Everything and More: The Prospects of Whole Brain Emulation

Eric Mandelbaum

CUNY (Baruch College & The Graduate Center)

Forthcoming in *The Journal of Philosophy*

(Penultimate Version. Please cite whatever you want, as the Jphil version won't be out in print until right before the heat death of the universe)

Abstract

Whole Brain Emulation (WBE) has been championed as the most promising, well-defined route to achieving both human-level artificial intelligence and superintelligence. It has even been touted as a viable route to achieving immortality through brain uploading. WBE is not a fringe theory: the doctrine of Computationalism in philosophy of mind lends credence to the in-principle feasibility of the idea, and the standing of the Human Connectome Project makes it appear to be feasible in practice. Computationalism is a popular, independently plausible theory, and Connectomics a well-funded empirical research program, so optimism about WBE is understandable. However, this optimism may be misplaced. This article argues that WBE is, at best, no more compelling than any of the other far-flung routes to achieving superintelligence. Similarly skeptical conclusions are found regarding immortality. The essay concludes with some positive considerations in favor of the Biological Theory of consciousness, as well as morals about the limits of Computationalism in both artificial intelligence and the philosophy of mind more generally.

The Promise of Whole Brain Emulation

Whole Brain Emulation (WBE) has been proposed as the most promising avenue for creating human-level artificial intelligence, creating superintelligence, and even for achieving immortality.[1] The basic goal behind WBE is to create a software model of one's mind which could then be uploaded to a storage system. The resulting software model could then be downloaded to new hardware, and the uploaded model can, so the idea goes, be used to recreate a functional isomorph of the original brain from which it was copied. Doing so would then, by hypothesis, replicate all the psychological features that were present in the original individual whose brain was copied.

What level of detail might be needed for an upload? *Prima facie*, all one would need is a "connectome."[2] A connectome is an anatomical wiring diagram that charts the connections between each neuron, giving us an overall wiring map of the brain. The idea is that by replicating a wiring diagram of one's brain, we would thereby replicate one's psychology.[3]

Connectomics and WBE are natural partners, and together they appear to offer tantalizing possibilities. Connectomics is a well-defined research program, one which is well underway: The

---

[1] Sandberg, A., and Bostrom, N. 2008. Whole Brain Emulation: A Roadmap. Technical Report, 2008-3, Future of Humanity Institute, Oxford University, Oxford.

[2] Sporns, O., Tononi, G., and Kötter, R. 2005. The Human Connectome: A Structural Description of the Human Brain. *PLoS Computational Biology*, 1(4): 42.

[3] Seung, S. 2012. *Connectome*: *How the Brain's Wiring Makes Us Who We Are*. New York: Houghton Mifflin Harcourt.

Human Connectome Project is an interinstitutional, reputable, amply funded research project.[4] The success of the project could itself underwrite WBE, as the wiring diagram of the connectome may be thought to replicate the functional characteristics of one's brain. Consequently, WBE seems particularly well-placed among all transformative technologies as, crucially, it needn't rely on any conceptual breakthrough to ensure its success.[5] Here are Anders Sandberg and Nick Bostrom, two prominent futurists, on WBE's promise:

> WBE represents a formidable engineering and research problem, yet one which appears to have a well-defined goal and could, it would seem, be achieved by extrapolations of current technology. This is unlike many other suggested radically transformative technologies like artificial intelligence where we do not have any clear metric of how far we are from success.[6]

Sandberg and Bostrom suggest that WBE allows for a "clear metric" because we can understand how far we are from replicating a complete brain's wiring diagrams. In this way WBE stands alone as the only route to superintelligence that we currently appear to understand, at least in broad strokes. Moreover, we have made progress on this front, with some simple species' connectomes already mapped (e.g., *C Elegans'* connectome was mapped over thirty years ago).[7]

---

[4] See http://www.humanconnectomeproject.org/ for details.

[5] This claim assumes that we wouldn't conceptual breakthroughs in neuroscience to understand the connectome.

[6] Sandberg, A., and Bostrom, N. 2008, p8.

[7] White, J., Southgate, E., Thomson, J., and Brenner, S. 1986. The Structure of the Nervous System of the Nematode Caenorhabditis Elegans. *Philosophical Transactions of the Royal Society London B Biological Sciences* 314(1165):1–340.

At first blush, WBE seems tantalizing. Understanding the brain and mind is far too difficult a task to accomplish in any reasonable amount of time. WBE however, holds promise of being able to sidestep this worry: "A key assumption, characteristic of the WBE approach to AI, is nonorganicism: total understanding of the brain is not needed, just understanding of the component parts and their functional interactions" Sandberg 2013, p257).[8] WBE holds out hope that we can emulate the brain's functional apparatus without understanding (e.g.,) how neural structure itself leads to intentionality, consciousness, or intelligence. Because of how successful Connectomics has been, Sandberg estimates a 50% confidence level in the proposition that WBE will arise by 2064.[9] Even the rosiest optimists amongst us would need to put the chances that we have achieved anywhere near full understanding of the brain, never mind the mind, by then as infinitesimal. Thus, WBE seems much more promising than, say, creating complete models of the mind based on understanding everything about our disparate mental faculties and capacities (e.g., our characters, personalities, or ways of acquiring beliefs).

WBE is also tantalizing as it allows for the possibility of extremely lofty goals. If reproducing the connectome would reproduce the functional properties of the mapped brain, then WBE might hold the key for immortality. One's identity might be thought of as the totality of one's psychology—their personality, memories, emotions and the like. But the promise of WBE allows for other lofty goals besides immortality, as WBE seems like the clearest route to achieving

---

[8] Sandberg, A. 2013. Feasibility of Whole Brain Emulation. In Vincent C. Müller, ed., *Theory and Philosophy of Artificial Intelligence*, 251–264. Berlin: Springer.

[9] Sandberg, A. 2014. Monte Carlo Model of Brain Emulation Development. Future of Humanity Institute Working Paper, Oxford University.

superintelligence too.[10] WBE would allow for relatively cheap uploading and storage of human-level intelligence (which itself would constitute "weak superintelligence," human-level intelligence that can operate at much greater speeds).[11] As human capital is the central driver of economic growth, having large amounts of readily available human-level intelligences will make for enormous technological and societal enhancement.[12]

Since few technologies hold the promise of such transformative ends as immortality and superintelligence, the question of the feasibility of WBE is pressing, even if one's a priori intuitions of the chances of achieving it are more pessimistic than Sandberg's. My goal in this paper is to provide that analysis. I argue that one should have a healthy skepticism as to the fecundity of WBE. Moreover, the problems with WBE are not specific to it—showing the problems inherent in WBE will illuminate fissures in the doctrine of Computationalism writ large.

Whole Mind Emulation

I am interested in two questions: 1) would a connectome of a single mind suffice to instantiate a broad range of psychological features, features such as one's personality, character, intelligence, and

---

[10] Superintelligence is generally glossed as intelligence that far exceeds human capacities in every domain (Bostrom 2010). It is unclear how much a system must exceed human intelligence in order to qualify as superintelligent.

[11] Bostrom, N. 2006. How Long before Superintelligence? *Linguistic and Philosophical Investigations* 5(1): 11–30.

[12] See Hanson, R. 2008. Economics of the Singularity. *IEEE Spectrum*, 37–42; Hanson, R. 1994. If Uploads Come First: The Crack of a Future Dawn. *Extropy* 6(2): 10–15; Hanson, R. 2016. *The age of Em: Work, love, and life when robots rule the Earth*. Oxford: Oxford University Press.

phenomenology?;[13] and 2) assuming an affirmative response to the first question, would the connectome suffice for personal identity? Would the "software" to be uploaded ensure duplicating a given mind? Would uploading my connectome suffice for uploading *me*?[14]

To put it in Sandberg and Bostrom's terms, assume a brain emulator is a piece of software. Our question is then whether brain emulation so understood would entail *mind emulation*—a model that is "detailed and correct enough to produce the phenomenological effects of a mind" *inter alia* (Sandberg and Bostrom 2008, p7).

WBE can be a success even if it could not preserve full personal identity. To be successful depends on one's aims: if a connectome sufficed for establishing propositional attitudes, then uploading a connectome could allow for creating artificial intelligence.[15] If this intelligence can be

---

[13] A Connectome is a representation, so it may be better to speak of an instantiation of a connectome. For convenience I'll speak as if connectomes are instantiations as the distinction won't affect my argument moving forward. Similarly, when I speak of uploading connectomes, one may prefer to think about uploading instantiations of connectomes (see, chapter 8 of Schneider, S. 2019, *Artificial You* Princeton: Princeton University Press).

[14] Or would uploading my connectome suffice for even uploading a token of the type that is me? See Schneider ibid. for discussion.

[15] Propositional attitudes perhaps aren't, strictly speaking, necessary for intelligence. Perhaps an agent could be intelligent without believing or desiring anything. And perhaps, strictly speaking, attitudes aren't sufficient either: maybe there could be a creature with beliefs and desires but no combinatorial apparatus for generating rational thought or behavior. That said, having full-blooded propositional attitudes seems to make intelligence likely.

harnessed, the WBE might serve as the catalyst for superintelligence, even if WBE could not ensure immortality.

Physicalism, Multiple Realizability, and WBE

WBE seems *prima facie* feasible. It is a natural bedfellow of Computationalism, the idea that the mind is just a computer of sorts, where mental processes are understood as transformations of mental representations.[16] In a canonical formulation of computationalism, mental representations are taken to be symbols, and mental processes compute over the formal properties of the symbols.[17] Computationalism gains inspiration from the Church-Turing thesis which holds that any computable function can be computed by a Turing machine. Any software duplicate will be Turing machine equivalent. Unless one believes that the mind is somehow outside of the physical realm altogether then there should be no a priori restriction to the feasibility of WBE.

---

[16] Some may balk at just how close bedfellows these should be (e.g., Schneider 2019 which argues against the canonical reading of computationalism while still holding a broadly computationalist theory).

[17] See, e.g., Fodor, J. 1998. *Concepts: Where Cognitive Science Went Wrong.* Oxford: Oxford University Press; Block, N. 1995. "The Mind as the Software of the Brain," in D. Osherson, L. Gleitman, S. Kosslyn, E. Smith, and S. Sternberg, eds., *An Invitation to Cognitive Science.* New York: MIT Press. Of course, some other theorists that might be considered broadly computationalist—e.g., certain connectionists—would balk at discussions of symbolic computation. For a skeptical take on classical computationalist models see Schneider (ibid), and Schneider, S 2011. Language of Thought: A New Philosophical Direction. Cambridge: MIT Press.

To make the case as strong as possible for WBE, let us assume token physicalism in what follows so that the Church-Turing thesis holds over all neural events.[18] Given that, what are the roadblocks to WBE?

Sandberg and Bostrom write that WBE only requires that we find "a 1-to-1 model where all relevant properties of a system exist" (2008, p7). But what are the relevant properties of one's brain? This question is pressing. WBE relies on the idea that the end goal of computational neuroscience is to provide a neuroinformatic map of the brain. The detail of the map matters: if WBE requires a level of detail equivalent to molecule-for-molecule duplication, then it is far too information rich a plan to be feasible in the short term (where short term includes times as early as Sandberg's estimate of 2064). WBE proponents understand that they need to abide by "nonorganicism", and instead suggest that only a certain level of functional understanding should be necessary for WBE to be viable. For example, Sandberg writes, "For the current paper we will focus on simulations that attempt to achieve full functional equivalence – all relevant behavioral properties and internal causal links of the original system are replicated" (Sandberg 2013, p253). Part of this project entails modeling the interactions of "neurons and brain systems, and the emergent dynamics between them" (ibid., p252). But what are the *brain systems* referred to here? If they are merely wiring diagrams

[18] Token Physicalism itself might be too strong a thesis for some—many non-dualists, and most property dualists—seem to reject it. The a priori feasibility of WBE only needs a weaker thesis, something such as the mental supervening on the physical. Nevertheless, assuming token physicalism will not much matter above helping explication (and anyhow, will make the WBE proponent's case even stronger). For more on token Physicalism see Stoljar, J., "*Physicalism*", The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), Edward N. Zalta (ed.), URL = < https://plato.stanford.edu/entries/physicalism/>.

between neurons, then although mapping out the connections is an extremely difficult engineering task, it nonetheless is one that seems feasible. However, if more than the connectome matters, if instead lower-level, finer grained details, such as ones that involve neurochemical elements, or other substances that correspond to our "hardware" are germane, then the road to emulation is much less clear. The problem in front of us is to identify whether there are relevant aspects of cognition broadly construed (phenomenology, intentionality, intelligence, and personality, at a minimum) that are not merely dictated by the connectome.

Would the connectome suffice for replicating functional competence? To answer affirmatively is to presuppose a version of machine Functionalism[19] as well as the Multiple Realizability thesis,[20] the idea that psychological properties can be realized from a wide array of structural properties.[21] Both views are closely related: Machine Functionalism dictates that all essential properties of the mind are functional (and not structural) properties;[22] that is, it assumes

---

[19] Putnam, H. 1975. The Nature of Mental States. In *Mind, Language, and Reality*. Cambridge: Cambridge University Press.

[20] Fodor, J. 1974. Special Sciences (or: The Disunity of Science as a Working Hypothesis). *Synthese* 28(2): 97–115.

[21] This is close to right, though there is a bit of slippage. The former thesis—whether the connectome would suffice for replicating functional competence—is about behavioral competence, whereas the latter—Machine Functionalism—is about the essence of the mind. All of those who answer the latter question affirmatively will do the same for the former, but some who answer the former affirmatively may be silent on the latter.

[22] Which properties count as structural—say, the fusiform gyrus, or an electron—depends on one's explanatory ends. Properties that look structural from one vantage point (e.g., the prefrontal cortex

that (e.g.,) to be a belief is to just be a mental state that serves a certain role, the role that belief generally serves.[23] This function is the essence of the mental state. As such, there is nothing in the essence of (e.g.,) belief that makes it seem as if it had to be realized by a particular substrate. In which case perhaps there could be intelligent creatures that had silicon "brains." If these creatures had beliefs, this would prove that belief is multiply realizable, since it could be realized in brains like ours or in heads filled with silicon.

Problems for Whole Brain Emulation

Let us start with a seemingly pressing, though relatively easy, problem for the Multiple Realizability thesis: the embodied mind and extended mind theses. These theories hold that our minds extend beyond on our skulls, and not just because of (e.g.,) an externalist semantics that dictates that content is not only in the head. Proponents of the embodied mind posit that the body is integral to the functioning of the mind. Locomotion is interpreted as a central cognitive function, not one that is just useful for aiding in cognitive development but instead partially constitutive of cognition itself. Similarly, extended mind theorists claim that objects outside of one's body entirely—say one's cellphone—count as partially constitutive of one's cognitive apparatus.[24]

---

from the standpoint of intentional psychology) look functional from another (e.g., the prefrontal cortex from the standpoint of biochemistry). See Lycan, W. 1987. *Consciousness*, Cambridge: MIT Press.

[23] Ironically, some of the biggest proponents of the Connectomics also hold that neural structure and function are closely related, resulting in a rather precarious dialectic position (see, e.g., Chen, B., Hall, D., and Chklovskii, D. 2006. Wiring Optimization Can Relate Neuronal Structure and Function. *Proceedings of the National Academy of Sciences* 103(12): 4723–4728).

[24] Clark, A., and Chalmers, D. 1998. The Extended Mind. *Analysis* 58(1): 7–19.

Both the embodied and extended mind theses seem in tension with the Multiple Realizability thesis, which presupposes that one can just upload the cognitive software into any number of hardware realizers. But if embodied and extended cognition theorists are right, then there are real restrictions on the types of programs one could be uploaded into—for instance, a brain in a vat would not suffice for cognition.

Nonetheless, the embodied mind is not a deep obstacle for the feasibility of WBE. For one thing, the restrictions that would apply are, in the scheme of things, relatively trivial—they are not restrictions on the type of hardware that would be needed for uploads, but are instead restrictions on the type of environments the hardware would have to be embedded in. Adding the analogs of perceptual inputs and motor outputs, as well as some objects to interact with, is far less challenging than successfully reproducing an entire functional copy of a brain. In the case where we are envisioning that we can already do the latter, the former should be a small roadblock at best.

However, there is a more serious problem lurking, one that questions the scope of Functionalism in its entirety. Functionalism about mental states—beliefs, desires, hope, and the like—seems appealing because the functional role that each state plays seems essential to its character. For example, if you found a state that was not caused by perception, did not interact with motivational states to produce behavior, and did not serve as premises in inferences then it would be hard to see how it could count as a belief.[25] Though that functionalist intuition is reasonable enough, extending it to other states—say, phenomenological and motivational states—is a much more

[25] Quilty-Dunn, J., and Mandelbaum, E. 2018. Against Dispositionalism: Belief in Cognitive Science. *Philosophical Studies* 175(9): 2353–2372.

tenuous proposition. Beliefs seem functional, but e.g., experiencing something as green seems less so.[26]

Thus we again face the question: what level of granularity is necessary in order for WBE? What properties are the relevant ones that need to be recreated in order for emulation to be successful? One's take on whether the connectome will recapitulate cognition writ large will depend on one's theory of consciousness, even among Physicalists.

The theories of consciousness that matter for evaluating the prospects of WBE are theories of what makes a state phenomenally conscious. There are three major ones: 1) the Higher Order Thought theory (HOT); 2) the Global Workspace Theory; and 3) the Biological Theory.

All these theories give different answers to the question of what makes a state conscious. The HOT theory states that a first-order mental state becomes phenomenally conscious when a higher-order state takes the first state as its content.[27] In this way, consciousness is reduced to thoughts about mental states: a thought becomes conscious when we have another thought about it; a feeling becomes conscious when we have a thought about that feeling; and so on. HOT theory is friendly to WBE because the essential relation posited—that of a thought monitoring another mental state—is a functional one.[28] Monitoring mentions nothing of implementation or machinery

---

[26] See also Maudlin, T. (1989). Computation and consciousness. *The Journal of Philosophy*, *86*(8), 407-432.

[27] Rosenthal, D. 2005. *Consciousness and Mind*. New York: Oxford University Press.

[28] One could, in theory, be a HOT theorist but not be a functionalist (if say one rejected Functionalism about intentionality and had a non-functional specification of monitoring). That said, I cannot think of any non-functionalist Hot theorists.

at all, so should be able to be instantiated in many different ways, thus amenable to Multiple Realizability.

Global Workspace theory posits that any state that is "globally broadcast" is ipso facto phenomenally conscious. To be globally broadcast is a dispositional property: it is to be a state that is ready to be utilized by a varied array of other mental processes, such as reasoning, linguistic, and motoric processes.[29] Popular versions of global broadcasting posit competitive neural networks where sensory and frontal areas compete for resources, with the winner becoming conscious.[30] But none of the neural details, not even the use of neural nets, are essential to the view; instead they are just one way to flesh the view out. At its core, the Global Workspace theory is a functionalist view: it hypothesizes that to be phenomenally conscious is just to be a representation that is available to a host of consuming mental mechanisms (e.g., language production). In us, such consumption may involve details about competition between sensory and frontal cortices, however, that is a detail (most likely) about consciousness in us and not phenomenal consciousness *simpliciter*. Thus, Global Workspace theory allows that if we uploaded our connectome, we could replicate consciousness in a nonbiological substrate. In other words, Global Workspace, like HOT, is an essentially functionalist theory, one compatible with the Multiple Realizability thesis and WBE.

However, the third theory of consciousness—the Biological Theory—is where deep problems for WBE arise. The Biological Theory posits that the coding and interchange of

---

[29] Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

[30] Dehaene, S., Changeux, J., Nacchache, L., Sackur, J., and Sergent, C. 2006. Conscious, Preconscious, and Subliminal Processing: A Testable Taxonomy. *Trends in Cognitive Science* 10(5): 204–211.

information between electrical and chemical formats gives rise to consciousness, and that the specific neural hardware we use is essential to phenomenal consciousness.[31]

Some prominent arguments for the Biological Theory come from Ned Block: one argument relies on the explanatory gap, and the other perceptual overflow. The explanatory gap is the thesis that we do not have any idea of how a subjective state (such as seeing red, or hearing middle C on a piano) could be identical to an objective state (such as having a certain pattern of neuronal activation).[32] The thesis does not claim that in fact humans cannot in principle explain how objective states could give rise to subjective states. Instead, it is a theory about our current epistemic position, one which claims that at this moment we have no clue how psychophysical identities could be true.[33] The idea is that we do not yet possess the concepts to bridge this gap (although one day we may).

The existence of the explanatory gap is fairly untendentious, though the morals one should draw from it are more controversial. The Biological Theory takes the existence of the explanatory gap as support, as neither the HOT nor the Global Workspace view can explain why, if consciousness if a functional property, we should have an Explanatory Gap and the subsequent Hard Problem (Block 2009).[34]

---

[31] Block, N. 2009. Comparing the Major Theories of Consciousness. In M. Gazzaniga, ed., *The Cognitive Neurosciences*, 1111–1122. Cambridge, Mass.: MIT Press.

[32] Levine, J. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64: 354–361.

[33] Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(4), 354-361.

[34] I do not quite see how the explanatory gap is supposed to help the Biological Theory here, as it seems to also fall prey to the gap.

Another argument Block puts forward in favor of the Biological Theory is that it is the only view that can explain phenomenal overflow. "Phenomenal overflow" describes situations where one's phenomenal consciousness—generally in perceptual situations—overflows cognitive access. Perception and phenomenal consciousness more generally seem *richer* than what cognition can conceptualize. Parade examples use the logic of a partial report paradigm.[35] Subjects see an arrangement of letters (e.g., three rows of four letters) for a brief period of time. The letters then disappear and the subjects are cued to one of the rows. Subjects can report three or four letters from any cued row. But if subjects are asked to report as many letters as possible without any cue, they can still report only three or four letters. That is, subjects appear to consciously see all of the letters during the presentation but can only consciously access three or four total letters from the twelve-letter array. The rest of the letters are consciously perceived—they add to one's phenomenology—but they are not consciously accessed.[36]

Theorists like Block (2011) use overflow to argue for the Biological Theory.[37] They argue that any functional view of consciousness, such as HOT or Global Workspace, would place the unseen letters in the appropriate functional role as dictated by those theories. Take Global

---

[35] Sperling, G. 1960. The Information Available in Brief Visual Presentations. *Psychological Monographs* 74(11): 1–29.

[36] For competing takes on overflow see Phillips, I. 2016. No Watershed for Overflow. *Philosophical Psychology* 29(2): 236–249 and Gross, S., and Flombaum, J. 2017. Does Perceptual Consciousness Overflow Cognitive Access? The Challenge from Hierarchical Processes. *Mind and Language* 32(3): 358–391.

[37] Block, N. 2011. Perceptual Consciousness Overflows Cognitive Access. *Trends in Cognitive Science* 15(12): 567–575.

Workspace theory: the letters are originally conscious (because they add to phenomenology); since they are conscious, they should be reportable because, by hypothesis, to be conscious is just to be available in the workspace, which entails being available to report. But the letters are not reportable even though they are conscious, thus, Block reasons, Global Workspace must be wrong.

The Biological Theory of consciousness is the only non-functionalist of the theories canvassed, and as such it can explain the richness of perception and experience by interpreting that richness as overflowing access. What makes a state conscious is not its dispositional properties (e.g., being available to report or being the content of another thought) but merely the state being caused (or realized) by the specific biological machinery we have.

The Biological Theory also finds support outside of any of the overflow arguments. The connectome is the level of grain that most theorists find plausible for positing as the functional basis of the mind.[38] But the connectome is just an anatomical wiring diagram—even electrical connections between neurons are left out.[39] A fortiori Connectomics is committed to the view that a sub-neuronal difference should not lead to a functional difference. But sub-neuronal differences do appear to lead to psychological differences. What causes the vast individual differences in phenomenology is extremely unclear at the moment. But the contribution of sub-neuronal properties is integral in a way that is rarely appreciated in the literature. Serotonin, dopamine, norepinephrine, histamine, and countless neuropeptides are not accounted for in the connectome; they count as part of the 'hardware' of our system. These neurochemical properties act as neuromodulators, affecting neuronal connections in fundamental ways, even changing basic

[38] Seung, ibid.

[39] Morgan, J., and Lichtman, J. 2013. Why Not Connectomics? *Nature Methods* 10(6): 494–500.

neuronal functions.[40] In the connectome of *C Elegans*—a vastly easier connectome to understand than the human one—every neuron and synapse was subject to neuromodulation (ibid.).[41] The effect of neuromodulation is enormous in all nervous systems[42]:

> Modulators can qualitatively alter the neuron's intrinsic properties, transforming neurons from tonic spiking to those generating plateau potentials or bursts. The effect of neuromodulators can activate or silence an entire circuit, change its frequency, and/or the phase relationships of the motor patterns generated.[43]

And again, this holds in much simpler creatures than human beings (e.g., in worms); it is reasonable to suppose that in the more baroque case of the human brain, neuromodulators (to say nothing of glial cells) take on an even greater role. After all, we depend on intervening on neuromodulators to change affective and motivational states—serotonin reuptake pharmaceuticals are not targeting neuronal connections, but neurochemicals. To put it mildly, it seems implausible that every neuron can have its basic function changed by its instantiation base yet also hold that the instantiation base would have no effect on any cognitive property.

Moreover, we have good evidence that some sub-connectomic properties do matter for psychology. For instance, steroids from the adrenal cortex, as well as from sex organs, are not

---

[40] Bargmann, C. I., and Marder, E. 2013. From the Connectome to Brain Function. *Nature Methods* 10(6): 483–490.

[41] *C Elegans* connectome is perhaps not the most favorable piece of evidence for WBE enthusiasts. It was mapped in 1986 and yet we still have little idea what function any of its neuronal connections subserve, even though it only has ~300 neurons as opposed to our 100 billion or so neurons.

[42] Marder, E. (2012). Neuromodulation of neuronal circuits: back to the future. *Neuron*, *76*(1), 1-11.

[43] Bargmann, C. I., and Marder, E. 2013, p486.

captured by the connectome.[44] But increases in (e.g.,) testosterone plainly do affect a wide range of behavior, such as testosterone's ability to predict aggression (cortisol and serotonin do too).[45] Even some of Connectomics biggest proponents seem to see this problem, though perhaps not the consequences of it: "The ability of pharmacological agents to rapidly induce sleep, tranquility, excitement, hallucinations and so on means that the behavioral state can be dramatically altered probably without any modification to the connectome."[46] Of course, to be excited or tranquil is to be in a particular psychological state.

This is not to say that the Biological Theory is true. In consciousness studies—as elsewhere in science—ruling out false theories is the goal, whereas finding true theories is a bit idealistic. Perhaps the best response for functionalists is to become subneural functionalists, where the properties that matter for functional realization are lower-level than neuronal—perhaps much lower level (e.g., perhaps biochemical, or perhaps subatomic). This would be an interesting discovery—the idea that neural properties are not the functional realizers of the mind is, at the very least, very surprising. Moreover, the resulting dilemma itself, that one is forced to be a Biological Theorist or a subneural functionalist, is an interesting enough endpoint.

But becoming a subneural functionalist is also rather destructive to the idea that WBE is the best chance to achieve Superintelligence or immortality. Subneural functionalism contravenes the "nonorganicism" that allow futurists to champion WBE in the first place. The reason WBE is so

---

[44] Morgan and Lichtman 2013, p496.

[45] Montoya, E., Terbrug, D., Bos, P., and van Honk, J. 2012. Testosterone, Cortisol, and Serotonin as Key Regulators of Social Aggression: A Review and Theoretical Perspective. *Motivation and Emotion* 36(1): 65–73.

[46] Morgan and Lichtman 2013, p497

appealing to transhumanists, futurists, and the like is that it seems much less far-fetched than all the other routes to posthuman intelligence. WBE is supposed to be a data saver; it supposes that all we need to do is upload the functional properties, so we do not need to know how the whole brain works (or the whole body works, or the whole species works, or the whole universe works). However, the more low-level the functional properties are, the more we will need to know (and the more information we would need to upload), meaning we would be much further away from achieving uploading than even skeptics might assume. If the relevant level of detail demands molecule-for-molecule duplication, then WBE looks to be entirely unfeasible as an engineering project in even the medium-to-far term (and possibly computationally intractable).

So, if subneural functionalism is true, then the viability of WBE is in trouble. But we can go further still, for if the Biological Theory is true, much deeper theoretical revisions will be needed. If the Biological Theory is true, Multiple Realizability, Computationalism, and even Functionalism cannot be true of the entire mind. These theories may be true of propositional attitudes, or some other aspect of cognition, but they are not true of consciousness, in which case the mere possibility of machine consciousness and WBE is imperiled. This moral has not been lost on the proponents of the Biological Theory, such as Ned Block: "The Biological Theory says that only machines that have the right biology can have consciousness, and in that sense the biological account is less friendly to machine consciousness" (Block 2009, p1119). Of course, we are not in a position to say that the Biological Theory is true. But it's enough to note that it is, at this time, still very much alive, one of the very few live theories we have of consciousness, even if it is extremely underspecified.

Should Whole Brain Emulator Optimists Care about Consciousness?

In discussing the viability of WBE, Sandberg opined that "there doesn't seem to be any convincing knock-down arguments within the philosophy of mind against WBE" (Sandberg 2013, p261).

Although there is not a knock-down argument against it, there is reason to have serious skepticism about WBE's viability and this, in turn, reveals some deeper problems in the metaphysics of mind.

Before concluding, let us take a step back to consider the big picture: what did we want WBE for anyway? Only two ends have been put forth. The first is as a step towards achieving superintelligence, and the second is for achieving immortality. I take these in turn.

As a reminder, the route to superintelligence went through using WBE to upload human-level intelligence. Once we have a cheap and easy way to produce and store human intelligence, we can create an enormous amount of uploads and then put them to the task of discovering the breakthroughs that can lead to superintelligence.

How much would consciousness matter for this program? Say the Biological Theory is only true for phenomenal consciousness. Could the rest of cognition then be captured by the connectome, in which case WBE could still lead to superintelligence? The question turns, in part, on whether there can be intentionality without phenomenology. Having some unconscious intentional states—like beliefs—is a commonly enough held position.[47] But could there also motivation, or desire, without any phenomenology? That seems much less clear. What it is to desire something seems to involve feeling a certain way. Likewise, what it is to be motivated has an aversive quality to it, which is just to say that some motivations appear to have some phenomenology.

[47] Mandelbaum, E. 2014. Thinking Is Believing. *Inquiry* 57(1): 55–96. For a defense of phenomenal intentionality see Kriegel, U. 2013. The Phenomenal Intentionality Research Program. In U. Kriegel (ed.), *Phenomenal Intentionality* (pp. 1-26). OUP, 2013.

If we want uploads to do anything, they will have to be motivated.[48] Cognition without conation is just a spinning wheel connected to nothing. Having a billion more human-level intellects available to work on a problem will only help solve the problem if they are designed to solve the problem or motivated to do the work. Part of the appeal of uploads is that we wouldn't have to design any particular goal for them, for doing so takes us far beyond merely uploading a connectome. Since we won't be able to design uploads with the goal of solving any particular problem, uploads will only act if they are intrinsically motivated to. If they have no motivations, then they will not do anything on their own.

The problems for WBE get even worse. Many theories of the attitudes dictate that to have any beliefs at all, one must have other propositional attitudes, particularly desires and motivations.[49] If there are no desires, then uploads may not even have beliefs, for, so the thought goes, part of the functional role that is constitutive of belief is that they interact with desires to cause action. If uploads do not have beliefs, it is hard to see how they could ever engage in thinking as they would lack the premises of thoughts (and the desires to go through the bother of transitioning from thought to thought).[50]

---

[48] One may argue that cars and calculators do things without being motivated, but they do so at the behest of intelligent, motivated designers and users. Even Bostrom's paperclip maximizer has to be seen as either having the motivation to turn everything into paperclips or had been given the function to do so. Bostrom, N. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

[49] Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge: MIT Press.

[50] Quilty-Dunn, J., & Mandelbaum, E. 2018. Inferential transitions. *Australasian Journal of Philosophy*, *96*(3), 532-547.

There is an even more exotic argument against the existence of beliefs that are totally disconnected to phenomenology. It starts by noting that our beliefs matter to us. When we encounter disconfirming information it *hurts* and immediately causes us to readjust our beliefs, often perversely increasing credence in the proposition under attack.[51] Some hold that this is a defining feature of belief, so that any state that did not act this way would not be a belief.[52] If this is right, then if the connectome did not include valences, uploads could not have beliefs. And this argument generalizes for any mental state where valence plays a constitutive role.

If uploads lacked beliefs and desires, then they would just be giant calculators that we neither know how to control nor understand the mechanics of. Recall that the appeal of WBE was its nonorganicism, which allows that we could copy the brain without the need to understand how all of it works—this is what was supposed to move up the timetable of feasibility for WBE versus any other technologies. Then once we had the uploads, we could reason with them the way we would with any belief/desire-based agent. But if uploads don't have the normal attitudes, we will have no idea how motivate them to do anything—it's not even clear that they would be able to be motivated. In that case, we'd have to go back to a more fine-grained stance to affect their behavior, which would demand another conceptual breakthrough.

WBE's promise for immortality raises even murkier questions. We generally think the issue of immortality and uploads boils down to the question of whether uploading your mind without consciousness would suffice for immortality. But even smaller questions about consciousness fester:

---

[51] Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language*, *34*(2), 141-157.

[52] Porot, N., & Mandelbaum, E. (2021). The science of belief: A progress report. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(2), e1539.

might one's particular type of phenomenology matter for capturing identity? Does one's character intimately involve the kind of phenomenology they have? Maybe you could be you even with a different character. This is not totally implausible—people can change their personality throughout their lifespan (though whether that actually makes a change in personhood is tendentious).[53] Yet some of the properties that seem deeply central to our self-conception would be left out. Above we noted that tranquility, excitement, and the like will be left out of the connectome, but these properties are plainly not just properties at the edges of our identity but instead are often integral to who we are. People think of themselves as, e.g., deeply energetic, or extremely calm and patient. But those personality traits would be left out of the connectome. Could your connectome duplicate you even if it was (e.g.,) a sickly sloth while you are a dynamo bursting at the seams with energy and ideas?

Even without taking a stand on what exactly personal identity amounts to, it appears that what it is like to be you does have some bearing on what it is to be you. And if that is the case, then the biggest roadblock to the grandiose promise of WBE—uploads—is that our biological machinery itself may be responsible for a good deal of our cognitive life. The problem is not just that you (e.g.,) see deep purple whereas the upload version of you would experience periwinkle. It is that to exist as you would involve some of the full panoply of emotions, feelings, depths, and depravities of everyday life, and these would be left out of the uploads.

This doesn't mean that we should endorse Mysterianism or be sure that uploading is necessarily impossible. The world never ceases to surprise. Perhaps one day we will be able to upload full wiring diagrams into hardware just like ours. But if so, that would no longer be emulating

[53] Strohminger, N., and Nichols, S. 2014. The Essential Moral Self. *Cognition* 131(1): 159–171.

whole brains, but cloning and recreating them from scratch, in which case the feasibility of achieving it should seem that much further off than current futurists prognosticate.[54]

[54] Helpful comments and discussion were received from Ron Avni, Adam Bradley, David Chalmers, Tim Crane, Cian Dorr, Jackson Kernion, Jessica Moss, Jake Quilty-Dunn, Shen Pan, Kate Ritchie, David Rosenthal, Fiona Schick, Henry Schiller, and David Udell. They are all hereby thanked for their collegiality, patience, and friendship.