example, if we ever manage to create superintelligent AGIs, machines that dramatically outperform humans on all or most behavioral measures of intelligence, might they have souls of a different or "higher" kind than ours? On some views, both humans and non-human animals have souls, but human souls are of a higher grade, in virtue of their rational powers. Perhaps superintelligent AGIs would have souls of a still higher grade, souls with powers that stand to rationality as rationality stands to mere sentience.

Fourth, there would be interesting theological questions, given certain theological background assumptions. If human souls continue to exist after our bodily death, might the same hold true for AI souls? If human souls have spiritual capacities, such as a basic power to know and experience God, might the same be true of AI souls?

No doubt many other questions could be added to the list. The goal of this paper has been modest: to clarify the AI ensoulment hypothesis and show that it is worth taking seriously. Once

[46] Parfit, *Reasons and Persons*.

taken as a live possibility, it opens many avenues for future research and raises a host of pressing ethical, metaphysical, and theological questions. There is much more work to be done.[47]

University of Notre Dame

References

Baars, Bernard. 1988. *A Cognitive Theory of Consciousness* (Cambridge University Press).

Beckwith, Francis J. 2007. *Defending Life: A Moral and Legal Case Against Abortion Choice*. (Cambridge University Press).

Block, Ned. 1978. "Troubles with functionalism." *Minnesota Studies in the Philosophy of Science* 9: 261-325. https://hdl.handle.net/11299/185298

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

Campbell, Keith. 1970. *Body and Mind* (Doubleday).

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press).

Chalmers, David. 2010a. *The Character of Consciousness* (Oxford University Press).

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

Clark, Andy and David Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1):7-19. https://doi.org/10.1093/analys/58.1.7

Collins, Robin. 2011. "A Scientific Case for the Soul." In *The Soul Hypothesis: Investigations Into the Existence of the Soul*, edited by Mark C. Baker and Stewart Goetz (Continuum Press), 222-246.

Cutter, Brian. 2015. *Sensory Experience and the Sensible Qualities*. Ph.D. Dissertation. http://hdl.handle.net/2152/31630

Dalbey, Bryce, and Bradford Saad. 2022. "Internal Constraints for Phenomenal Externalists: A Structure Matching Theory." *Synthese* 200 (5): 1–29. https://doi.org/10.1007/s11229-022-03829-1

Davidson, Donald. 1987. "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 60 (3): 441–58.

[47] Thanks to Alexander Pruss, Ben Page, and Parker Settecase for valuable discussion. Thanks to Brandon Rickabaugh and two anonymous referees for helpful comments.

Dretske, Fred. 1995. *Naturalizing the Mind* (MIT Press).

Feser, Edward. 2018. “Aquinas on the Human Soul.” In *The Blackwell Companion to Substance Dualism*, edited by Jonathan J. Loose, Angus John Louis Menuge, and J. P. Moreland (Wiley Blackwell), 88–101.

Fodor, Jerry. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (MIT Press).

Foster, John. 1991. *The Immaterial Self: A Defence of the Cartesian Dualist Conception of the Mind* (Routledge).

Grace, Katja, John Salvatier Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. "When Will AI Exceed Human Performance? Evidence from AI Experts." *Journal of Artificial Intelligence Research*, 62, 729–754. https://doi.org/10.1613/jair.1.11222

Hasker, William. 2001. *The Emergent Self* (Cornell University Press).

Hofstadter, Douglas. 2007. *I am a Strange Loop* (Basic Books).

Huemer, Michael. 2021. *Knowledge, Reality, and Value: A Mostly Common Sense Guide to Philosophy.*

Inman, Ross D. 2017. *Substance and the Fundamentality of the Familiar: A Neo-Aristotelian Mereology.* (Routledge).

Jackson, Frank. 1982. “Epiphenomenal Qualia.” *Philosophical Quarterly* 32 (April): 127–36. https://doi.org/10.2307/2960077

Kane, Robert. 1996. *The Significance of Free Will* (Oxford University Press).

Lamme, Victor. 2006. "Toward a True Neural Stance on Consciousness." *Trends in Cognitive Sciences* 10 (11):494-501. https://doi.org/10.1016/j.tics.2006.09.001

Lowe, E.J. 1996. *Subjects of Experience* (Cambridge University Press).

Lycan, William G. 2001. “The Case for Phenomenal Externalism.” *Philosophical Perspectives* 15: 17–35. https://doi.org/10.1111/0029-4624.35.s15.2

Millikan, Ruth. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism* (MIT Press).

Mills, Eugene. 1996. “Interactionism and Overdetermination.” *American Philosophical Quarterly* 33 (1): 105–15. https://www.jstor.org/stable/20009850

Moreland, J.P. 2012. "The Argument from Consciousness." In *The Blackwell Companion to Natural Theology*, edited by William Lane Craig and J.P. Moreland (Wiley-Blackwell), 282-343.

Moreland, J.P. 2018. “In Defense of a Thomistic-Like Dualism.” In *The Blackwell Companion to Substance Dualism*, edited by Jonathan J. Loose, Angus John Louis Menuge, and J. P. Moreland (Wiley-Blackwell), 102–122.

Moreland, J.P. and Scott Rae. 2000. *Body and Soul and the Crisis in Ethics* (Intervarsity Press).

Neander, Karen. 1991. Functions as Selected Effects: The Conceptual Analysts' Defense. *Philosophy of Science* 58 (2):168-184. https://doi.org/10.1086/289610

Oderberg, David S. (2007). *Real Essentialism* (Routledge).

Parfit, Derek. 1984. *Reasons and Persons* (Oxford University Press).

Pautz, Adam. 2021. *Perception* (Routledge).

Popper, Karl, and John Eccles. 1977. *The Self and Its Brain: An Argument for Interactionism* (Springer).

Pruss, Alexander. ms. *Norms, Natures, and God*.

Putnam, Hilary. 1967. "Psychological Predicates." In *Art, Mind, and Religion*, edited by W. H. Capitan and D. D. Merrill (University of Pittsburgh Press), 37-48.

Rickabaugh, Brandon. 2018. "Against Emergent Dualism." In *The Blackwell Companion to Substance Dualism*, edited by Jonathan J. Loose, Angus John Louis Menuge, and J. P. Moreland (Wiley-Blackwell), 73–86.

Rosenthal, David. 2005. *Consciousness and Mind* (Oxford University Press).

Schneider, Susan. 2021. *Artificial You:AI and the Future of your Mind* (Princeton University Press).

Schwitzgebel, Eric, and Mara Garza. 2015. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39 (1): 98–119. https://doi.org/10.1111/misp.12032

Swinburne, Richard. 2007. *The Evolution of the Soul, Revised Edition* (Oxford University Press).

Swinburne, Richard. 2018. "Cartesian Substance Dualism." In *The Blackwell Companion to Substance Dualism*, edited by Jonathan J. Loose, Angus John Louis Menuge, and J. P. Moreland (Wiley-Blackwell), 131–151.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59 (October):433-60. https://doi.org/10.1093/mind/LIX.236.433

Tye, Michael. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (MIT Press).

Tye, Michael. 2021. *Vagueness and the Evolution of Consciousness: Through the Looking Glass* (Oxford University Press).

Unger, Peter. 2005. *All the Power in the World* (Oxford University Press).

van Inwagen, Peter. 1983. *An Essay on Free Will* (Oxford University Press).

Zimmerman, Dean. 2011. “From Experience to Experiencer.” In *The Soul Hypothesis: Investigations Into the Existence of the Soul*, edited by Mark C. Baker and Stewart Goetz, (Continuum Press) 168-201.