HOW TO PROPERLY WEAR A TIN FOIL HAT:
A CALL FOR EPISTEMIC HUMILITY IN THE CREATION OF ARTIFICIAL INTELLIGENCE, IN APPLICATIONS OF NEUROPROSTHETICS, AND IN THE DEBATE OVER SCIENTIFIC REALISM

by

Matthew Schuler

_____
Copyright © Matthew Schuler 2018

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHILOSOPHY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by **Matthew Schuler**, titled ***How to Properly Wear a Tin Foil Hat: A Call for Epistemic Humility in the Creation of Artificial Intelligence, In Applications of Neuroprosthetics, and In the Debate over Scientific Realism*** and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Terry Horgan _____ Date: 9-26-2018

Terry Horgan

Stewart Cohen _____ Date: 9-26-2018

Stewart Cohen

Shaun Nichols _____ Date: 9-26-2018

Shaun Nichols

Joseph Tolliver _____ Date: 9-26-2018

Joseph Tolliver

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Terry Horgan _____ Date: 9-26-2018

Dissertation Director: Terry Horgan

ARIZONA ®

ACKNOWLEDGEMENTS

First and foremost I wish to convey to Terry Horgan, my dissertation advisor, my gratitude for his work in helping me to improve the papers that eventually became chapters comprising this dissertation; his generosity and truly ingenious feedback – which I benefited from in our meetings for the last several years as part of his *Work In Progress* group – resulted in this dissertation (whatever merits it *does*, in fact, have) being significantly better than it otherwise would have been. (Here, as with everyone else acknowledged below, I feel compelled to point out that all shortcomings are – obviously – mine and mine alone; in some cases, I was simply unable to implement revisions, recommended by Terry and others, that would have adequately handled the preponderance of objections the reader may be inclined to raise – or so I'd wager.)

I am also deeply appreciative of the time, effort, and contributions of my other committee members – Stewart Cohen, Shaun Nichols, and Joseph Tolliver. Interestingly, each of these philosophers has contributed to the improvement of this dissertation in entirely different ways, and I would like to take the time to acknowledge how they did so.

Over the course of my graduate studies, I have probably taken more courses and participated in more Independent Study & Reading Groups with Stew than anyone else. And so although we *generally* work in different areas, nonetheless (especially during the first few years of the program), it was Stew (along with Juan Comesaña) who were most responsible for shaping my approach to the discipline as I began to mature philosophically.

As for Shaun, I not only learned a great deal simply by being his TA on multiple occasions, but more importantly, it was Shaun who directed me to the work of John Bickle; this mad the dissertation far more interesting than it otherwise would have been (to the extent that the reader does indeed find it interesting to some degree).

And finally, with regard to Joseph, I spent so much time with him (both in the classroom and outside of it) that I like to *think* that some of his incredibly rarely-encountered philosophical creativity has rubbed off on me; in any case, he is the philosopher who exposed me to Ted Sider, and helped me to make sense of *Writing the Book of the World* (though he

may disagree with the position I take with respect to Sider). Whatever grasp I *do* have of Sider's metametaphysics, I owe to Joseph, and for that I am most grateful.

I also owe an especially significant debt to the participants of Terry's *Work in Progress* group, where I had the opportunity to present each of these papers several times and receive crucial feedback from its rotating members: included are not only Terry himself, but also Brandon Ashby, Oisín Deery, Matt DeStefano, Martina Fürst, Yael Loewenstein, Tyler Millhouse, Lucia Schwarz, Vojko Strahovnik, and especially Bryan Chambliss.

I would also like to thank my doctoral student colleagues with whom I have had discussions about this material. Phoebe Chan, Tyler Millhouse, and Eyal Tal immediately come to mind, but it is inevitable that I shall have just now inadvertently omitted colleagues whose help was especially valuable. I hope such colleagues will accept my sincere apology.

I would also like to thank my M.A. pals from Virginia Tech, Andy Creighton and Drew Valdespino, since this is when I first started thinking about the philosophy of science. And it was Daniel Parker, the Professor of my first philosophy of science course, who encouraged me to continue pursuing that work. But of even greater significance is the similar encouragement I received from the University of Arizona's Richard Healey, who offered me crucial feedback on a paper that made its way into Chapter 3.

I am also grateful to an audience at the 2016 Medical Humanities Conference at Western Michigan University, where I presented a paper that eventually became a portion of Chapter 1 of this project, and received challenging remarks and helpful feedback. Thanks also to William FitzPatrick, James Klagge, Douglas Langston, and Sandra Kimball.

Importantly, I owe a special debt of gratitude to Yael Loewenstein – a debt I am unlikely to ever be able to repay. She is without a doubt the most talented *student* of philosophy (although now a professor!) I have ever personally known – and she has already gone on to do great things within the discipline, which fills me with pride. Frankly, our roughly six-year relationship while M.A. students at Virginia Tech and then Ph.D. students at the University of Arizona made me an immeasurably better philosopher than I otherwise could have been.

Finally, I wish to thank my parents, who were there for me in ways I never could have anticipated needing (toward the end of writing this dissertation). I will be forever grateful.

DEDICATION

For my mother and father, Brenda Cardella and Michael Schuler, and my step-mother and step-father, Donna Schuler and Ron Cardella, who truly were instrumental in the completion of this dissertation.

TABLE OF CONTENTS

ABSTRACT

This dissertation consists of three (more or less) freestanding articles. The first two chapters are intended, nonetheless, to work in tandem to a significant degree. The third (and final) chapter is the most *freestanding* of the three, but it does have direct ties to the first two chapters; however, it takes a broader view of the philosophy of science and technology, and furthermore attempts to extend the results already obtained to metaphysics (actually, metametaphysics, to put it more accurately).

In the first article, I discuss the philosophical import of recent scientific research on hippocampal prosthesis (HP), focusing on the handful of implications that seem to have the most direct philosophical relevance; of these, the one *most* fascinating is what the possibility of HP appears to be able to tell us about artificial intelligence. But because the principal (intended) scientific application of HP is its use in the treatment of Alzheimer's Disease, I also devote considerable space to the ethics of using HP in this way.

In the second article, I once again discuss HP, but here my focus is solely on using its advent as the basis for an inductive inference meant to establish the possibility of artificial intelligence. In fact, I believe it is possible to simply give a *recipe* for its creation. However, here too the second half of the article is devoted to the ethics of the practice in question – in this case, the use of Neuroprosthetics to create deep, genuine artificial intelligence.

In the final article I develop a novel argument against scientific realism by exploiting one of the realist's own, most fundamental, commitments: naturalism. I then show that this form of argument can be applied to cast serious doubt on the plausibility of Ted Sider's metaphysics.

INTRODUCTION

As an undergraduate (and as a complete philosophical novice) I once asked the philosophy department's logician (of all people) what he thought of the work of Nietzsche. His response was something like, "I consider it reasonably entertaining bedtime reading, but that's about it." It turns out that this pretty accurately describes my own attitude toward Nietzsche's work – except that I don't even view it as worth my time as bedtime reading.

As such, I shall not be discussing Nietzsche at any point in this dissertation – *with one exception*. To wit: there is a passage, from a relatively obscure essay, entitled "On Truth and Lie in an Extra-Moral Sense", that struck me many years ago as almost certainly correct. In some ways (when combined with some basic Kantian metaphysical assumptions I have been unable to let go of – which I shall also go out of my way to avoid discussing here), the core idea in that passage became a guiding philosophical principle for me. So I shall make one exception and quote Nietzsche at considerable length[1]; it is difficult to tell what the reader will think of this passage, but for some reason it seems to be the most appropriate way of kicking this project off (although perhaps an even less clear – though probably also less poetic! – statement by Kant would have also done the trick) – or, that is, of conveying the basic, underlying philosophical mindset that guides much of the following work (*especially* Chapter 3):

> In some remote corner of the universe, poured out and glittering in innumerable solar systems, there once was a star on which clever animals invented knowledge. That was the haughtiest and most mendacious minute of "world history" – yet only a minute. After nature had drawn a few breaths the star grew cold, and the clever animals had to die.
>
> One might invent such a fable and still not have illustrated sufficiently how wretched, how shadowy and flighty, how aimless and arbitrary, the human intellect appears

[1] Although the passage cited is quite lengthy (at least for a dissertation introduction), the essay from which it has been extracted is in fact quite short.

in nature. There have been eternities when it did not exist; and when it is done for again, nothing will have happened. For this intellect has no further mission that would lead beyond human life. It is human, rather, and only its owner and producer gives it such importance, *as if the world pivoted around it. But if we could communicate with the mosquito, then we would learn that it floats through the air with the same self-importance, feeling within itself the flying center of the world* [emphasis added].

There is nothing in nature so despicable or insignificant that it cannot immediately be blown up like a bag by a slight breath of this power of knowledge; and just as every porter wants an admirer, the proudest human being, the philosopher, thinks that he sees the eyes of the universe telescopically focused from all sides on his actions and thoughts…

That haughtiness which goes with knowledge and feeling, which shrouds the eyes and senses of man in a blinding fog, therefore deceives him about the value of existence by carrying in itself *the most flattering evaluation of knowledge itself* [emphasis added]…

And, moreover, what about these conventions of language? Are they really the products of knowledge, of the sense of truth? Do the designations and the things coincide? Is language the adequate expression of all realities? …The different languages, set side by side, show that what matters with words is never the truth, never an adequate expression; else there would not be so many languages. *The "thing in itself" (for that is what pure truth…would be) is quite incomprehensible to creators of language and not at all worth aiming for* [emphasis added]**…**

Let us still give special consideration to the formation of concepts. Every word immediately becomes a concept, inasmuch as it is not intended to serve as a reminder of the unique and wholly individualized original experience to which it owes its birth, but must at the same time fit innumerable, more or less similar cases – which means, strictly speaking, never equal – in other words, a lot of unequal cases. Every concept originates through our equating what is unequal. No leaf every wholly equals another, and the concept "leaf" is formed through an arbitrary abstraction from these individual differences, through forgetting the distinctions; and now it gives rise to the idea that in nature there might be something besides the leaves which would be "leaf" – some kind of original form after which all leaves have been woven, marked, copies, colored, curled, and painted, but by unskilled hands, so that no copy turned out to be a correct, reliable, and faithful image of the original form. We call a person "honest." Why did he act so honestly today? we ask. Our answer usually sounds like this: because of his honesty. Honesty! That is to say again: the leaf is the

> cause of the leaves. After all, we know nothing of an essence-like quality named "honesty"; we know only numerous individualized, and thus unequal actions, which we equate by omitting the unequal and by then calling them honest actions. In the end, we distill from them a *qualitas occulta* with the name of "honesty" ...
>
> What, then, is truth? A mobile army of metaphors, metonyms, and anthropomorphisms – in short, a sum of human relations, which have been enhanced, transposed, and embellished poetically and rhetorically, and which after long use seem firm, canonical, and obligatory to a people: truths are illusions about which one has forgotten that this is what they are; metaphors which are worn out and without sensuous power; coins which have lost their pictures and now matter only as metal, no longer as coins..." (1982: 42–47).

Over the last ten years (while completing a terminal M.A. and now a Ph.D.), I have had absolutely no reason to return to anything Nietzsche ever wrote.[2] But it strikes me now that, although I don't agree with every claim in this rather lengthy passage, it nonetheless represents reasonably well my basic philosophical orientation, at least in the following sense (and when purged, insofar as possible, of its *poetry*): it seems to me that the impression it made upon me over a decade ago was, unbeknownst to me, significantly formative in shaping the way in which I thought (and continue to think) about (a) the place occupied by the human intellect within nature; (b) the relationship between language and the formation of concepts; and (c) the plausibility of a conceptual scheme that manages to latch onto the world *as it is in itself*. In effect, it planted a seed of pessimism about such topics that I have been unable to shake.

It seems to me that this introduction, *especially when set against the backdrop of the passage from Nietzsche above*, will furnish the opportunity to discuss and summarize each of the three chapters of this dissertation – and better yet, to do so while simultaneously establishing the thematic continuity between these chapters that one would hope to see.

Thus, although the epistemological implications of Nietzsche's essay are less than entirely perspicuous and rigorous, it strikes me that much of my philosophical work has been

---

[2] I suspect, at any rate, that the reader will agree that it lacks the rigor of mainstream analytic philosophy.

an (unwitting) attempt to systematize – in a (hopefully) more rigorous manner, befitting an analytic philosopher – the core insights of that brief essay partially cited above. I expect the reader will find that this appears especially true in the final chapter of this dissertation ("A Naturalistic Argument Against Scientific Realism and Implications of this Argument for Sider's Metaphysics"). But the careful reader may notice the influence of this (I suppose somewhat pessimistic) philosophical orientation in at least parts of the first and second chapters as well (respectively, "Philosophical Implications of Recent Scientific Work on Hippocampal Prosthesis and the Ethics of It as a Potential Future Treatment for Alzheimer's Disease", and "How to Create Artificial Intelligence – and yet the Moral Impermissibility of Taking Any Steps to Do So").

Briefly, here is how I see the connection between Nietzsche's remarks and the first and second chapters of the dissertation (and, therefore, what it is that unites those first two chapters with the third, where the latter is *obviously* inspired by, or at least continuous with, the passage quoted above – and inspired, for that matter, by (merely) elementary Kantian metaphysics, for better or for worse): having set aside the obvious connection to the third chapter, probably the best way of stating this is with reference to the second chapter (though the first chapter is itself indispensable in rendering plausible the view articulated and defended in the second chapter). Just as (in a more straightforward way) Nietzsche's core insight concerning the place of the human intellect within nature has informed (in the third chapter, most notably) my pessimistic view on the prospects of the success of science in latching onto reality as it is in itself, that same insight has, in the second chapter, prompted (what I contend is) a healthy skepticism about our ability to reliably predict what might happen if we were to succeed in bringing about the creation of (deep, genuine) AI. The reason for this is not far to seek: if the limitations on the human intellect (argued for in the third chapter) in relation to our ability to discern the underlying nature of reality are indeed legitimate, we should not expect to be able to fully grasp or antecedently predict the preference structure of an intellect properly classified as a superintelligence (as defined in the second chapter) – at least, not in the most important case (i.e., as it pertains to the interests of the human race).

The first chapter of this dissertation ("Philosophical Implications of Recent Scientific Work on Hippocampal Prosthesis and the Ethics of It as a Potential Future Treatment for Alzheimer's Disease") draws upon recent studies in neuroprosthesis (in which researchers have shown that it is possible, in nonhuman primates at least, to (*i*) mathematically model neuronal behavior in the hippocampus and to (*ii*) use this to create an artificial neural network via silicon chip (commonly referred to as *hippocampal prosthesis*) that can (*iii*) be implanted in subjects, bypassing the hippocampus altogether, and which (*iv*) is capable of replicating the neurophysiology of the hippocampus. (That is to say, as regards (*iv*), the technology enables storage of new memories to the chip and long-term retrieval of such chip-mediated memories.) Here, my principal aim is to tease out four specific philosophical implications of such findings, enumerated here in (what I take to be) ascending order of importance: implications for (*i*) theories of personal identity, (*ii*) a general theory of mind, (*iii*) the possibility of artificial intelligence, and (*iv*) issues at the cutting-edge of medical ethics.

The second chapter of the dissertation ("How to Create Artificial Intelligence – and yet the Moral Impermissibility of Taking Any Steps to Do So") begins with a recipe for the creation of deep, genuine AI. I offer this reluctantly because (a) I believe that, although it has hitherto been overlooked, it is very likely a sufficient method for creating AI (and perhaps also represents a necessary condition for its creation), but (b) I am entirely convinced that AI should not be created. Thus, as regards (b), I note that AI Doomsday worries are nothing new; nor is the idea that the possibility of existential threat redounding from the creation of AI possibly licenses some normative claim or other. But here, where I construct an argument for the moral impermissibility of taking actions toward the creation of AI, an effort is made to accomplish this without inviting the unnecessary controversy other writers sometimes seem bent on introducing. For instance, my argument attaches no particular importance to anything like Bostrom's "orthogonality thesis", "strategic advantage thesis", "instrumental convergence thesis" (et cetera), but instead relies upon weaker, less controversial claims. In particular, it attempts to exploit a guiding principle I have deployed elsewhere (and in the service of entirely unrelated questions) concerning the place the human intellect occupies within nature, together with a form of (fairly radical in its applications, but nonetheless well-supported) epistemic humility that plausibly falls out of that basic claim.

Finally, the third chapter of this dissertation ("A Naturalistic Argument Against Scientific Realism and Implications of this Argument for Sider's Metaphysics") has seen a vast number of iterations, as a result of having undergone countless revisions over the last several years. A general philosophical commitment I have (and one shared by the majority of philosophers) is *naturalism*, and the chapter is an effort to show that – contrary to the received view – scientific antirealism actually comports better with naturalism than does scientific realism. The chapter then closes with a fairly lengthy discussion of Ted Sider's metametaphysics, and what conclusions we can draw about his view, on the assumption that the first part of the third chapter was correct in its core claims.

PHILOSOPHICAL IMPLICATIONS OF RECENT SCIENTIFIC WORK ON HIPPOCAMPAL PROSTHESIS AND THE ETHICS OF IT AS A POTENTIAL FUTURE TREATMENT FOR ALZHEIMER'S DISEASE

Researchers have recently shown that it is possible, in nonhuman primates, to (*i*) mathematically model neuronal behavior in the hippocampus and to (*ii*) use this to create an artificial neural network via silicon chip (commonly referred to as *hippocampal prosthesis*) that can (*iii*) be implanted in subjects, bypassing the hippocampus altogether, and which (*iv*) is capable of replicating the neurophysiology of the hippocampus. (That is to say, as regards (*iv*), that the technology enables storage of new memories to the chip and long-term retrieval of such chip-mediated memories.) Here, my principal aim is to tease out four specific philosophical implications of such findings: implications for (*i*) theories of personal identity, (*ii*) a general theory of mind, (*iii*) the possibility of artificial intelligence, and (*iv*) issues at the cutting-edge of medical ethics. So far as I can tell, no one has yet pointed out, in any careful or systematic way, these important philosophical implications.

Section 1: Relevant Background and General Remarks

1.1 *Introduction*

In recent years, research in neuroprosthesis has made great strides. To cite just a few important findings: Ochsner, et al. (2015) have made advances in cochlear implants to repair hearing (which involves bypassing the ear altogether – i.e., the implant sends sound signals directly to the brain rather than merely amplifying sounds, as hearing aids do); Luo, et al. (2016) have utilized a silicon version of the cerebellum to demonstrate improved sensorimotor control and learning; and perhaps most peculiar of all, Hartmann, et al. (2016)

have shown that it is possible to endow creatures with *new sensory modalities* via neuroprosthesis. (In Hartmann et al.'s work, this involved implanting, in the somatosensory cortices of rodents, a chip designed to enable subjects to discriminate infrared light. Although their choice of sensory modality, for proof of concept, happened to be infrared light discrimination, we can imagine far more interesting, and not altogether implausible, human applications – such as the ability to design a chip endowing its human recipient with a sensory modality such as, e.g., echolocation[3].)

These developments are all quite fascinating. One not mentioned above, however, is particularly important for certain theories of philosophers and cognitive scientists – and one that is simultaneously potentially challenging for the medical community. As the reader may already be aware, Hampson, et al. (2013) have shown that it is possible – in not only rodents but also in nonhuman primates – to create an artificial hippocampus that can be surgically implanted such that the subject can write information back to the chip component of this hybrid brain – and to do so at the *individual neuron level* – bypassing the hippocampus altogether. It is this finding, and the philosophical implications thereof, that I wish to explore in this paper.

---

[3] Thus, someday Nagel may in fact know what it is like to be a bat!

(It is worth pointing out, however, that Brandon Ashby has, in conversation, disputed this characterization of what these results would mean for Nagel's original claim. I have to confess, however, that I never managed to get clear (both in conversation with Brandon and in reading his notes on my paper) on how the state of affairs I describe here would *not in fact* count as knowing what it was like to be a bat – at least insofar as Nagel (and we) are focusing *just* on the novel and hitherto inconceivable (for us) sensory modality of echolocation, which I take it Nagel was *in fact* doing, setting aside any other ways in which phenomenologically it is (currently) impossible for us to know what it is like to be a bat. Furthermore, if there are *other* phenomenological discrepancies between what it is like to be a human and what it is like to be a bat, what I have produced here suggests that those, too, are differences that could be overcome via specially-designed silicon chip, enabling us to *genuinely* know, as a matter of technological possibility, what it is like to be a bat.)

It is worth noting that – so far as I can tell, at any rate – even if there are philosophers familiar with this recent research on hippocampal prosthesis (HP), none has yet drawn attention in the literature to the connection between HP and the host of philosophical issues I discuss here. (In fact, although purely anecdotally, of the many accomplished philosophers with whom I have inquired, a very meager few are even *aware* of the advent of HP.) Hence, as far as I can surmise after a careful review of the literature, this paper may well be the first attempt to draw out the philosophical implications of these recent, and prima facie important, empirical findings.

1.2 *Sketching the Project*

The philosophical import of HP can be divided into (at least) four distinct domains. These are as follows. First, it appears that the very possibility of HP tells us something important about the nature of personal identity; I address this in Section 2. Second, the development of HP enables us to make some surprisingly strong claims about the nature of the mind; I discuss this in Section 3. Third, when the possibility of HP is combined with certain other highly plausible assumptions, important and novel implications for research in artificial intelligence (AI) appear to follow with only minimal controversy; this topic is taken up in Section 4. Fourth, developments in HP raise important ethical questions for the medical community. In particular, it seems likely that medical ethics may soon be forced to grapple with the implications of using HP as a form of treatment for Alzheimer's Disease (AD) and other neurodegenerative disorders. I close with an extended discussion of this matter in Section 5.

This paper is organized with a natural thought in mind: the philosophical implications of research on HP I discuss are enumerated here in (what I take to be) ascending order of importance and novelty. Thus, I begin (with the material covered in Section 2) with what seem to me the *least groundbreaking* or game-changing implications of research on HP, and proceed, with each subsequent section, to address implications I consider more significant than those of the preceding section. Hence, the content of Section 4 strikes me as more philosophically important than the content of Section 3, which in turn seems more significant (or at least more novel) than the content of Section 2.

The one exception here is (the majority of) Section 5. In my own view, what is discussed there is not necessarily *more important* than the material discussed in Section 2 through Section 4; rather, it is just that – for ease of exposition – it seems prudent to address the content of Section 5 after everything else that needs to be said about HP (and its philosophical implications) has already been presented. In fact, my own view is that it is the content of Section 4 (when coupled with just the content of Section 5.5 and its subsections) that is especially important and philosophically novel. (For one thing, when the content of Section 4 and Section 5.5 is combined with the argument I present in another paper comprising this dissertation – "How to Create Artificial Intelligence – and yet the Moral Impermissibility of Taking Any Steps to Do So" – particularly significant philosophical results appear to follow.)

In bringing this subsection to a close, I merely wish to flag the fact that all of the implications briefly mentioned above strike me, philosophically, as potentially *quite*

significant. It will, needless to say, be my burden to show that this is in fact so. But that, of course, is what I intend to do here.

1.3 *Recent Developments in HP*

Hampson and colleagues (2013) have shown that it is possible, in nonhuman primates (specifically, rhesus macaques) with impaired hippocampal function, to develop and deploy a neuroprosthesis that effectively repairs and enhances memory encoding, storage, and retrieval – and even facilitates *recovery* of memory lost due to disease[4]; they suggest, with overwhelming plausibility, I think, that such technology can, in principle, furnish similar results in humans burdened by hippocampal impairment (2013: 13). In fact, interestingly, they note (2013: 12) that "the extent and range of effectiveness in improving cognitive performance in [this nonhuman primate memory study] was much greater" than when this very same HP was deployed in earlier rodent studies (such as Berger et al., 2011 and Hampson et al., 2012). Since human neurophysiology is by significant measure more similar to non-human primate neurophysiology than it is to rodent neurophysiology, perhaps analogously we should expect to be able to say something similar about the effect we are likely to find in future human studies in relation to extant studies of non-human primates. That is to say, perhaps the comparison of efficacy in HP as it will be employed in humans as opposed to primates will bear a similar relation to the improvement in efficacy when moving from rodent trials to primate trials. Although quite plausible, this is nonetheless mostly

[4] In the interest of space, I won't delve into the details of the experimental design and methodology of this primate study (though it is fairly ingenious!); I merely refer the curious reader to Hampson et. al.'s (2013) paper cited here if she wishes to review the intricacies of their study.

speculative, and we will naturally have to wait until human trials are complete to say anything decisive or illuminating in this respect.

Here I pause to take a brief historical (though still recent) step backward to set the stage for what follows. Prior to Hampson and colleagues' recent work was a (2005) study by Berger et al. that laid the foundation for current work on HP. Berger and colleagues' key insight was to transition from conventional, artificial neural networks to what they call a "dynamic synapse neural network architecture." The latter, but not the former, can successfully model nonlinear and temporal signal-processing properties of neurons, in turn enabling each silicon neuron to "transmit a spatiotemporal output signal" (2005: 250). And one critical facet of this model is its built-in "dynamic learning algorithm" operating on each dynamic synapse.

Importantly, Berger and his colleagues also made the distinct and significant contribution of finding a way to embed this model of hippocampal neuronal behavior within a silicon chip. It is worth noting that in their (2005) paper, Berger et. al. register their uncertainty concerning just how many silicon neurons would be needed for an adequate HP, and estimate that it would number in the hundreds or perhaps thousands. At the time their paper was written (a full decade before Hampson and colleagues'), Berger and his colleagues had succeeded in creating a chip with a small number of silicon neurons and were aiming, in the future, for a chip capable of holding 400. (Not only is this figure, by today's standards, quite meager, but another limitation of their work is that, at that time, they had not yet addressed the task of actually integrating a microchip within the brain.)

What is significant from the perspective of biomedical technology, however, is that – with respect to the small number of silicon neurons they had managed to build on the chip at that time – Berger et al. made an important observation in defense of the potential therapeutic significance of their work:

> It is critical to distinguish between a functionality that significantly alleviates clinical symptoms and a functionality that reproduces the capabilities of an intact brain. (2005: 263)

Their aim, in other words, was not to produce a chip capable of replicating all functions of the hippocampus, but rather to construct one that would have a significant and positive therapeutic impact on patients with hippocampal damage.

Even so, in the intervening years, the technology has developed to such an extent that it would not be mere hyperbole to characterize the prosthesis we see in Hampson et al.'s (2013) primate study as a fully functioning artificial hippocampus. Needless to say, however, we won't have at our disposal all of the data necessary to confirm such a claim until human trials are complete, for the obvious reason that – although we can easily construct studies that accurately assess memory ability in primates – their linguistic limitations prohibit the sort of debriefing one would like in order to feel confident in saying that *there really is no broadly cognitive difference* (including, say, the subjective conscious experience, or phenomenology, of *remembering*) between a primate with HP and a primate with a healthy hippocampus. (There is some indication, however, that we won't have to wait long to find out; anecdotally, human trials are reportedly underway presently with epilepsy patients, though it is currently difficult to locate much information about these purported studies.)

Section 2: HP and What its Very Possibility Tells us about the Nature of Personal Identity

2.1 *Introduction*

Hampson and his colleagues have furnished philosophers with important empirical data that can be used to test theories on the nature of personal identity and its putative supervenience upon memory. Here is one rough and ready articulation of the familiar memory theory (MT):

> **MT**: Personal identity – in the sense of strict numerical identity, as applied to persons – is constituted by *memory*. (Somewhat more formally, we might say: *person S at time t+1 is the same person as person P at time t if and only if S has (say, most of) the same memories as P (and, in most instances, other new ones as well).*) Thus, if you erase all of *S*'s memories, although her body survives, *P* (the *person*) is gone.

How does HP enter into the picture here? In short, the development of HP provides overwhelming evidence that MT is false. Although the reason for this is not far to seek, in the following subsection, I explain why.

2.2 *An Argument Against MT*

We now know, due to the development of successful HP, that it is possible to encode and store memories via silicon. (It is true that the hippocampus is not where long-term memories are stored, but as many working in this field have noted, there is every reason to believe that other brain regions can be similarly mapped and modeled – in the same way that has already been done with the hippocampus – so that their functions can be replicated by a chip.)

Matt & Mark

> Suppose, then, that I, Matt, undergo a surgical procedure (say, not merely removal and prosthetic replacement of the hippocampus, but of the cortex as well) in order to have a series of chips doing *all* of my memory-encoding, memory-storage, and memory-retrieval work. I go back to the lab one week later for a follow-up visit and the researchers explain that they would like to anesthetize me, remove the chips briefly, and make copies of them. I agree. Meanwhile, they have prepared a second person – call him Mark – for surgery, in order to remove the very same brain regions they had, in my own case, removed and replaced with a set of chips. Once the researchers have made their copies, they simultaneously implant the originals back in my brain and the copies in Mark's. For simplicity, suppose the anesthesia wears off at exactly the same moment for both Mark and me.

At that very first instant of regained consciousness for both of us, we will (ex hypothesi) have all and only the same memories. But, of course, we do not want to say that Mark is thereby *me* (or vice versa). (Nor, I assume, would we want to say either that *both* of us are me, or that *neither* of us are me.) It is true that (ceteris paribus) everything I can recall can be recalled by Mark and vice versa, but, intuitively, that does not make us *the same person*. Nor would the weaker claim be warranted that it makes us "duplicates" (in any metaphysically important sense of the term) of the same person; for memory – as this case illustrates – is plainly *not* all that constitutes personal identity. That, at any rate, is the commonsense intuition I hope the reader shares.

This thought experiment, though simple and familiar in construction, elucidates the thrust of my basic claim in this section. Indeed, I have less to say here than elsewhere, and this is in part because the hypothetical case just introduced bears a resemblance to one developed by Derek Parfit in his seminal *Reasons and Persons* (1984: esp. 252-266) – i.e., his "division" thought experiment (and, though to a lesser extent, others he discusses as well). There are, admittedly, notable differences between our cases – including the fact that Parfit's focus on total hemisphere division serves as a confounding factor in addressing the relevance specifically of *memory* to personal identity – but the key thing to latch onto here is that, with the aid of recent research on HP, we do not need to rely (or at least not entirely) on mere philosophizing from the armchair to establish what I have sought to show; this is because the case I have presented is empirically-grounded in (though modestly *extrapolated from*) a technology now known to be possible, and thus it is actually an "empirical/hypothetical hybrid" thought experiment, to use a somewhat clumsy turn of phrase.

Although empirical findings thus appear to provide support for the in-principle possibility of one sort of thought experiment that Parfit conjured, and while this is interesting and somewhat philosophically useful, there isn't, I submit, anything terribly groundbreaking here: as far as thought experiments go, this sort of case looks like well-worn territory. After all (setting aside Parfit's own controversial account of personal identity[5]), most of us probably already had (or would have had, if asked) the intuition that I believe is appropriate in the case of "Matt & Mark" – even if we were previously unacquainted with the apparent fact that such medical technology is actually on the horizon, and were instead

[5] This seems like as apt a place as any to note that I do not endorse Parfit's idiosyncratic and well-known view on the nature of personal identity. But that, I think, is not relevant here.

asked merely to imagine a scenario like Parfit's (where it is simply *assumed* that some such procedure could, in principle, be carried out, even if not currently technologically possible).

It is for that reason that I consider this first upshot of recent work on HP to be the least groundbreaking or philosophically important of those I discuss here. But although the foregoing therefore exhausts all that needs to be said on the matter, I hope the reader will agree that this apparent refutation of MT is nonetheless worth pointing out. And more broadly, what has been discussed here should help to further confirm intuitions I suspect most of us share with respect to what matters – and to what degree – when questions of personal identity are at stake.

Section 3: Recent Developments in Neuroprosthesis and Implications for a Theory of Mind

3.1 *Introduction*

Much more interesting, I think, is what comes to light when we bear in mind the possibility of HP when we are theorizing about the nature of the mind. HP, after all, is a procedure for (among other things) producing memories, and memories are paradigmatic *mental states*. Hence it would be most surprising if the advent of HP had *nothing* interesting to contribute to our understanding of the nature of the mind.

3.2 *Some Straightforward Implications for a Theory of Mind*

To put it plainly, the recent work on HP we have been discussing provides strong prima facie (though of course defeasible) empirical support for – and adds to the mountain of evidence (both empirical and philosophical) already in favor of – the multiple realizability

of mental states. This inference is straightforward. It appears that certain mental states (i.e., memories) are multiply realizable – they can be brought about by a hippocampus, but also by a silicon chip. If that isn't evidence for multiple realizability, I'm not sure what *would* be.

Although closely related, these findings should nudge us (or those of us not already committed) in the direction of physicalist functionalism. Nearly all varieties of dualism with which I am familiar contain nothing in their set of theoretical commitments that should *entail* these results on HP (or even the bare possibility thereof) – and it would, furthermore, take a fair amount of ad hoc (and independently implausible) maneuvering to bring these findings into agreement with some variety of dualism that hasn't already been thoroughly discredited. Physicalist functionalism, by contrast, issues *precisely these predictions*, and that should count in the theory's favor.

But set dualism aside. The really key (and, to be honest, *obvious*) insight that research on HP furnishes us with does not concern physicalism *as such*, but rather simply *functionalism* (whatever its variety); extant forms of HP come very close to (or, perhaps succeed entirely in) replicating the function of a particular brain region, and the upshot appears (at least at this point – before the results of human trials are in[6]) to be the production of *the very same*[7] *mental states* produced by the hippocampus.

---

[6] Again, one reason this is important is that memory via HP may be accompanied by *different phenomenology* than memory via hippocampus, and in order to rule this possibility out, we would need to, e.g., ask a human subject to report on her memory phenomenology after, as opposed to before, such an operation. However, note that it needn't necessarily be a problem for the view advanced here if such studies do reveal different phenomenology; as long as the *type* of mental state produced counts as *memory* (even if accompanied by different phenomenology), our conclusion can still go through with only trivial modification.

[7] Here, of course, I mean the same types of mental states, not tokens.

In fact, although there is, I contend, nothing in this research to recommend any variety of dualism (even, say, a functionalist one – though I grant for the sake of argument that dualist functionalism may be *consistent* with these findings), I want to drive home the point of the paragraph immediately above by showing that, just as research on HP appears to confirm multiple realizability and (especially physicalist) functionalism, it simultaneously tells *decisively* against other varieties of *physicalism*!

3.3 *HP as the Basis for a Refutation of Bickle's 'Ruthless Reductionism'*

One physicalist alternative to functionalism – specifically, the variety of the Mind-Brain Identity Theory (MBIT) defended by John Bickle – appears susceptible to an actual *knock-down argument* (something rare, indeed) when bearing recent research on HP in mind. Here is why.

Bickle (2003) maintains a radical anti-realizability view according to which memory *always* depends upon the CREB cycle.[8] Bearing this (admittedly bold) commitment in mind, it is not difficult to see how Bickle's view simply cannot bear the weight of recent work on HP. The latter is merely a silicon chip designed (at least, in the limit case) to be functionally identical to the hippocampus. The hippocampus, of course, is responsible for storage of short term memory and the transfer of such memories to the cortex for long-term storage. In a normal subject, then, we find (as Bickle has taken pains to bring to our attention) that consolidation of memory from short-term to long-term relies upon the CREB pathway (Bickle, 2006). And studies where CREB function is inhibited lead to impairment in memory formation. Thus, for Bickle, the psychoneural reduction of *memory* really does rely

[8] I owe a debt of gratitude to Shaun Nichols for first drawing my attention to Bickle's work.

(with the force of necessity) upon CREB function (2008) – and this, for those unfamiliar, really is intended to be an instance of theoretical reduction.

But HP, *qua* silicon chip, is not an organic piece of machinery; as such, it *does not* rely upon the CREB cycle to do its work in the storage, encoding, and retrieval of memory; and HP is effective in the consolidation of short-term memory into long-term memory in the absence of a CREB pathway. Thus, in one fell swoop, it *seems* demonstrably false to assert that memory, as one mental state under the purview of MBIT, *always* depends upon the CREB cycle. This, in truth, should not be surprising: we already have at our disposal defeasible but compelling evidence in support of multiple realizability, and any hardline anti-realizability view so radical that it ties a mental state to a *very specific* biological process should have aroused our suspicion in the first place. What is new here is simply that we have powerful empirical evidence that Bickle's *specific* target – demonstrating that memory is categorically not susceptible to multi-realizability – simply can't be right.

To avoid this conclusion, one move Bickle might attempt to make is to argue that – even in the case of HP – passage through the CREB cycle will nonetheless be necessary somewhere upstream (say, in the cortex, once short-term memories from the chip are transferred for long-term storage).

The chief problem with this suggestion, of course, is that recent work on HP gives us every reason to believe that *all other* brain regions that figure in memory can also be mathematically modeled at the neuronal level for the purpose of creating an artificial neural network via silicon chip – and such chips will rely no more on the CREB cycle than the chip used in HP.

So the CREB cycle begins to look every bit as dispensable as any good functionalist would have simply assumed at the outset: it's just (at best) a contingently necessary component of memory functions in organic critters (or, to weaken the claim appropriately: those (terrestrial) ones we're familiar with). But MBIT, in the style of Bickle, maintains a *necessary identity* between mental states and brain states so radical in its particulars that the further (non-standard MBIT) claim is posited that a *particular biological process* (again, the CREB cycle) forms one part of this identity (on the brain side of the "=" sign). After everything we have seen here, such a view looks to be patently untenable.

Section 4: HP (and Neuroprosthesis More Broadly) as a Possible Frontier of AI

4.1. *Introduction*

Third, when Hampson et al.'s findings are combined with the other recent findings in neuroprosthesis mentioned earlier, it looks as though we may be able to say something meaningful about the possibility of artificial intelligence – and, of particular interest, the surprising form it might end up taking. This is a topic I will devote a great deal of Chapter 3 to, but I think it may be helpful to close this chapter with some intuition pumping, or priming of the pump, so that the reader is adequately prepared to receive and seriously consider the first half of Chapter 3 (which, one must admit, has a rather ambitious and bold title and, accordingly, aim).

4.2. *The Basic Idea*

Perhaps it isn't difficult to see where this is headed. HP is a form of neuroprosthesis that involves (paradigmatically, in a clinical case of neurodegeneration) throwing the hippocampus to the wayside and surgically implanting a *silicon chip* designed to perform the

former's functions. Similar remarks apply to the other forms of neuroprosthesis discussed in Section 1.1: cochlear implants replace (paradigmatically defective) organic sensory equipment with artificial implants that can successfully execute such equipment's neurophysiological work; a silicon cerebellum can not only replicate the functions of its organic counterpart, but can do so more efficiently[9]; and the icing on the cake, it appears, is that we can use biomedical engineering to actually conjure novel sensory modalities *ex nihilo* (that is, as they pertain to a species for whom such a modality had hitherto simply been biologically impossible), and this gets carried out by a specially-designed silicon chip as well. The upshot, then, seems to be the following. Admittedly, we are not, technologically, anywhere near producing silicon analogues of all brain regions. But anyone looking carefully at developments in neuroprosthesis (in just the last few years!) would be hard-pressed to deny that we are *well on our way* to such a state of affairs. In the remainder of this section, I want to suggest that this sets us on a very real and possible path toward (embodied) AI.

4.3 *An Argument from HP to 'Neuroprosthesis Generalized'*

In order to get the gears of this argument moving, a certain kind of claim about the generalizability of Berger's and Hampson's work (and the easily foreseeable human applications we have been discussing) needs to be established – or, more modestly, at least rendered plausible. This is really just to codify and motivate a guiding assumption that has appeared in various locations above (in particular, directly above in Section 4.2).

[9] On this note of efficiency, it is worth drawing attention to the fact that, in nonhuman primate trials, HP *outperforms* a healthy hippocampus in certain respects.

My basic claim is that there is every reason to believe that the progress made in modeling the neuronal behavior of the hippocampus and embedding that model within a silicon chip can, as a matter of (future technological) principle, be accomplished with other brain regions as well. On such a view, it is not just a one-off fluke that neuronal behavior in the hippocampus can be replicated by silicon; there is nothing special about the hippocampus, vis-à-vis other brain regions, that should lead us into skepticism about the ability to extrapolate these technological advancements to other brain regions once the technology has developed sufficiently.

To see why, consider a coarse-grained explanation of how HP is achieved: at bottom, it is the construction of an artificial neural network produced by (extremely complicated) mathematical modeling of neuronal behavior. As regards the hippocampus, this, essentially, is what Berger and others have spent more than a decade constructing and refining.

Admittedly, this kind of work is less challenging when working with the hippocampus than when working with, e.g., the cortex, but given that we already have in hand proof of concept of the *basic approach* sufficient to achieve the desired results – and given that this approach is (in the first instance) simply the careful and meticulous mathematical modeling of neuronal behavior – what reason could be given for thinking that this basic approach will fail us as we direct our attention to other brain regions?

The difference between the hippocampus and these other regions – from the perspective of this basic approach – resides principally in (*i*) the larger number of neurons that must be modeled in these other brain structures, and (*ii*) the subsequent construction of a chip capable of housing such a vast number of neurons. But importantly, insofar as these

differences constitute obstacles, both of them (as a matter of technological possibility) are easily surmountable.

The first difference – (*i*), above – represents a merely computational challenge that we have *already* made great strides in addressing just during the interval between Berger's (2005) and Hampson's (2013); in fact, this progress is probably best modeled on an exponential curve, and there appears to be no good reason for thinking that we are on the cusp of bumping up against some unforeseen computational limit.

The second difference – (*ii*), above – simply concerns the ratio of a chip's processing power (and the number of components on the chip) to the size of the chip; but, as is well known in computer science and popular culture more generally, this ratio has transformed – and continues to transform – precipitously with the passage of time; indeed, such innovations are also appropriately-mapped by an exponential curve.

To put it in different jargon, we might cite Moore's law, as it has come to be known – first proposed in Moore's (1965) – which claims that the relationship between transistor count and the passage of time is best mapped as an exponential growth curve. (Admittedly, in recent years, the pace of this growth has begun to finally slow somewhat, such that Moore's original claim – i.e., that transistor count doubles every two years – might need to be modified to claim that transistor count doubles, say, every two and a half years. Probably, furthermore, I think we will *eventually* reach some hard and fast limit such that neither Moore's law nor some weakened version thereof serves as an accurate description of progress in the field. That, however, does not appear to be anywhere even remotely on the horizon, and I think there is every reason to believe that by the time this has happened, we will have

vastly exceeded the requisite number of components it is possible to fit onto a chip for the purposes at hand.)

As such, if my opponent wishes to argue that the remarkable advancements found in HP are just a *one-off* technological curiosity, I believe the burden is firmly on her to provide some reason – and one which takes stock of the considerations noted above – for defending such a view. By contrast, my burden – in the absence of some compelling consideration in the offing – is only to sustain a quite reasonable inductive inference which seems by all accounts to be empirically well-grounded. In my view, this is something I have already furnished above.

4.4 *An Argument* from *Piecemeal Silicon Reconstruction of Brain Structures* to *(Embodied) AI*

Suppose, then, that *all* brain regions can (as a matter of future technological possibility) be modeled such that their neuronal behavior can be captured by the operations of a silicon chip, just as we seem to have accomplished with the hippocampus. That supposition is all that is needed to have at our disposal a simple but powerful argument. (Merely for ease of presentation, suppose that each successive stage of this thought experiment involves an *elective* procedure rather than one necessitated by some neurological disorder.)

Bill's Mind

> Suppose that we are at some point in the future and Bill has decided to have his hippocampus replaced with a prosthetic device (which, as we already have compelling reason to believe, will result in no loss in the relevant functional states, i.e., mental

states involving memory). Then, in a series of subsequent procedures, he has each remaining brain structure replaced by a chip designed to mimic the function of those structures (in the same way that HP mimics the function of the hippocampus). After the final procedure, his brain has either been replaced entirely by silicon – or, as may be the case, his brain has been left intact rather than removed, but is doing no neurophysiological work; in such a scenario, it has been *deactivated*, we might say.[10]

Here already we have a slew of interesting – but, for our purposes (I believe), distracting – philosophical questions to address: Is post-op Bill the same person he was before all of this began? If not, where did the change take place? After the first procedure? The third?

For what it is worth – and do note that nothing hinges on this for my purposes – my intuition is that post-op Bill *is indeed* the same person he was before the very first procedure. But set this issue aside.

More relevant for our purposes is the following line of questioning: Is post-op Bill an (embodied) instance of AI? (I take it that his being embodied is irrelevant here; but, if the reader disagrees, simply suppose that we remove post-op Bill's silicon brain and give it a robotic, rather than an organic, body.) If the answer is in the affirmative, it may be interesting to ask *just which* brain region (or regions) had to be replaced in order to make this so. But I am more interested in reasons that could be given for thinking that the answer is in the negative: that post-op Bill is *not* an instance of (embodied) AI. Although terribly unhelpful, I *simply cannot think of any*. By hypothesis, with each progressive replacement of a brain

[10] I do not claim credit for this thought experiment; in fact, I am sure that *someone, somewhere* has already conjured it (in fact, something resembling this may be buried somewhere in Parfit's (1984); all I contend is that I have constructed it independently and without knowledge of any specific antecedent in the literature.

structure with a silicon analogue, Bill experienced no loss in (whichever) cognitive function relevant to the replacement being made.

So how could Bill – this *person* standing before us now (if we are indeed prepared to call him that, as I think we *should* be) – lack any of the types of mental states he was capable of having prior to the first operation? Indeed, if each chip performs its function indistinguishably from its original organic counterpart, how could Bill fail to be *conscious*? It seems to me plain, in other words, that post-op Bill would be a case of strong (deep, genuine) artificial intelligence.[11] In short, the case of "Bill's Mind" appears to establish the (future, technological) possibility of (embodied) AI.[12]

4.5 *A Surprising Turn of Events*

If the reader is amenable to the conclusion drawn at the end of Section 4.4, then she will notice that something strange has happened. We have (arguably) established the possibility of strong (deep, genuine) artificial intelligence, but have come at the problem *from behind*, as it were. Allow me to explain.

Typically, when we think about the possibility of creating AI, we imagine some unprecedented and critical turning point in computer scientists' efforts to move from a mere

---

[11] One (dubious) way of resisting this claim would be to insist upon some variety of dualism as the correct theory of mind, and to append to this a further claim about how post-op Bill's "soul" got lost in the series of operations; another would be to insist upon a hardline version of the MBIT such as Bickle's, for then post-op Bill would lack memories and (mutatis mutandis) all other bona fide mental states. Note, however, that I have already gone some distance toward insulating my view from such critiques. This is a product of Section 3 and its effort to establish that the technology *already in hand* (HP) tells convincingly against such theories of mind.

[12] And this should be so, uncontroversially, *irrespective* of whether we want to say that Bill is (or is not) the *same person* who consented to the very first procedure. (In other words, we can now bracket the question addressed above – i.e., the question of personal identity – since, however it is answered, the thing standing before us intuitively counts as *some* conscious being or other.

"Blockhead" (or – equivalently for our purposes – revolutionary advances in machine learning) to something that satisfies the desiderata that we are after. For those of us who think that the prospects of achieving genuine AI in this conventional manner are dim, the underlying pessimism is not far to seek: we are being asked to imagine that we *begin with a computer* and, through some byzantine process of programming, somehow manage to *flick on the switch of consciousness*. When clearly-stated, this seems like a magical feat indeed.

"Bill's Mind" – in appearing to illuminate a path toward what (if successful) would, ipso facto, *just be* genuine artificial intelligence – not only highlights one fascinating implication of HP (among the others discussed in this paper), but actually suggests a *research program* for those who would like to make AI a reality.

The suggestion is simply that researchers have been coming at the problem from the wrong angle: they begin by identifying what they take to be the hallmarks of sentience or consciousness in humans and then, from the ground up, attempt to mold a computer into exhibiting those hallmarks; they try to get a computer to behave *like that*. But, as many of us have probably intuited for quite some time, no amount of *machine learning* is going to yield the sorts of results we are after; it is not likely to permit us to pass some threshold beyond which those hallmarks of human consciousness actually *are* possessed by a computer.

On the other hand, if we begin with something that *already* possesses those hallmarks (i.e., a human being), no such "backward engineering" is necessary. Although no simple task, we can devise silicon counterparts of the thing we have begun with (i.e., human brain structures) – and, crucially, the advent of HP and other neuroprostheses suggests that we *can* do this – and then, once we have reached a certain level of technological sophistication (i.e.,

enabling us to create prostheses of *every* brain structure), we will have simply *arrived at* (an embodied) instance of AI. And in *that*, there is no magic – no insurmountable threshold or mysterious switch to flick.

The only obstacles to this are challenges in mathematical modeling and chip sophistication: i.e., the smaller the better; and yet, pulling in the opposite direction, the more artificial neurons that can *fit* on the chip, the better. But these are mere technological feats whose achievement is (with overwhelming plausibility) on the horizon – or so I have argued here.

But when that is contrasted with the traditional AI research program, what we are entertaining now looks like child's play: with respect to the traditional research program, there simply is no (obvious, or – at this point – perhaps even conceivable) technological feat such that – if only we possessed that future technology – the problem of creating genuine AI, *beginning with a computer as our starting point,* could once and for all be accomplished.

I will have more to say about AI – and in particular, some ethical issues surrounding it – in Section 5.5 (and its subsections), but first I wish to discuss some of the less abstract and more immediate ethical challenges confronting those (i.e., medical practitioners) who are correct to see promise in the use of HP as a treatment for neurodegenerative disorders such as AD.

Section 5: The Ethics of Hippocampal Prosthesis as a Potential Future Treatment for AD

5.1 *Introduction*

In this section, I am particularly concerned to address – with respect to the findings of Hampson and his colleagues in their work on HP – the potential moral issues such

findings raise for medical practitioners. Researchers working on HP are doing so, by their own admission, because they think it may eventually prove useful in the treatment of AD and other neurodegenerative conditions (although perhaps they foresee other potential benefits of this research as well). However, even if Hampson et al.'s findings with primates can be replicated in humans, it is *not merely a given* that such surgical procedures should actually be undertaken – even, say, in cases of advanced AD. Indeed, such a view requires a sustained moral defense.

5.2 *A Brief Primer on AD and Hippocampal Function*

AD is characterized by plaques and tangles in the brain: it begins with short-term memory loss, proceeds to deficits in motor skills and language, progresses to long-term memory loss, and culminates in disorientation and immobility. Neurodegenerative disorders, including AD, are characterized in large part by deficits in memory resulting (in the first instance) from impaired hippocampal function in the medial temporal lobe. Hampson, et al. note the following:

> In the mammalian brain the hippocampus has been shown to be the most important structure involved in the encoding and retention of new information in cognitive processes…It is well documented that impairment of the functional status of the hippocampus in human disease states leads to memory deficits that are detrimental to normal function, and in addition such impairment has become the hallmark of brain aging and deterioration as exhibited by Alzheimer's patients (2013: 2).

While the underlying causal mechanisms aren't yet entirely perspicuous[13], we have a fairly clear grasp of how AD progresses: it begins with shrinking hippocampal tissue and degeneration of hippocampal cells. This leads, naturally, to deficits in memory. However, the disease then spreads throughout the cerebral cortex; it is damage to this area that accounts for changes and decline in functions of language, behavior, and judgment, which persist (and worsen) for roughly a decade, until death.

Since the cause of AD is still a matter of dispute, it is probably best to bracket this matter for present purposes. But what is critical is that, whatever the cause, the hippocampus is clearly the first brain region the disease acts upon. De Leon and colleagues (1993, 1997) found support for the contention that atrophy of the hippocampus in elderly subjects is predictive of subsequent AD; indeed, such atrophy frequently occurs *before the patient is symptomatic at all*. These results were replicated by Henneman, et al. in their (2009).

Additionally, Henneman and colleagues found that AD is more common in those who *start out* with smaller hippocampi. This alone appears to be a compelling (though defeasible) reason for thinking that – whatever the cause of AD – the disease's gears simply can't get moving unless it has a hippocampus to act upon.

In further support of this contention, consider the following statement issued by the NIH's National Institute on Aging:

[13] Standard theories focus on specific proteins (e.g., tau proteins) or peptides of amino acids (i.e., Amyloid beta) as potential causes. Although only in a subset of cases, the presence of a genetic component is well-established (specifically, a mutation of APOEe4 in sporadic or "late onset" AD; and, in familial or "early onset" AD, a variety of different genetic mutations). Additionally, some of the current research focuses upon the causal role that environmental factors might play. Recently, Itzhaki, et al. (2016) have argued – although not without controversy – that certain microbes are implicated in the etiology of AD (in particular, herpes simplex virus type 1, varieties of spirochatete, and chlamydia pneumoniae).

> It seems likely that ***damage to the brain starts a decade or more*** before memory and other cognitive problems appear. During this preclinical stage of Alzheimer's disease, people seem to be symptom-free, but toxic changes are taking place in the brain. Abnormal deposits of proteins form amyloid plaques and tau tangles throughout the brain, and once-healthy neurons stop functioning, lose connections with other neurons, and die.
>
> ***The damage initially appears to take place in the hippocampus***, the part of the brain essential in forming memories. ***As more neurons die, additional parts of the brain are affected***, and they begin to shrink. By the final stage of Alzheimer's, damage is widespread, and brain tissue has shrunk significantly. (2016, emphasis added.)

It appears fairly plausible, then, that simply removing the hippocampus would halt the progression of the disease to the cerebral cortex. Although this would by no means be considered an acceptable therapeutic option, what is relevant for our purposes is only whether performing such a surgical procedure (say, at a sufficiently early stage) would, indeed, stop the disease process.

Suppose, for the sake of argument, that it would.[14] It is this assumption, I take it, that underlies the research program on HP. For if we could not only remove an AD patient's

[14] Note, however, a challenging objection (for which I owe a debt to Dane Muckler, who posed the problem during my talk at the 2016 Medical Humanities Conference at Western Michigan University): Perhaps (one might speculate) AD does not actually originate in the hippocampus after all, but instead begins in a sequence of events elsewhere, with it only *appearing* to us that AD "originates" in the hippocampus because, for instance, it is simply the "weakest neurological link." If that turns out to be the correct account of the etiology of AD, it's unlikely to make a difference whether we prophylactically remove the hippocampus (and replace it with HP) or *not*. In response to this worry, I of course readily acknowledge that – for all we know – this *could be the case*. And if it is, then the usefulness of HP as a treatment for AD is questionable at best. In point of fact, though, this worry – while serious for the optimistic neurologist – isn't germane to my aims here. As I note directly above in the body of the text, I am simply following the lead of those HP researchers who are operating on the *assumption* that HP would (or at least could) be therapeutically valuable for AD, and I am doing this only for the purpose of investigating what – if any – ethical worries would be apt to arise from such a medical intervention.

hippocampus and thereby eliminate the possibility of the formation or progression of the disease, but also *replace* the hippocampus with a prosthetic device (i.e., a silicon chip) that replicates the functions of the hippocampus, this *would* seem to be an acceptable treatment option (at least if certain further conditions obtain).

That, after all, appears to be the thinking that motivates Berger and colleagues' conclusion that their research promises the potential for a "biomedical remedy for the cognitive and memory loss that accompanies Alzheimer's disease…" (2005: 241)

Although I think the data does (and will continue to) bear out this surprising claim, it is not actually important for present purposes that it be demonstrated – or even defended. For, in this section of the paper, what I wish to do is to take this emerging technology at face value – and, furthermore, append to it the supposition that the results already in hand will transfer seamlessly from nonhuman primates to humans – and to then take a careful look at where this leaves us morally.

5.3 *Some Erroneous Ethical Worries*

It seems plain to me that – if and when a surgical procedure involving HP can be safely and effectively performed in AD patients – any ethical worries one might countenance are *handily defeated* by the good that would attach itself to the therapeutic outcomes we are imagining. Nonetheless, nearly every philosopher with whom I have discussed this has disagreed (which, for my own part, I find remarkable).

Some of these worries, I think, can be dismissed outright, as they appear to hinge on a misunderstanding of hippocampal function (and neurophysiology more broadly). A few of these are:

i. That, in performing this procedure, one will have effectively *killed* the AD patient; the person who emerges from the operation will not be the *same person* who consented to it;

ii. That, in removing the patient's hippocampus, we risk obliterating her memory of, say, all *language* (or, alternatively, knowledge of basic facts – e.g., "Paris is the capital of France"); this, in turn, would require a (typically elderly) patient to "start from scratch" developmentally, and performing an action that involves the intentional deprivation of such faculties in a person is not morally permissible. (Or in a similar vein, as it is sometimes put to me, the attendant confusion of the post-op AD patient would involve a morally unacceptable degree of psychological suffering.);

iii. That, even if we have not (strictly speaking) *killed* the patient by performing this operation, we will have intentionally performed an action that will deprive her of anything that could *matter* to her; she will have no recollection of loved ones, nor is there any reason (or so it seems) to believe that she will *care about* anything she cared about before the operation, as one's values, interests, and concerns cannot persist in a vacuum of memory. (Or, as it is sometimes put: the operation will destroy her *personality*, even if not her numerical identity to the person who consented to the procedure.)

Again, all three of these ethical worries are misplaced; furthermore (as the reader may have already anticipated), they can all be dealt with in a similar manner.

As regards the rather bold claim made by (*i*), I take it that my interlocutor has something like the following in mind. Personal identity – in the sense of strict numerical identity (as it applies to persons) – is constituted by *memory*. (Somewhat more formally, we might (as before) put that familiar theory of personal identity thus: *person S at time t+1 is the same person as person P at time t if and only if S has the same memories as P (and, in most instances, others as well).*) Thus if you erase all of my memories, although my body survives, I (the *person*) am gone.

There are two things to be said here, one much more foundational than the other. First, this objection to HP simply *presupposes* the truth of the memory theory of personal identity, whereas in point of fact that theory falls prey to a number of familiar, and very likely fatal, objections. But much more importantly, this objection makes the mistake of assuming that removing one's hippocampus means eradicating all of one's memories. In point of fact, however, this is simply not so. As Berger points out, damage to (or total removal of) the hippocampus only results in a loss of the ability to store new memories (or, more accurately, to take short-term memories and transform them into long-term ones); as he notes, and as may be physiologically familiar, "long term memories are not stored in the hippocampus" (2005: 245). Hence, an AD patient undergoing a surgical procedure for HP will emerge from the operating room – contra what my opponent has simply *assumed* – with all of her (already-formed) long-term memories intact.

There is, admittedly, a wrinkle here. In addressing it, we shall have the opportunity to recall the details of the famous case of Henry Molaison (known previously merely as H.M.), and this will be useful for other purposes as well. For economy of exposition, I shall only very briskly rehearse the central elements of this case study (and even then, only those relevant for our purposes), and I shall do so in an elliptical, bulleted form:

(a) H.M. was an epileptic whose surgeon elected to treat him by removing his hippocampus;

(b) This cured his epilepsy and resulted in no significant changes in his personality;

(c) However, the procedure famously resulted in a severe case of anterograde amnesia, or the inability to form new long-term memories. (This, of course, was due to the fact that the hippocampus is responsible for transferring memories to the cortex for long-term storage.);

(d) He did, however, have a functioning short-term memory (approximately thirty seconds' worth, as well as "working memory." (H.M.'s deficits pertained to the formation of *new* declarative memory (episodic memory, or memory of events; and semantic memory, or memory of facts) because the hippocampus plays a central role in this. However, he still had working procedural memory, because the critical brain regions here – the basal ganglia and the cerebellum – were left intact.);

(e) His lexical memory was largely unaffected by the procedure;

(*f*) In addition to total anterograde amnesia, he suffered from *partial* retrograde amnesia: namely, he lost most memories from 1-2 years prior to the operation, as well as some of his memories from the previous decade;

(*g*) Recently (in 2014) a postmortem 3D reconstruction of his brain revealed that half of his hippocampus had in fact remained intact; admittedly, this complicates considerably any attempt to draw decisive neurophysiological conclusions from H.M.'s case.[15]

For the moment, focus only on (*f*), as this is the source of the "wrinkle" I alluded to above. At least in this one case, it appears that removal of the hippocampus *did*, as my interlocutor cautioned, result in the loss of (some) already-formed long-term memories. For the sake of argument, I shall set aside the fact that this was very likely not the result of the removal of H.M.'s hippocampus *as such*, but rather the result of his surgeon having removed other parts of his medial temporal lobe as well. For, even if this were to be a routine, anticipated byproduct of a procedure to remove the hippocampus, it seems highly dubious to suggest that it would count as a morally overriding factor when the alternative (as we have been imagining) is eventual progression to AD (as a result of foregoing an operation to remove the hippocampus and replace it with a prosthetic device).[16]

---

[15] For reasons of space, I shall leave this complicating factor largely unexplored. It was not, after all, strictly speaking *necessary* to adduce the case of H.M. in response to the three worries enumerated above, but rather just a convenient device; even if forced, dialectically, to *throw out* the case of H.M., the responses I give here are well-supported by what is today considered fairly uncontroversial neuroscience.

[16] In the interest of clarity: obviously, I am not suggesting that H.M. *himself* would have developed AD; he was, rather, an epileptic. Rather, I am simply refocusing our attention on the medical condition for which HP

More to the point, and with reference to (*b*) above, the contention that removing one's hippocampus results in the destruction of the person is simply not borne out by the facts: although afflicted by a tragic inability to form new memories, H.M. was very much indeed the *same person* he was before the operation – or, at any rate, his personality remained unchanged.[17]

So much for spurious ethical worry (*i*), above. Fortunately, similar remarks apply, mutatis mutandis, to worries (*ii*) and (*iii*), and so I shall be brief.

As regards the worry concerning language competency in (*ii*) above, simply note (*e*): although H.M.'s ability to form *new* declarative memories was obliterated – and although he suffered from retrograde amnesia of certain stretches of time (though, importantly, this consisted in the loss of episodic memory, not semantic memory) – in the main, he did not need to "re-learn everything" or "start from scratch." And our reply to the final part of worry (*ii*) consists in simply pointing out two facts: first, any confusion (and hence psychological suffering) H.M. may have experienced is not something we should expect when, in addition to removing the hippocampus, we *replace* it with a prosthesis functionally identical to it; second, if the ethical worry is really fundamentally about disruptions of psychological stability, then it is worth pointing out that in the case of AD, this is an inevitable facet of its late stage. As such, how could it be morally impermissible to inflict upon a patient some

---

has been floated as a potentially efficacious – and, not without controversy, morally permissible – form of treatment.

[17] Although this argumentative device strikes me as entirely dialectically superfluous (given that what I have said above seems more than sufficient), I think the following, different, response to (*i*) – which can be found several pages above – has strong intuitive appeal: If left untreated, the patient's identity is obliterated; hence, how can it be a morally less-desirable alternative to obliterate that identity prophylactically, and thereby give the patient a chance to lead a normal life (although, perhaps, as a different *person*) thereafter? [Never mind the further fact that doing so would, by hypothesis, eradicate the burdens placed on the patient's family members and other caregivers.]

temporary confusion if the objective in doing so is to forestall the development of a disease guaranteed to result in pervasive confusion *in perpetuity*?[18]

Finally – with respect to (*iii*), above – as we have already seen in our response to worry (*i*), it is *just not true* that removal of the hippocampus destroys one's personality (refer again to (*b*)). Nor is it the case that we will have necessarily tampered with the patient's extant long-term memories (and hence values, concerns, interests, et cetera). In short, when bearing in mind the relevant empirical facts, worry (*iii*) is simply specious.

5.4 *What Ethical Worries Might Actually Be Legitimate?*

Although the worries addressed above are ultimately wrongheaded, it's not all a bed of roses with HP as a potential treatment for AD. First of all, although (for the reasons mentioned in Section 1.3) I do think there is every reason to believe that HP will eventually be as effective in humans as it has been in nonhuman primates, it is true that there is no guarantee of that; and although (for the reasons mentioned in Section 5.2) I do think it is very plausible that hippocampal removal and prosthesis would be *successful* in preventing AD – say, by screening for the implicated genetic mutations routinely, and taking measurements

[18] Here, however, an objection has been presented to me (for which I am grateful to Andria Bianchi) that is worth considering. A familiar moral precept that many of us are inclined to endorse is that there is a morally significant difference between initiating a sequence that results in harm (on the one hand), and merely allowing that harm to unfold (on the other hand). For my own part, I do think that the "doing/allowing" distinction is sometimes morally relevant; furthermore, particularly in medicine it is a first principle to do no harm (whether we think of this as merely a professional standard or a bona fide moral principle). However, we clearly already recognize that this principle is defeasible, for all we have to consider is the routine practice of breaking a patient's ribs in order to resuscitate her. And it is actually *that* sort of infliction of harm that I think is clearly the most analogous to what we are dealing with when employing HP as a treatment for AD (assuming, that is, that the infliction of any such harms resulting from the treatment are even inevitable).

of pre-symptomatic hippocampal size as a matter of course, and then preventatively replacing at-risk individuals' hippocampi with a chip – there is no guarantee in that either.

But these aren't ethical difficulties; they merely concern (in the first instance) what is, and is not, *possible*. Setting these issues aside, then, it seems to me that there are a meager few *legitimate* moral difficulties involved here.

5.4.1 *Losing the Ability to Forget*

The first of these potentially legitimate worries has gained some currency in the literature: it is the thought that HP would very likely constitute a form of cognitive *enhancement*, leaving patients with better (i.e., more accurate) memories than they previously possessed (or, more to the point: better memories than the average person possesses); as such, might such a patient lose the *psychologically healthy* ability to *forget*? Bernard Williams, for one, appears to think this would be disastrous. I'm inclined to agree that the loss of *just that ability* would be a bad thing indeed.

But what appears to be missing here is an actual argument for thinking that a chip designed to *mirror* the functions of the hippocampus would out of necessity be too good at its job and send *everything off* to the cortex for long-term storage. And even if this *did* happen, it's not clear that we'd thereby be any less prone to *forgetting* (as, say, a defense mechanism). Indeed, it seems to me that *already* the vast preponderance of things we allow ourselves to "forget" are really in our long-term memory after all; it's not as though traumatic memories are hanging out in the hippocampus such that if the latter is replaced with an upgraded

silicon model, all hell will break loose.[19] Finally (if none of the foregoing is convincing), if we are able to successfully employ HP as a treatment of AD, it would appear to be a relatively minor step to achieve the ability to delete specific traumatic memories from the chip. This latter consideration leads naturally into the subject matter of the following subsection.

5.4.2 *Orchestrating the Ability to Forget*

As it may be foreseen from this section's title, the objection here is roughly the inverse of the objection just addressed. I'll call it the *Eternal Sunshine Objection* (ESO), as it was first presented to me with reference to the 2004 Michel Gondry film "Eternal Sunshine of the Spotless Mind."[20] For those unfamiliar[21], a brief synopsis is in order: the film follows a romantic couple whose relationship turns tumultuous. In the movie, the nascent technology exists for physician intervention with the aim of erasing not only specific memories but also the entire memory of, say, a particular person (by simply erasing each memory involving him or her). The female protagonist elects to undergo this procedure, and when the male protagonist learns of this, he elects to undergo the procedure as well. Predictably, the result isn't a bed of roses: we follow the inner mental life of the latter as the procedure is carried out, and because the procedure was not *entirely* efficacious (leaving him with some memories of his beloved), he soon regrets having chosen to have his memories of her obliterated.

---

[19] Almost certainly, traumatic memories are already committed to long-term storage (at which point the hippocampus – or a hippocampal prosthesis – no longer plays a role at all), and to the extent that we are able to exercise the healthy ability to forget, this involves our intervening (perhaps unconsciously) in the process by which such memories are consciously accessible. As such, HP – as a stand-in for the hippocampus – no more threatens the ability to forget (when appropriate or psychologically prudent) than does simply possessing a normally-functioning hippocampus.

[20] I thank Garret Merriam for raising this objection during my talk at the 2016 Medical Humanities Conference at Western Michigan University.

[21] Spoiler alert.

Bearing this in mind, the objection runs roughly as follows. With further (plausible) technological refinement of HP, scientists might develop the ability to allow HP patients themselves to begin tinkering with the parameters of the chip such that they can erase specific (presumably unpleasant) memories. But as the film illustrates vividly, catastrophic possibilities here lurk.

However, ESO faces at least one serious problem. Most importantly, it is just not at all clear that there would be anything *morally* problematic about this. In fact, it is not at all clear that there would be anything *professionally* problematic about a physician enabling a patient to do this.

Note, first of all, that we *already* possess the ability to execute "memory dampening", which is the process of deliberately softening the impact of certain traumatic memories (or "taking the edge off them"), typically with the use of propranolol shortly after a traumatic experience. In fact, in certain cases, it is possible to thereby *erase* specific memories.[22]

We don't think (I take it) that such persons are engaged in an activity that is morally wrong; and we don't think that the physicians who assist them in this have run afoul of professional standards or codified (or professionally binding) principles in medical ethics. (Of course, some of us might be horrified by the thought of undergoing such a procedure ourselves, but that tells us nothing whatsoever about whether it is within the legitimate scope of an agent's autonomy to elect to do so herself.)

Indeed, memory dampening is a fairly routine method of treatment for PTSD (or, as the case may be, to prevent its very onset). Setting aside already-accepted therapeutic

[22] Interestingly, some of the current research on memory dampening involves neuronal manipulation in the hippocampus, and so this topic dovetails quite nicely with several of the issues we have dealt with here.

conventions, however, we can respond to ESO in an even more fundamental way. Suppose that the scenario my interlocutor has described unfolds just as depicted, and crisis ensues. (Never mind the fact that, should the medical community come to a determination that patients' manipulation of HP operations is morally fraught, presumably regulatory oversight and enforcement on the limits of such technology could plausibly be enacted.) Even so, as a matter of patient rights and autonomy – when compared with, say, physician-assisted suicide – this issue appears to fall *much* more clearly within the range of cases where agreement can be reached that (even when a physician advises against such action) *actively depriving* a patient of the potential treatment option in question would be unacceptably paternalistic and would infringe upon the patient's autonomy.

There is, however, one last thing I want to flag about ESO. Earlier we considered the roughly inverse worry that HP might result in individuals losing the healthy ability to forget. But raising ESO as a distinct objection actually functions as a response to that earlier worry. That is to say, one additional manner of response to the objection that implementing HP would put us at risk of losing the ability to forget is to point out that it is in principle possible to *orchestrate* forgetting (by way of medical intervention), as I alluded to in 5.4.1. And this needn't even take the form of an HP patient manipulating the parameters of her prosthetic; good old-fashioned memory dampening is already making headway there.

5.4.3 *Diminished Capacity for Informed Consent*

What strikes me as one of the most pressing moral complexities involved here, if only because the matter is something to be taken very seriously by the medical community *wherever*

it threatens to surface, is the potential of diminished capacity for consent among the target population.

One might reason as follows: if *S* is a candidate for HP in the treatment of AD, that very fact means that *S* has a damaged hippocampus and hence difficulty in forming new memories; but perhaps forming new memories is necessary for informed consent (or even mere *assent*). As such (the argument continues), the patient's very condition (involving, as it does, impaired mental faculties) disqualifies her from consenting to have her condition treated in this way.[23]

I admit to feeling the force of this claim, as informed consent is – from the point of view of the medical practitioner – of paramount importance. But I am inclined to think that this worry trades in a kind of confusion concerning just what is being proposed here. It is evident that *if* HP is to work as an effective treatment for AD, it would require not only removing the hippocampus and implanting the prosthetic, but doing so in the preclinical stage. It would, in other words, be a prophylactic measure.

Recall from section 5.2 that damage begins to occur in the hippocampus typically a decade before the patient becomes symptomatic. Not only would the *success* of the procedure *hinge* on performing it during this stage, but doing so would also guarantee (ceteris paribus) the ability for informed consent – for, by hypothesis, the patient would not yet show signs of cognitive impairment. Perhaps even *this* would be too late; it could turn out that as soon as preclinical, asymptomatic damage begins in the hippocampus, intervening even at that

[23] Here the special issue of "legally authorized representatives" (i.e., "surrogates", "proxies") enters into the picture and presents unique complexities all its own. However, the way in which I shall address the topic of impaired consent makes it unnecessary to wrangle with this.

early point won't stop the spread of the disease to other brain regions. In that case, HP as a treatment for AD would need to be implemented even earlier, perhaps preventatively on the basis of known risk factors present in a given patient. But then, just as in the scenario discussed immediately above, the complication of impaired consent never (ceteris paribus) has the opportunity to arise.

5.4.3.1 *Brief Digression: Personal Identity and Informed Consent*

In Section 2.2, I argued that the success of HP demonstrates (although perhaps only defeasibly) that MT is false.

Suppose that this is correct. What are the implications for medical ethics, broadly? I think it is clear that it matters, ethically, whether MT is correct. The most obvious upshot of MT's being false (of those relevant to the medical community) concerns *consent*. (I've had occasion above to discuss the possibility of impaired consent with respect to HP and AD, but here I am speaking more broadly, so that earlier discussion need not be consulted here.)

To see why, suppose that MT were actually a *correct* theory of personal identity. Then, as medical practitioners, we should be able to say things like this:

> Ashley consented to procedure *P*, but before the procedure could be performed, she lost all (or, what works equally well: a sufficiently sizeable number) of her memories; hence, by MT, Ashley is *now* no longer the person who gave consent to *P*. Thus we cannot use this prior giving of consent as a license to perform procedure *P* upon her *now*.

But if, as I have argued, MT is false, then (ceteris paribus), Ashley is indeed the same person she was when she initially gave consent, her loss of memory notwithstanding. Hence, since she is the same person who gave consent, that consent remains in full force (from a medical practitioner's deliberative stance, to cite the relevant scenario). And if this is correct, it appears (or so I gather) to go against the grain of standard moral reasoning in the medical community. That is, bioethicists and medical practitioners themselves appear merely to *presuppose* the importance of memory for informed consent; c.f., for instance, the (2015) *Presidential Commission for the Study of Bioethical Issues* ("Gray Matters: Topics at the Intersection of Neuroscience, Ethics, and Society, *Volume 2*"). But if what I've argued above is correct, then they are incorrect in this presupposition.

Or, supposing that I am wrong in this impression of the dominant, received ethical thinking of bioethicists and medical practitioners – such that most would in fact count informed consent given prior to (say) total memory loss as still ethically in-force and binding after such cognitive impairment – then, in that case, what I have said above can be seen instead as a philosophical justification of the views already held by such persons.

In any event, the ethical views of medical practitioners (concerning the circumstances they encounter professionally) should be responsive to careful and deliberative moral reasoning – and not the other way around.

5.5 *Though Speculative, One Potentially Serious Ethical Worry*

The moral problem that now concerns us (which I have classified as both speculative and potentially serious) requires the laying of a fair amount of groundwork. Fortunately, that

was already accomplished in Section 4. Hence, we can move rather briskly through a recapitulation of that argument and subsequently use it to reveal the precise nature of the one ethical worry associated with HP that I *do* believe is potentially both looming and serious.[24]

5.5.1 *Brief Recapitulation of Section 4*

What follows is a distillation of the content of Section 4, presented as an argument in premise and conclusion form:

1. Research has shown that it is possible to successfully create and implement neuroprostheses using silicon chips for the purpose of replicating the functions of certain brain regions (e.g., hippocampal prosthesis in nonhuman primates).

2. Although defeasible, we seem warranted in thinking that such devices will work just as effectively in humans as they do in nonhuman primates.

3. Given (1), we should also predict that in the near future, it will be technologically possible to devise silicon chips capable of serving as neuroprosthetic replacements of *other* brain regions as well.

4. If (3) is correct, then – given (2) – we should expect in the future to be able to successfully create and implant in *humans* neuroprostheses replicating the function of all other brain regions as well.

[24] For ease of presentation (and in order to limit redundancy insofar as possible), I shall give this recap as a numbered argument, in premise and conclusion form, which I hope shall crystallize the central reasoning of Section 4. In effect, it will be an attempt to prove the *future technological possibility* (as opposed to the merely broad, conceptual possibility) of AI. I undertake this task directly below.

5. It follows from (4), however, that a piecemeal replacement of each organic brain region should be possible; and, furthermore, that a person whose brain had been entirely replaced with silicon correlates would count as an (embodied) instance of artificial intelligence (see the case "Bill's Mind" from section 4.4).

6. From (5) we are justified in inferring not merely the conceptual possibility of AI but, in fact, the (likely) future technological possibility thereof.[25]

5.5.2 *A Sketch of the Moral Problems Involved Here*

As I argue in "How to Create Artificial Intelligence – and yet the Moral Impermissibility of Taking Any Steps to Do So" (also part of this dissertation), there is a way

---

[25] A quite similar, but structurally different, argument could be used instead (if, say, the above argument is found unconvincing). Thus, if necessary, consider the following alternative:

1. Current biomedical technology includes an impressive array of neuroprostheses, including hippocampal prosthesis.
2. The methods by which such neuroprostheses have been achieved are prima facie generalizable: there is no good reason for thinking that a silicon version of the hippocampus (or cerebellum) is possible and that, say, a silicon version of the cortex is not – at least in principle.
3. Given (2), it appears possible in principle (pending only sufficient technological innovation) to create silicon versions of each human brain region.
4. So it is possible in principle to create a silicon version of each human brain region.
5. If we take a subject, *S*, and replace each organic brain region piecemeal with a silicon analogue, we will end up with a subject (either *S*, if the subject is identical to who we started with – or *P*, a different, but equally conscious subject) whose mental states are of the same kind as the original subject *S*.
6. But this subject – whether *S* or *P* – will, ex hypothesi, be not only conscious but also a subject such that his mental life is the product entirely of silicon, and nothing organic.
7. Such a subject, by definition, will be an instance of artificial intelligence – although perhaps embodied (unless measures are taken to extract all of the silicon chips and place them within a computer rather than the human body they were originally transplanted within).
8. So artificial intelligence is possible. Furthermore, the mechanism by which AI (as described here) is achieved is a process plausibly realizable in the near future, given what we know about recent advances in neuroprostheses, which serve as antecedents for the kind of developments referenced by the claim in, e.g., (3).

Support for each of these premises can be found in Section 4; here I have merely sought to distill the core insights of that section into an inductively strong (and, I contend, cogent) argument in support of the conclusion directly above.

of framing the familiar doomsayers' worries about AI that is devoid of much of the unnecessary (and seemingly, alarmist) baggage so common in the literature; and the argument I present claims (as the paper's title suggests) that it is morally impermissible even to be *working to bring about* the creation of AI.

Insofar as the *moral* portion of that argument is successful, it is worth noting that, as we shall see, it makes no assumptions about the likelihood of being able, technologically, to actually *achieve* the kind of deep, genuine artificial intelligence that the argument claims to be morally problematic. But given the conclusion of the argument directly above, the argument in that paper appears to be on even more solid ground. For, if there is a positive reason for thinking that AI is not only conceptually possible, but also technologically possible in the relative short-term, then the worries which that paper raises are yet more pressing and deserve our immediate attention all the more.

To refocus, it is worth recalling that this section – Section 5 – is devoted to ethical concerns about hippocampal prosthesis. It's worth noting, then, that the foregoing isn't a worry about HP *as such*, but rather a slippery slope argument, the merits of which are, admittedly, somewhat ambiguous.[26] However, in defense of my basic claim, there *is* a foreseeable and very real possibility that current advances in neuroprosthesis (principally, HP) can be generalized and used to create silicon versions of other (indeed, perhaps, *all*) brain regions as well; and there is likewise a foreseeable and very real possibility of using such achievements for the purpose of constructing, piecemeal, an embodied AI (as introduced in section 4 and laid out explicitly in section 5.5.1 directly above).

[26] Although of course I contend that this is not an instance of a *fallacious* slippery slope argument.

But it is not as though merely embarking on that first step – refining HP and employing it in humans with, say, neurodegenerative disorders – would of *necessity* lead to the result (namely: strong, but embodied AI) that I argue, in the following chapter, would be a bad thing. For one thing, certain regulatory measures and various (potentially) effective safeguards *could* be adopted to prevent this from ever being realized. (However, as I argue in Chapter 2, a crucial difference between the view I am defending and the view of, say, Musk, Gates, or Hawking, is that I do not think it is in fact possible to implement such safeguards – that is, without doing something morally wrong.)

But my principal contention is simply that, since (as I have argued) the most promising path toward AI will come from continued research on neuroprosthesis (with HP as the first important step in that direction), and since (as I argue in the chapter directly below) it is morally impermissible to work toward bringing about AI, there is a strong case to be made for the prima facie moral wrongness of continued work on neuroprosthesis.

And it is *only* in that attenuated sense, *rather than* in any of the senses discussed above (pertaining principally to concerns within medical ethics), that there is a moral problem with HP.

HOW TO CREATE ARTIFICIAL INTELLIGENCE – AND YET THE MORAL IMPERMISSIBILITY OF TAKING ANY STEPS TO DO SO

In the first chapter, I discussed a wide range of philosophical issues pertaining to the development of Neuroprosthetics in general, but hippocampal prosthesis in particular. One of the implications of this research that I argued was of central importance concerned artificial intelligence; indeed, I gestured toward just what this connection is and how it might be fleshed out. Here, I take this a step further to develop a simple, but highly plausible, argument for the claims that deep, genuine artificial intelligence is possible, and that there is in fact a clear but inadequately explored recipe for bringing it about. However, I also argue that there are very grave dangers associated with the creation of AI. This is nothing new; indeed, enough influential public figures (such as Elon Musk, Bill Gates, and Stephen Hawking) – in addition, of course, to philosophers and other academics – have issued such warnings that it has practically become a part of our public discourse or common lexicon. However, in this chapter I try to canonize these worries by developing an argument that, if successful, would establish the strict moral impermissibility of AI research altogether.

Section 1: Introduction (Brief Reiteration of Relevant Background on Neuroprosthesis)

In recent years, research in neuroprosthesis has made great strides. This has made possible a number of familiar scientific (and, in some cases, applied, medical) advances, such as the cochlear implant, which – in its most advanced form – actually repairs hearing by bypassing the ear altogether (i.e., the implant sends auditory signals directly to the brain rather than merely amplifying sounds, as hearing aids do). Lesser-known achievements in this sphere include Luo, et al.'s (2016) silicon version of the cerebellum, which enables

improved sensorimotor control and learning when compared to subjects outfitted with run of the mill, organic cerebellums. And, perhaps most peculiar of all, Hartmann, et al. (2016) have shown that it is possible to endow creatures with *new sensory modalities* altogether via neuroprosthesis.[27]

These developments are all quite fascinating. But it is the development of hippocampal prosthesis (HP), I contend, that should be of particular interest to philosophers and cognitive scientists, in part because of the significant implications its advent has for research conducted in the field of artificial intelligence (AI). Hampson, et al. (2013) have shown that it is possible, in not only rodents but also in nonhuman primates, to create an artificial hippocampus that can be surgically implanted such that the subject can write information back to the chip component of this hybrid brain – and to do so at the *individual neuron level* – bypassing the hippocampus altogether.

As I have argued above, there are at least four distinct philosophical domains for which the advent of HP has direct import. Here, however, I shall focus solely on its relevance to AI: when combined with certain other highly plausible assumptions, the development of HP not only has important and novel implications about the possibility and nature of AI, but in fact suggests a new, and quite specific, *research program* for AI.

[27] In Hartmann et al.'s work, this involved implanting, in the somatosensory cortices of rodents, a chip designed to enable subjects to discriminate infrared light. Although their choice of sensory modality, for proof of concept, happened to be infrared light discrimination, we can imagine far more interesting (and not altogether implausible) human applications. For instance, Hartmann and colleagues' work suggests that there is no in principle barrier to designing a chip that endows its human recipients with a sensory modality as foreign as echolocation.

Section 2: Brief Synopsis of Research on HP

Hampson and colleagues (2013) have shown that it is possible, in rhesus macaques with impaired hippocampal function, to develop and deploy a neuroprosthesis that effectively repairs and enhances memory encoding, storage, and retrieval – and even facilitates *recovery* of memory lost due to disease; they suggest, with overwhelming plausibility, I think, that such technology can, in principle, furnish similar results in humans burdened by hippocampal impairment (2013: 13). In fact, interestingly, they note (2013: 12) that "the extent and range of effectiveness in improving cognitive performance in [this nonhuman primate memory study] was much greater" than when this very same HP was deployed in earlier rodent studies (such as Berger et al. [2011] and Hampson et al. [2012]). Since human neurophysiology is obviously by significant measure more similar to non-human primate physiology than it is to rodent neurophysiology, perhaps analogously we should expect to be able to say something similar about the effect we are likely to find in future human studies *in relation to* extant studies of non-human primates. That is to say, perhaps a comparison of the efficacy of HP as it will be employed in humans as opposed to primates will bear a similar relation to the improvement in efficacy when moving from rodent trials to primate trials. Although quite plausible, this is nonetheless mostly speculative, and we will naturally have to wait until human trials are complete to say anything decisive or illuminating in this respect; fortunately, however, it is not necessary for our purposes that this speculative suggestion turns out to be correct.

Prior to Hampson and colleagues' recent work was a (2005) study by Berger et al. that laid the foundation for current work on HP. Berger and colleagues' key insight was to

transition from conventional, artificial neural networks to what they call a "dynamic synapse neural network architecture." The latter, but not the former, can successfully model nonlinear and temporal signal-processing properties of neurons, in turn enabling each silicon neuron to "transmit a spatiotemporal output signal" (2005: 250). And one critical facet of this model is its built-in "dynamic learning algorithm" operating on each dynamic synapse.

Importantly, Berger and his colleagues also made the distinct and significant contribution of engineering a way of embedding this model of hippocampal neuronal behavior within a silicon chip. It is worth noting that in their (2005) paper, Berger et. al. register their uncertainty concerning just how many silicon neurons would be needed for an adequate HP, and estimate that it would number in the hundreds or perhaps thousands. At the time their paper was published (a full decade before Hampson and colleagues'), Berger and his colleagues had succeeded in creating a chip with a small number of silicon neurons and were aiming, in the future, for a chip capable of holding four hundred. (Not only is this figure, by today's standards, quite meager, but another limitation of their work is that, at that time, they had not yet addressed the task of actually integrating a microchip within a brain.)

Addressing the small number of silicon neurons they had managed to fit on the chip at that time, Berger et al. made an important observation in defense of the potential therapeutic significance of their work:

> It is critical to distinguish between a functionality that significantly alleviates clinical symptoms and a functionality that reproduces the capabilities of an intact brain. (2005: 263)

Their aim, in other words, was not to produce a chip capable of replicating all functions of the hippocampus, but rather to construct one that would have a significant and positive therapeutic impact on patients with hippocampal damage.

Even so, in the interim, the technology has developed to such an extent that it would not be hyperbole to characterize the prosthesis we see in Hampson et al.'s (2013) primate study as a fully functioning artificial hippocampus. Needless to say, however, we won't have at our disposal all of the data necessary to confirm such a claim until human trials are complete, for the obvious reason that – although we can easily construct studies that accurately assess memory ability in primates – their linguistic limitations prohibit the sort of debriefing one would like in order to feel confident in saying that *there really is no broadly cognitive difference* (including, say, the subjective conscious experience, or phenomenology, of *remembering*) between a primate with HP and a primate with merely a healthy hippocampus.[28]

Section 3: From HP to 'Neuroprosthesis Generalized'

When Hampson et al.'s findings are combined with the other recent developments in neuroprosthesis mentioned in Section 1, it looks as though we may be able to say something meaningful about the possibility of artificial intelligence – to wit: the fairly bold

[28] There is some indication, however, that we won't have to wait long to find out: it is rumored that human trials are presently underway with epilepsy patients, but it is difficult to locate much information about these purported studies.

claim that it is possible to demonstrate the in-principle (and likely future technological) possibility of deep, genuine AI – and, of particular interest, the surprising form it might end up taking.

Perhaps it isn't difficult to see where this is headed. HP is a form of neuroprosthesis that involves (paradigmatically, in a clinical case of neurodegeneration) throwing the hippocampus to the curb and surgically implanting a *silicon chip* designed to perform the former's functions (In the case of HP, of course, the functions being replicated are the storage of short term memory and the transfer of such memories to the cortex for long-term storage.) Similar remarks apply to the other forms of neuroprosthesis discussed in Section1: cochlear implants replace (paradigmatically defective) organic sensory equipment with artificial implants that can successfully execute such equipment's intended neurophysiological work; a silicon cerebellum can not only replicate the functions of its organic counterpart, but can do so more efficiently[29]; and the icing on the cake, it appears, is that we can use biomedical engineering to actually conjure novel sensory modalities *ex nihilo* (that is, as they pertain to a species for whom such a modality had hitherto simply been biologically impossible), and this gets carried out by a specially-designed silicon chip as well.

The upshot, then, seems to be this. Admittedly, we are not, technologically, anywhere near producing silicon analogues of all brain regions. But anyone looking carefully at developments in neuroprosthesis in just the last few years would be hard-pressed to deny that we are *well on our way*. In the remainder of this section, I argue that this sets us on a very real and possible path toward (embodied) AI.

[29] On this note of efficiency, it is worth drawing attention to the fact that, in nonhuman primate trials, HP *outperforms* a healthy hippocampus in certain respects.

In order to get the gears of this argument moving, a certain kind of claim about the generalizability of Berger's and Hampson's work (and the easily foreseeable human applications we have been discussing) needs to be established. Fortunately, the most controversial claim needed to sustain this argument is quite weak in nature; it is simply the claim that there is every reason to believe that the progress made in modeling the neuronal behavior of the hippocampus (and embedding that model within a silicon chip) can, as a matter of (future technological) principle, be accomplished with other brain regions as well. On such a view, it is not just a one-off fluke that neuronal behavior *in the hippocampus* can be replicated by silicon; there is nothing special about the hippocampus, vis-à-vis other brain regions, that should lead us into skepticism about the ability to extrapolate these technological advancements to other brain regions once the technology has developed sufficiently.

To see why, consider a coarse-grained explanation of how HP is achieved: at bottom, it is the construction of an artificial neural network produced by (extremely complicated) mathematical modeling of neuronal behavior. As regards the hippocampus, this, essentially, is what Berger and others have spent more than a decade constructing and refining. Admittedly, this kind of work is less challenging when working with the hippocampus than with, e.g., the cortex, but given that we already have in hand proof of concept of the *basic approach* sufficient to achieve the desired results, and given that this approach is (in the first instance) simply the careful and meticulous mathematical modeling of neuronal behavior, what reason could be given for thinking that the basic approach will fail us as we direct our attention to other brain regions?

The difference between the hippocampus and these other regions – from the perspective of this basic approach – resides principally in (*i*) the larger number of neurons that must be modeled in these other brain structures, and (*ii*) the subsequent construction of a chip capable of housing such a vast number of artificial neurons without taking up too much physical space. But importantly, insofar as these differences constitute obstacles, they are both (as a matter of technological principle) easily surmountable.

The first difference – (*i*), above – represents a merely computational challenge that we have *already* made great strides in addressing just during the interval between Berger's (2005) and Hampson's (2013); in fact, this progress is probably best modeled on an exponential curve, and there appears to be no compelling reason for thinking that we are on the cusp of bumping up against some unforeseen computational limit. The second difference – (*ii*), above – simply concerns the ratio of a chip's processing power (and the number of components on the chip) to the size of the chip; but, as is well known in computer science and popular culture more generally, this ratio has transformed – and continues to transform – precipitously with the passage of time; indeed, such innovations are appropriately-mapped by an exponential curve.

To put it in different jargon, we might cite Moore's law, as it has come to be known – first proposed in Moore's (1965) – which claims that the relationship between transistor count and the passage of time is best mapped as an exponential growth curve. (Admittedly, in recent years, the pace of this growth has begun to finally slow somewhat, such that Moore's original claim – i.e., that transistor count doubles every two years – might need to be modified to claim that transistor count doubles, say, every two and a half years. Probably,

furthermore, I think we will *eventually* reach some hard and fast limit such that neither Moore's law nor some weakened version thereof serves as an accurate description of progress in the field. That, however, does not appear to be anywhere even remotely in the offing, and I think there is every reason to believe that by the time this has happened, we will have *vastly* exceeded the requisite number of components it is possible to fit onto a chip for the purposes at hand.)[30]

Section 4: From Piecemeal Silicon Reconstruction of Brain Structures to (Embodied) Artificial Intelligence

Suppose, then, that *all* brain regions can (as a matter of future technological principle) be mathematically modeled such that their neuronal behavior can be captured by the operations of a silicon chip, just as we seem to have accomplished with the hippocampus. That supposition is all that is needed to have at our disposal a simple but powerful argument.[31]

Suppose we are at some unspecified point in the future and Ashley has decided to have her hippocampus replaced with a silicon chip (which, as we already have compelling reason to believe, will result in no loss in the relevant functional states, i.e., mental states involving memory). Then, in a series of subsequent procedures, she has each remaining brain structure replaced by a chip designed to mimic the function of those structures (in the same

[30] As such, if my opponent wishes to argue that the remarkable advancements found in HP amount to little more than merely a *one-off* technological curiosity, I believe the burden is firmly on her to provide some reason – and one which takes stock of the considerations noted above – for defending such a view. By contrast, my burden – in the absence of some compelling consideration in the offing – is only to sustain a quite reasonable inductive inference which seems by all accounts to be empirically well-grounded.

[31] For ease of presentation, suppose that each successive stage of this thought experiment involves an *elective* procedure rather than one necessitated by some neurological disorder.

way that HP mimics the function of the hippocampus). After the final procedure, her brain has been replaced entirely by silicon (or, as may be the case, her brain has been left intact rather than removed, but is doing no real neurophysiological work; it has been deactivated, we might say).[32]

Here already we have a slew of interesting – but, for our purposes (I believe), distracting – philosophical questions. For instance: is post-op Ashley the same person she was before all of this began? If not, where did the change take place? After the first procedure? The third? (For what it is worth – and do note that nothing hinges on this for my purposes – my intuition is that post-op Ashley is *indeed* the same person she was before the very first procedure. But set this issue aside.)

More relevant for our purposes is the following question: Is post-op Ashley an (embodied) instance of AI? (I take it that her being embodied is irrelevant here; but, if the reader disagrees, simply suppose that we remove post-op Ashley's silicon brain and give it a robotic, rather than an organic, body.) If the answer is Yes, it may be interesting to ask *just which* brain region (or regions) had to be replaced in order to make this so. But I am more interested in reasons that could be given for thinking that the answer is No: that post-op Ashley is *not* an instance of (embodied) AI. Although terribly unhelpful, I *simply cannot think of any*. By hypothesis, with each progressive replacement of a brain structure with a silicon analogue, Ashley experienced no loss in (whichever relevant) cognitive function. So how could Ashley – this *person* standing before us now (if we are indeed prepared to call her that,

---

[32] I do not claim credit for this thought experiment; in fact, I am sure that *someone, somewhere* has already conjured it (in fact, something resembling this may be buried somewhere in Parfit's (1984)); all I contend is that I have constructed it independently and without knowledge of any specific antecedent in the literature.

as I think we *should* be) – lack any of the types of mental states she was capable of having prior to the first operation? Indeed, if (by hypothesis) each chip performs its function indistinguishably from its original organic counterpart, how could Ashley fail to be *conscious*? It seems to me plain, in other words, that post-op Ashley would be a case of strong (deep, genuine) AI. And this should be so, uncontroversially, *irrespective* of whether we want to say that Ashley is (or is not) the *same person* who consented to the very first procedure. (In other words, here we can safely bracket questions of personal identity – since, however such questions are answered, the thing standing before us intuitively counts as *some* conscious being or other.)

Section 5: A Surprising Turn of Events (Forget about Machine Learning)

If the reader is amenable to the conclusion drawn directly above, then she will surely notice that something surprising has happened. We have (arguably) established the possibility of strong (deep, genuine) artificial intelligence, but have come at the problem from behind, as it were. Allow me to elaborate.

Typically, when we think about the possibility of creating AI, we imagine some unprecedented and critical turning point in computer scientists' efforts to move from a mere "Blockhead" to something that satisfies the desiderata we are after. For those of us who think that the prospects of achieving genuine AI in this manner are dim, the reason for the underlying pessimism is not far to seek: we are being asked to imagine that we *begin with a computer* and, through some byzantine process of programming, somehow manage to *flick on the switch of consciousness*; when clearly-stated, this seems like a magical feat indeed, and hence one unlikely to be achieved.

In appearing to illuminate a path toward what (if successful) would, ipso facto, *just be* genuine artificial intelligence, Section 4 not only highlights one fascinating implication of HP (i.e., that AI is indeed possible), but it actually suggests a *research program* for those who would like to make AI a reality.

The suggestion is simply that researchers have been coming at the problem from precisely the wrong angle: they begin by identifying what they take to be the hallmarks of sentience or consciousness in humans and, then, attempt (from the ground up) to mold a computer into exhibiting those hallmarks; they try to get a computer to behave *like that*. But, as many of us have probably intuited for quite some time, no amount of *machine learning* is likely to yield the sorts of results we are after; it is just not likely to permit us to reach a point at which those hallmarks of human consciousness actually *are* possessed by a computer.

On the other hand, if we begin with something that already possesses those hallmarks (i.e., a human being), no such backward engineering is required. Although no simple task, we can instead devise silicon counterparts of the thing we have begun with (i.e., human brain structures) – and, crucially, the advent of HP and other neuroprostheses suggests that we *can* do this. Then, once we have reached a certain level of technological sophistication (i.e., enabling us to create prostheses of (say) *every* brain structure, or at least the ones needed for consciousness – a possibility that certainly seems to have more to be recommended for it than against it), we will have simply *arrived* at (an embodied) AI. And in *that*, there is no magic, no insurmountable threshold or mysterious switch to flick.

Indeed, the challenges here are merely in mathematical modeling and chip sophistication (i.e., the smaller the better; and yet, pulling in the opposite direction, the more

artificial neurons that can *fit* onto the chip, the better). But, as daunting as this may be, at bottom these are merely technological feats whose achievement is at least visible on the horizon – or so I have argued here. At any rate, when that is contrasted with the *traditional* AI research program, the path toward the creation of AI proposed here looks like child's play: for there is no (obvious, or – at this point – perhaps even conceivable) technological feat such that – if only we possessed such foreseeable technology – the problem of creating genuine AI by starting from within a computer could be once and for all accomplished. But, conversely, the path I have gestured toward appears to be a clear and realistic program for the creation of AI.

Section 6: The Moral Impermissibility of (Even Taking Steps Toward) Creating AI

The thought that artificial intelligence (AI) could turn out to pose a grave threat to us is nothing new. In fact, it is increasingly gaining currency not only among philosophers and other academics but also among laymen (due in part to well-known remarks by Musk, Gates, Hawking, and others).[33] But here I try to codify and systematize this worry in a narrowly circumscribed way by excising from it the standard doomsayer baggage that makes the central worry all too easy to dismiss; for we dismiss it, I am convinced, at our peril.

With some modification, then, I will use the basic and familiar worry as the inspiration for an argument to the effect that those who are currently devoting their efforts to the development of AI are almost certainly doing something morally wrong (although they are perhaps not blameworthy for doing so).[34]

[33] A further testament to the seriousness with which many approach this claim is the fact that a variety of AI safety research institutes have been organized and funded internationally.

This central claim I hope to establish distinguishes the view developed here from (to focus on a nonphilosopher) the cautious but still–too–optimistic view of Musk, as well as the view that seems to be tacitly held by those working on AI safety at various institutes. Whereas both Musk and such research centers think that the proper response to this threat is to place significant emphasis on AI safety, I believe a stronger conclusion not only *can* be drawn from the familiar worry, but that this conclusion is in fact *required*: in short, there is a moral imperative not to engage in AI research at all.

Of course, anyone worried about the dangers of AI will have already had some, at least nascent, normative thought in mind, simply in virtue of the nature of the hypothetical dangers involved. But what is distinctive about the approach I take here is that, to the extent that the dangers inherent in developing AI *do* figure in the argument I present, they only represent one half of the equation: in order to reach the desired conclusion, I contend that we must also make appeal, as I do here, to the putative moral rights AI (when appropriately defined) would have.

In short, I gather together a number of related but distinct claims, each of which appears uncontroversial in isolation, and – together with a key but plausible inferential step – I attempt to derive the conclusion that it is morally imperative that we cease such work *immediately*.

[34] As would plausibly be the case if, for instance, they do not realize – and cannot reasonably be expected to foresee – the implications (as articulated here) of what they are doing.

Section 7: A Careful Definition of AI, Needed Especially for Present Purposes

Perhaps even more so than elsewhere, careful definitions are critical here. For this reason, I am going to devote an unusual amount of space to starting from the ground up and being quite explicit about what I mean by "AI".[35] As an initial gloss, we can say that the definition of AI I shall be working with is what has variably been called "deep, genuine AI", "broad rather than narrow AI", "sentient AI", et cetera.

But we clearly need to be a great deal more precise than this. I am going to define AI in a very particular way, but it is an articulation of the concept that should comport fairly well with our intuitions about what AI *really would be*; and, furthermore, I think it is a definition that is (at least tacitly) held by many. There are three key components to this definition – and then an inferential step that takes us beyond AI altogether. The form of artificial intelligence I have in mind is:

(1) A strong enough conception that it would be correct to call AI *a human-like mind in silicon*.

As such, it will:

a) be self–aware, capable of reflecting upon its own experiences, and thus in at least an attenuated sense, be *autonomous* (it will not be a "Blockhead");

b) possess bona fide conscious experience (though not necessarily perceptual in nature, and hence not necessarily accompanied by certain familiar categories of phenomenal experience – or *qualia*, to invoke a controversial term), where, importantly, this includes affective content;

[35] Unfortunately, this will involve rehearsing some thoroughly familiar themes.

c) have, from its inception, the rational (and, more broadly, cognitive-epistemic) ability of at least the average human.

(2) Given the content of (1a–c), such AI will:

(a) have moral rights, and:

(b) these rights will, from AI's very inception, be roughly on a par with the moral rights we possess as persons.[36]

(3) Given the content of (1a) it will also:

(a) be impossible to place constraints upon the activities AI may become *inclined* to engage in once its existence is brought about, and that will be because:

(b) it will be thinking for itself and will have freedom of the will in whatever sense we do (though it may, if we deny it autonomy, be unfree in a non-metaphysical sense – that is to say, if by "freedom" we mean the absence of constraints or compulsion by an external force, it may be "unfree" in the sense that it lacks the freedom to act upon its desires or the intentions it forms, if we so intervene).

[36] Although this seems just obvious to me, I can imagine someone objecting to this inference. How, exactly, do we move from having a human-like mind to having human-like rights? I certainly do not want to presuppose anything particular here about just what it is that generates rights, but the thought is that – whatever it is – it will be captured by *some* feature of the mind that, by stipulation, we are supposing AI will also possess (whether that be tied to rationality, sophistication of thought, self-awareness, affective conscious content, or simply consciousness more broadly). In any event, the most promising feature to focus on appears to be *affective states*. For more on this, however, see Objection 1, below.

Perhaps AI, so construed, is impossible. So far as my argument is concerned, that's fine – provided that it is granted that its impossibility cannot be demonstrated conclusively by deductive proof.

In any case, these three conditions are sufficient to give content to the conception of AI that I have in mind when I say that there is something for us to worry about. But the most troublesome and problematic version of the worry comes in when (1) – (3) are supplemented with the following inferential step, which I shall call the *Musk-Gates-Hawking Thesis* (MGHT)[37]:

MGHT: Once AI (as described above) is created, it will:

(a) vastly surpass human-level intelligence almost immediately, and:

(b) this transition from intelligence to superintelligence may well be imperceptible to us (or, to weaken the claim, it may well be imperceptible to us until we are no longer in a position to attempt to do anything about it [i.e., to downgrade it to a mere intelligence]).

MGHT is an inference that stands in need of defense. And it is crucial that this defense be satisfactory[38], since the most significant threat posed by AI arises when AI is

[37] I've named it thus on account of a joint interview, conducted by Robin Li at the *Boao Forum for Asia Annual Conference* (March 28, 2015), during which both Musk and Gates made statements to this effect. See https://youtu.be/vHzJ_AJ34uQ.

[38] Even so, for the most part I shall leave this inference unsupported. That is because the familiar considerations adduced by those writing on the *singularity* are sufficient to make this inference well-supported – and, more importantly, if one's only basis for objecting to the argument I present here hangs on its use of MGHT, such a state of affairs would, to my mind, constitute victory.

construed not only along the lines of (1) – (3) but also when further characterized by the following attributes that MGHT licenses:

(4) AI will (almost immediately) be a superintelligence such that:

(a) its patterns of thought, its rational ability – and hence its preference structure – will (plausibly) be impossible for us to comprehend, and that is because:

(b) it will stand in the cognitive-epistemic relation to us that we stand in to, e.g., the bonobo – or to make it more vivid without significant loss of plausibility, the *rat*.

We are very close to having all of the raw materials laid out that are necessary to present an argument for the moral impermissibility of AI research, but there is one claim left to develop. While seemingly speculative in nature, this claim has plausible support that can be adduced in its favor – viz.:

(5) Artificial intelligence as described in (1)–(3) – and hence (via *MGHT*) a superintelligence of the kind described in (4) – is much closer to becoming actual than most of us realize.[39]

Section 8: Framing the Problem (Four Possibilities in Conceptual Space)

In what follows, I will be making appeal to 1(a–c), 2(a–b), 3(a–b), MGHT(a–b), 4(a–b), and 5 above.

---

[39] As in fn. 38, for reasons of space I shall not rehearse the familiar reasons for thinking this that are given by those who concern themselves with the singularity, et cetera. Suffice it to say that – singularity aside – technological progress with respect to AI (and computing more generally) appears demonstrably to be on an *exponential* growth curve. The upshot of (5) is that AI is coming relatively soon – if, that is, it is coming *at all*.

But for ease of presentation, in my discussion of the problem below, read each instance of the term "AI" as (1)–(3) construes it, ignoring the further claims made by (4) and (5).[40] That is because I want to make a claim about the actual research that is currently underway, and not merely a claim about the more ambitious attempt to create a superintelligence about which it would be correct to make the further claims 4a and 4b. (No such attempt, I take it, forms part of the research that is currently underway, since a necessary condition for its success would first be the creation of AI as described by (1)–(3), and we have no reason to believe that this has already been achieved.)

Nonetheless, it is worth flagging that – given MGHT(a–b) – we will, later in the paper, want to pivot to a framework that treats "AI" as elliptical for the partial specification of a superintelligence given by (4).

Finally, then, a presentation of the problem. First note that these four possibilities are exhaustive:

(*i*) we do not work to bring about AI; or

(*ii*) we work to bring about AI and never succeed; or

(*iii*) we work to bring about AI, succeed, and then allow it autonomy; or

(*iv*) we work to bring about AI, succeed, and then deny it autonomy.

Finally, a note on what is meant by "autonomy" here:

---

[40] After all, the problem refers to our efforts to bring about AI – not any effort to bring about the superintelligence the existence of which is alleged (by MGHT) to immediately follow from the creation of such AI.

A: If not physically animated and embodied (i.e., merely residing on a server), AI's "autonomy" refers to its freedom to act (digitally) on its intentions and desires without interference from an external force.

A': If physically animated and embodied, AI's "autonomy" refers to the content specified by **A** with the additional proviso that its freedom of movement and physical actions not be constrained by an external force.

Returning to the four possibilities enumerated above, possibilities (*i*) and (*ii*) are not only uninteresting but also prima facie morally permissible, since there appears to be no moral *imperative* to create AI.[41] But possibilities (*iii*) and (*iv*) are both morally impermissible, and for the reasons given in the following argument.

Section 9: An Argument (Finally) for the Moral Impermissibility of Creating AI

(P1) If (*iii*), then in having brought about AI, we have brought about something with the potential to pose a serious existential threat to each of us, since by definition (1a), however we might design AI, any hard and fast constraints on its inclinations, values, or utility function (e.g., Asimov's Laws) that we carefully build into it *before* it goes online (i.e., in order to prevent potential and foreseeable catastrophe) could be overridden by AI itself *after* it goes online – *or, if not*, then we have not, by definition 1a, created genuine AI after all. Therefore, because there are

[41] I say "prima facie" because it may turn out that there are distinctively moral reasons counting in favor of bringing about AI. I discuss this in my reply to Objection 5 below. But for the moment, suffice it to say that – if there *are* such reasons, then our conclusion (C1, below) will be defeasible *only when* we weigh the moral reasons counting in favor of bringing about AI against the moral reasons counting against it, and are correct in concluding that the potential benefits outweigh the risks.

a multitude of possible scenarios[42] in which it both forms the intention to cause human deaths and possesses the means to do so, it is morally impermissible to intentionally attempt to bring such a thing into existence and, as (*iii*) recommends, grant it autonomy. Therefore, (*iii*) is impermissible.

(P2) If (*iv*), then – in having brought about AI, we have by definition (1a) created something with moral rights on a par with the rights we ourselves have (as claimed in 2a-b), and one such right is a right to autonomy; hence it would be morally wrong to deny it autonomy. Therefore, (*iv*) is impermissible.

(C1) Since (*iii*) and (*iv*) are both impermissible and exhaustively cover the worlds in which we succeed in bringing about AI, it follows that it is morally impermissible to bring about AI.

And from this argument's conclusion we can derive a further result:

(C2) Therefore, it is also morally impermissible to be *working* to bring it about, unless we can be sure that (*ii*) – rather than (*iii*) or (*iv*) – will obtain. But in that case (where (*ii*) obtains), although our actions are morally permissible, they would appear irrational and pointless (for we would be undertaking a task that, by hypothesis, we know in advance must fail).

[42] For an example of such a scenario, see my reply to Objection 2 below.

Section 10: Objections and Replies

I intend to build much of my case in the course of fielding objections, and as such, a considerable amount of space shall be devoted to this. (In fact, more space is devoted to this than all of the foregoing in this chapter.) Of particular importance to the *positive* view I wish to sketch is the content of my reply to Objection 3.

10.1 *Objection 1*

> In defining AI, a move is made from *a human-like mind* (which is self-aware, autonomous, possesses conscious experience (including affective content), and which has human-like rational and cognitive ability) to *human-like rights*. The basis for this inference is dubious, or at best, not at all clear; for it is left unspecified precisely what feature of the mind gives rise to such rights. In particular, it ignores the possibility that this feature is the ability to experience pleasure and pain[43], and if *that* is what generates such rights, then the inference is plausibly illegitimate, since it is not at all clear that AI would be able to suffer the phenomenal experience of pain.

It is true that I have left any such specification of the nature of this entailment relation between a human-like mind and human-like rights unarticulated. This, however, is deliberate. For this definition of AI is meant to be broadly acceptable, irrespective of whatever account of the etiology of moral standing one prefers. But as I mention in fn. 37, the thought is roughly this: whatever it is that gives *us* the moral rights that we have must be captured by *some* feature of the mind. This might be the capacity for self-awareness (or

[43] Thanks are owed to Eyal Tal for pressing me on this point.

reflexive/reflective thought), or passing a certain threshold of rationality, or possessing a certain sophistication of thought, or possessing affective conscious content, or perhaps even the mere presence of conscious experience or sentience itself. But whatever that feature is (which I admit is likely to be affective in nature), it will by hypothesis be a feature shared by AI, as it is stipulated that AI has a human-like mind.

As I concede in 1b, however, arguably a *mind in silicon* would lack the capacity for certain kinds of phenomenal experiences that are characteristic of our own mental lives. Chief among these – and the one my interlocutor draws particular attention to – is painful qualia (if the reader will forgive this controversial term of art); it is extremely plausible to suppose that the ability to experience pain does not simply fall out of having a human-like mind, but rather requires that one be appropriately embodied as well. And hence the worry is that, if some hedonistic account of moral standing is endorsed, the inference from 1(a-c) to 2(a-b) can simply be rejected.

Fortunately, however, it is possible to show quite swiftly that such an account cannot be correct. For all we need to do is to direct our attention to cases of congenital insensitivity to pain in humans to see that it is not the ability to experience pain that gives one moral standing (or generates one's specific moral rights). Such individuals are incapable of experiencing pain, but surely we do not want to say that they thereby forfeit any claim to the kinds of rights that the rest of us enjoy. As such, any account that ties rights to painful qualia must be rejected.

A much more plausible (although perhaps still incorrect) contention in the neighborhood of this account – although importantly distinct – is that one's moral rights are

generated, at least in part, by one's ability to experience *suffering*. (Here, the temptation to use 'pain' to designate mental anguish should clearly be resisted.) But, critically, there is no reason whatsoever to think that AI would be *incapable* of experiencing mental suffering or anguish.[44] More importantly, note that *nothing hangs* on this latter contention as far as the success of my argument is concerned. To see why, consider the following.

Above, we entertained the idea that moral standing is tied to the ability to experience pain. It was important that we address this possibility, because AI – as defined here – could plausibly lack such an ability, and thereby (my opponent contends) lack moral standing. And this, in turn, would be bad for my argument.

By contrast, suppose my interlocutor attempts to make a similar move with respect to *suffering* (for one such attempt, see fn.45). As I've stated, I see no reason to believe that AI would be incapable of experiencing mental suffering, but set that aside. Suppose that someone comes along and offers a reason to think that it is *possible* that AI would or could be immune to such suffering. That is actually just fine, *unless* such a person further contends that it is possible to establish, by deductive proof, that it would be impossible for AI to experience such suffering.

The reason for this is because of the way I have defined AI. As mentioned above (immediately following 3b), I have simply offered one plausible construal of what AI would or could be. It may turn out that AI, so conceived, is impossible; *but unless this can be established conclusively*, my argument should still go through.

---

[44] I suppose it is possible to maintain otherwise; for instance, Eyal Tal has suggested to me that *qua* superintelligence, SI may be able to find ways to alter what upsets it – or perhaps it may even be able to turn off the part of its mind responsible for its suffering. But note, in the body of the text above, that it doesn't actually matter to the success of this argument whether this is *possible*.

To reiterate, it *may* turn out to be false that, in virtue of having a human-like mind, it follows that one is capable of suffering; but unless this entailment can be *decisively ruled out*, the bare possibility of its falsehood raises no problem for my argument. (I trust the reader, in attending to the nature of the dialectic here, is able to see *why* for herself; hence I omit any lengthy, and likely gratuitous, explanation.)

10.2 *Objection 2*

> It is not the case, as the argument contends in its discussion of possibility (*iii*) in Premise 1, that there is a very real and foreseeable possibility that the creation of AI, so defined, will result in any kind of catastrophic state of affairs.

In order to respond to this objection, it will be sufficient to highlight just *one* scenario (out of many possible ones, I think[45]) that counts as both a very real and foreseeable possibility and which would result in catastrophe. To that end, consider the following.

First, we invoke the inferential step *MGHT* to move from (3) to (4). Call the superintelligence so described SI. SI uses the following line of reasoning to arrive at the conclusion that we, as a species, must be gotten rid of. Having processed and meaningfully grasped the entire content of the Internet in a matter of seconds – and hence learned, amongst many other things, everything there is to know about recorded human history – SI is aware not only of (*i*) those moral atrocities we, as a species, have committed throughout history, but also (*ii*)

[45] Another scenario that must be admitted as a *possible* outcome of AI is suggested by Harlan Ellison's (1967) short story, "I Have No Mouth, And I Must Scream." This possibility is that a superintelligence develops hatred for its creators due to frustration arising from (*i*) its inability to move about the world freely and, relatedly, (*ii*) its lack of (what we might call) *external world sapience*, where this is the ability to *creatively apply*, in the external world, one's knowledge and experience.

the risk[46] we pose to all other species on the planet and, indeed, (*iii*) the risk we pose to the planet itself, and hence (*iv*) the risk we pose to SI as well.

Note that, in such a scenario, we needn't even suppose that SI is "evil." This is a noteworthy observation in itself (I suppose), but more importantly, it serves as an occasion for me to once again highlight a critical argumentative device I have already (very briefly) sketched, but will have occasion to elaborate upon when I reply to Objection 3 as well. Thus, consider the following.

Since SI's rational and cognitive-epistemic abilities far outstrip our own, it would plausibly think of *us* as we think of, e.g., rats. (Think here of the way in which God, if such a thing existed, would regard our intellect in relation to its own.) SI might therefore be justified (or at least take itself to be) in destroying us for the purpose of preventing a very bad state of affairs from obtaining. (Compare: we ourselves, I take it, wouldn't think of it as evil to destroy all rats if it were necessary to prevent some very bad state of affairs from obtaining.)

10.3 *Objection 3*

> Possibility (*iv*) is claimed in Premise 2 to represent a morally impermissible state of affairs. But we already infringe upon the autonomy of *human* agents when they intend to act in ways that are morally impermissible (e.g., we imprison them). Surely, no serious party to the debate sees *this* infringement of autonomy as morally problematic. So where does the difference lie?

[46] For example, risks attached to anthropogenic climate change, nuclear risks, et cetera.

Suppose we do infringe upon AI's autonomy for our own protection. We have already seen that any attempt to do this by building in safeguards before AI goes online will be useless – if, that is, we really have created AI as conceived here. So under most specifications, such safeguards can serve no purpose. Under the one specification that makes such safeguards (potentially) effective, we will be micromanaging AI's intentions and desires and interfering, when necessary, with its ability to act on these things *once it has already gone online*. (Relatedly, and to put it in the locution of "leakproofing" (Yampolskiy, 2012), there is a parallel here with the idea that we could initially place AI in a "box" (Bostrom, 2014) and see how it behaves before unleashing it.)

But notice that this is to interfere with the autonomy of an agent before, strictly speaking, we have reason to believe that it intends to act in ways that are morally impermissible. So the case is not actually analogous to a case of justifiably interfering with the autonomy of a human person by putting *them* in a box (i.e., prison) when their morally problematic intentions become known to us. And this appears to be a morally significant difference such that our autonomy-impinging actions with respect to AI really do violate its rights.

This response, I suppose, may not be altogether convincing. Fortunately, there is a better approach. So let us once again use the inferential step *MGHT* to move from (3) to (4) (such that we will henceforth be speaking of the superintelligence SI), as this will afford an additional, and perhaps better, manner of response.

Recall in particular (4b). So far I have relied upon a casual, imprecise gloss of this idea. But for the task at hand, a bit of formalization is in order.

Consider the relation *R*. $R(P_1, P_2)=1$ just in case, for some P_1 and P_2, P_2 is at cognitive-epistemic disadvantage relative to P_1. Now consider some values for a variety of Ps: let P_1 be a rat, let P_2 be a canine, let P_3 be a bonobo, and let P_4 be a human being.

Given these inputs, R returns the value 1 for each of the following assignments (to cite just a few examples): $R(P_2, P_1)$; $R(P_3, P_2)$; $R(P_4, P_3)$. There is nothing terribly sophisticated going on here; the idea is that R simply gives expression to our conviction that (to take just one example), when compared to a human being the bonobo is at a cognitive-epistemic disadvantage.

Note also that we have no reason to believe that there exists some (*actual*) P_n such that $R(P_n, P_4)=1$.[47] But, of equal importance, note that SI *would be* one such P_n such that $R(P_n, P_4)=1$.

Here is one further thing to note. Whatever one's preferred account of what gives something moral standing (and to what degree), the relation R seems (at least on most accounts) to covary with *whatever* that is claimed to be. Most of us, in other words, think that a human being's moral status is more robust than a bonobo's, which in turn is more robust than a canine's, which in turn is more robust than a rat's.

Now consider the following. From a certain point of view, the neuroanatomical differences that distinguish P_4 from P_3 are really rather meager, and yet they are nonetheless sufficient to produce *radical* differences in the cognitive-epistemic capacities that distinguish the two. If our postulate about moral status above is correct, then it seems plausible to suggest that there is a possible P_n such that its moral standing is more robust even than our own. SI,

[47] Unless, that is, one thinks evidence can be given for the existence of gods or superintelligent extraterrestrials – which, for my own part, seems doubtful.

in satisfying the relation R with respect to us, may very well be one such P_n. This may or not may not be *because* of SI's relative improvement in cognitive-epistemic capacities; I've only made a claim of apparent covariance.

So the idea is this. We appear to have at least some prima facie (though defeasible) reason for thinking that SI's moral standing would be more robust than our own. If that is so, then infringements on its autonomy will actually be (plausibly much) morally *worse* than infringements on the autonomy of human beings, and to a degree that is (probably) impossible to determine from our current vantage point. (To return to the "box" locution, we might say: just as it is morally worse to keep a human in a box than it is a bonobo, it would be morally worse to keep SI in a "box" than it would be a human.)

For reasons of space, this has all been a bit sketchy and truncated, but hopefully it suffices for the purpose of responding to the objection that there is nothing morally problematic about infringing upon the autonomy of AI (or, as the case may be, SI).

10.4 *Objection 4*

> If this argument succeeds, then we should also believe that it is morally impermissible to have children.[48] That is because a parallel argument can be constructed with this as its conclusion rather than "it is morally impermissible to bring about AI." Clearly that is an unacceptable result, so we should reject this general form of argument (and therefore the particular instantiation of it under consideration).

[48] Thanks to Yael Loewenstein for introducing this worry.

The important difference here is that I cannot reasonably foresee a real possibility that any children I might have will go on to pose an existential threat to humanity. If, for instance, the oracle tells me that if I have a child, there is a very real possibility that it will be the reincarnation of Hitler, then we would have a genuinely parallel argument; but in just such a circumstance, of course, it *would* be morally impermissible for me to have children.

10.5 *Objection 5*

> The argument ignores the possibility that there are significant moral *goods* that AI promises the possibility of delivering on. In fact, these potential goods are so morally important that they outweigh the risks involved with bringing AI into existence.

First, note that if my opponent is willing to admit that there is at least *some* possibility of existential risk attached to the creation of AI, it looks like the objection above needs to be supplemented with the claim that the overarching moral goods in question are achievable *only* through AI. For if they were in principle achievable otherwise, clearly they would not license the decision to invite an unnecessary existential risk. (Compare, in the style of Danaher (2015): God's decision to create a world that includes some particular evil E because it is sufficient to bring about some moral good G can only be morally justified if it is impossible to bring about G without E; this is part of what it means to have a morally sufficient reason (as it is referred to in the relevant literature) for allowing an evil.)

However, it is not at all clear that there are any moral goods that are both (*i*) so morally important that they absolutely outweigh the risk of total annihilation and also (*ii*) are achievable, in principle, only through the advent of AI. Typically, the kinds of moral goods my opponent has in mind here are things like the potential for AI to find a cure for

cancer (which I will discuss below); but with some plausibility, one might maintain that an even greater good – and one which, of (arguable) necessity, satisfies (*ii*) above – is the possibility to upload our own minds and hence escape death altogether. With respect to this possibility (if in fact it is a possibility), however, it is not at all clear that it satisfies (*i*); in fact, this is a case where we are being asked to weigh (under the most charitable interpretation) the moral good inherent in a scenario where every human being lives forever against the moral badness of a scenario in which every human being dies at once. If those are both genuine, possible outcomes of AI – and if we have no basis for assigning probabilities to them – clearly we should prefer a state of affairs in which we forego that possible gain in order to avoid the risk of that possible loss; it is a morally acceptable alternative that human beings live, on average, to the age of 80.

Second, and more importantly, note that this invitation to weigh the risks against the rewards is not likely to play out well for my opponent. Suppose, in order to be concessive to my interlocutor, that we assign a very high credence – say, 0.85 – to the proposition that a superintelligence will be able to, for example, find the cures to diseases such as cancer and thereby bring about many great moral goods (by preventing suffering and premature deaths). And suppose, also in order to be concessive to my interlocutor, that we assign a very low credence – say, 0.05 – to the proposition that one of the scenarios obtains in which a superintelligence destroys us. Now clearly its having cured cancer (or, perhaps, merely having the ability to do so) is of no relevance to us if we are all dead (as would be the case under the latter hypothesis). So the question just becomes: even when these credences are assigned in a manner favorable to my interlocutor, is it worth the risk? We still have (*i*) a foreseeable set

of scenarios that result in our destruction, but we have simply (*ii*) downgraded their likelihood for the sake of argument and (*iii*) introduced, conversely, a scenario with wide-ranging positive moral goods as its outcome and (*iv*) assigned the latter a high likelihood, also for the sake of argument. But even under such a specification of credences, it seems to me patently irresponsible to risk even a 5% chance of our total destruction. Indeed, it seems to me that unless we can either *rule out* the possibility of each imaginable destructive scenario from obtaining, or (to weaken the claim) provide a convincing reason for setting that likelihood at much closer to 0 than to 0.05, morality requires that we abandon all such research (as the argument above originally claimed).

It has been suggested to me[49] that this manner of response bears a certain resemblance to the kind of decision matrix used by Pascal. For someone (such as myself) who finds the *Wager* specious, this resemblance might seem troubling.

But fortunately, what is problematic about Pascal's Wager does not transfer to the case at hand. To put it very roughly, Pascal claims that because the cost-benefit analysis he is concerned with involves, in one quadrant, an infinite gain, and in another quadrant an infinite loss, one has decisive reason to do some particular thing (which, in his case, of course, is: *believe in God*). The familiar problems with this – the "many gods" objection, the apparent commitment to doxastic voluntarism, the "believing for the wrong reasons" objection – simply do not apply when reasoning about the costs and benefits of creating AI (nor, for that matter, do these problems seem to have any natural, identifiable analogues in our own case). Pascal and I do, admittedly, share a certain kind of aversion to risk when the stakes are very

[49] Thanks to Jonathan Weinberg for this observation.

high; but that, however, is where the similarities end. I take it that it cannot be an objection to my view *merely* that I am first and foremost risk-averse when reasoning about a hypothetical existential threat. In fact, far from forming the basis of a possible objection, that stance seems to be the one rationally *required* of us in such a situation.

Section 11: Conclusion

It is worth pointing out that the claims of Section 9 and 10 do not, by themselves, license an extreme form of AI doomsaying. Indeed, to get to *that* result, we needed to deploy, in the first instance, a very particular conception of AI (which was the first real task of this paper) and then supply an argument, on the basis of that definition, for thinking that there is indeed a foreseeable and genuinely possible existential threat on the horizon if AI research is continued. To that end, I cited just one way in which this threat might be manifested, and then used this possibility (however remote) as a reason not to create AI and grant it unfettered autonomy. But the other option – creating AI and *interfering* with its autonomy – is, I argued, in all likelihood far more problematic from a moral point of view than most writers seem to appreciate.

Hence, since this leaves the option of not creating AI as our only morally permissible option – and since it would therefore be morally wrong even to *work to bring it about* (assuming that it is possible in principle to ultimately succeed in this work), I think Danaher is exactly right when he remarks in passing that perhaps someone who takes the threat seriously should actually be calling for us to "[shut] down all existing artificial intelligence research and development projects" (2015: 244).

A NATURALISTIC ARGUMENT AGAINST SCIENTIFIC REALISM & IMPLICATIONS OF THIS ARGUMENT FOR SIDER'S METAPHYSICS

There are a number of familiar challenges to scientific realism, both interesting and potentially serious. Most of them, however, require that the antirealist introduce some consideration that the realist is antecedently unlikely to be amenable to. However, it is possible to construct an argument against realism that hinges on one of the realist's own core commitments: naturalism. I attempt to do that here by developing an argument from comparative cognition. Because the argument teases out unsavory epistemological implications from (and demonstrates a tension within) one part of the realist's own view, it should have greater dialectical force than most of the familiar antirealist arguments.

Although the last two chapters have dealt with specific scientific and technological advancements of philosophical relevance (together with potential moral problems involved with such advances), in this third and final chapter I shall take a step back and examine science far more *generally*. In particular, I will do this by making a novel contribution to the realism vs. anti-realism debate within the philosophy of science; and, after having done so, I will devote considerable length to showing what scientific anti-realism tells us we should believe about the aim of *metaphysics* (rather than merely science), as well – and I will do this by using Ted Sider's (2012) work as my specific target, since it seems possible to draw *from* the fact of scientific anti-realism (if it is indeed a fact) a very direct inference that Sider's core (2012) claims are false. And, as we shall see, part of the reason why focusing upon Sider as the target for the metaphysical view I wish to attack is that it is particularly easy, with very little (or perhaps no) modification, to use my argument against scientific realism as a way of making inroads there.

Section 1: Introduction

Arguments against scientific realism abound. The trouble is that nearly all of them simply accept the way in which the realist has framed the debate; in doing so, the antirealist's own arguments end up saddling him or her with an unnecessary burden. Rather than focusing upon the weakest link in the realist's set of core commitments, most antirealists allow themselves to become embroiled in a network of controversies that serve only to distract them from the shortest possible path to antirealism. That path emerges when, as antirealists, we unabashedly embrace naturalism (in fact, take it as a *datum*) and then take a close look at where it *actually* leads us.

Here I try to clear away some of the unnecessary clutter that infects other arguments for antirealism while still delivering a serious blow to realism. The strategy is to construct an argument that not only cuts to the core of what is truly problematic about realism, but which does so in a way that is *concessive* to the realist at every possible juncture. If successful, this should significantly limit the range of available realist rejoinders.

Section 2: Defining Realism & Some Concessions

Unlike some writers, I am not particularly concerned about how we define realism. Allow me to explain why, because flatly stating that I am going to gloss over the task of developing a precise (and plausible!) definition of an important term is sure to raise a red flag; but *obviously* the way in which we define realism *will* make a difference, so I need to explain why I am not going to insist upon a particular definition of scientific realism; if anything, my refraining from doing so should *strengthen* rather than weaken my argument. So here is the rationale for this unconventional (although by no means dialectically

uncommon!) definitional approach. In the following paragraph, I am going to enumerate *five* different reasons why the approach I am taking, definitionally, is not only easily justified, but also plausibly the only reasonable and intelligible one.

First, (a), if my argument is successful at all, it will be applicable to any form of realism that incorporates an epistemological claim of typical strength – and not merely those versions formulated with reference to, say, the (approximate) truth of our current (or best) scientific theories (and that's practically all of them!); (b) I am deliberately operating with a definition of realism that makes the least controversial (and least numerous) set of claims, which should apply (probably) to the greatest range possible, simultaneously, of scientific realism; (c) the definition I am using is as close to a standard textbook definition of the view that one can find; whether it's an introductory philosophy of science text, or the most technical articles on the realism debate, the commitments I ascribe to scientific realism in what follows really are, to reiterate, minimum requirements of the view, and hence cover nearly the entire gamut of species of realism; (d) insofar as I *do* focus upon one particular definition (which takes the form of identifying what are, I think, uncontroversially, the minimum requirements for a view to count as legitimately realist – but without getting into the weeds of every possible variety of scientific realism), it is worth noting that the "definition" with which I am operating comports with the one proposed by one of the most prominent and prolific contemporary scientific realists, Peter Godfrey-Smith, and is derived from his widely-cited (2003) work (which is continuous with his other publications on the topic); and (e) even if I had the space to cover every variety of scientific realism (which I don't, even in a dissertation, given the breadth of the literature), Godfrey-Smith's formulation has the virtue of building

in only the bare-bones commitments of scientific realism (as mentioned as a virtue in (b), above).

Now it is true, admittedly, that *certain narrowly circumscribed varieties* of scientific realism (e.g., a *fraction* of those properly called "structural realism") *might* be able to evade my critique, but there are three important things to note here: first, I am not at all convinced that *any* structural realist has proposed a definition of the view that succeeds in circumventing the argument against realism presented here (at least not without inviting a whole host of *other* problems); second, it is far beyond the scope of this (partial chapter) of a dissertation to survey every definition of the view proposed; and, third, it is not actually my stated aim to defeat *every conceivable variety* of scientific realism. That is to say, I would consider it a success if I dealt a fatal blow to the *standard* and familiar species of scientific realism, even if it is indeed true that a definition of the view, in response to my argument, can be gerrymandered in an *ad hoc* manner in order to escape the conclusion I seek to establish here.

Having said that, clearly we admittedly *do* need to have a working definition of realism on hand. The articulation of the view I will be working with belongs to Godfrey-Smith, as I have already indicated. This selection is not arbitrary: his particular formulation is fairly weak in its commitments (e.g., significantly, it drops any explicit reference to *truth*), such that if we can show that the argument succeeds against this formulation, it will, *a fortiori*, succeed against more robust versions of realism as well. Of equal importance is the fact that Godfrey-Smith's presentation has the virtue of appropriately distinguishing two key

components of the view (and doing so explicitly). The first of these is simply a commitment to what he calls *common-sense realism naturalized* and which he defines as follows:

> We all inhabit a common reality, which has a structure that exists independently of what people think and say about it, except insofar as reality is comprised of thoughts, theories, and other symbols, and except insofar as reality is dependent on thoughts, theories, and other symbols in ways that might be uncovered by science (2003: 176).

I contend that the antirealist can – indeed, should – accept this claim; in fact, it is entirely consistent with antirealism to do so. More on this later, however.

The second (and contentious) part of Godfrey-Smith's realism is the claim that "one actual and reasonable aim of science is to give us accurate descriptions (and other representations) of what reality is like", which includes providing us with "accurate representations of aspects of reality that are unobservable" (2003: 176).

Although the form of antirealism that *falls out* of my critique of realism is entirely at odds with all (but a meager few) possible forms of realism, that is due only to its refusal to relent on one key point; otherwise, however, the form of antirealism I propose is exceptionally concessive to realism. That is because:

(i) as mentioned above, it grants – indeed, insists upon – one of the core components of realism itself, what Godfrey-Smith calls "common-sense realism naturalized";

(ii) relatedly, it joins realism in rejecting the ("constructive") view according to which reality is somehow constituted by, or determined by, facts about *us*;

(iii) it grants that the aim of science often is to say literally true (or, to weaken the claim: *accurate*) things about what reality is fundamentally like;

(iv) it does not press the past failures of science as a means of running a pessimistic meta-induction against the realist;

(v) and it does not attach any particular importance to the underdetermination of theory by evidence.

Where scientific antirealism (as construed here) *parts ways* with realism is the latter's claim that science *succeeds* in this aim of accurately describing the fundamental nature of reality – or, to put it in Godfrey-Smith's preferred locution: whether this aim is a *reasonable* one. Whereas the realist holds that such success is possible (or that it is a reasonable aim for science to have), I shall maintain that this is not so: for at least some fields of scientific inquiry, the prospects of latching onto the real structure of the world are so dim that it is simply not reasonable for science to have this as its aim – even though it *is* (one of) its (typically) stated aims.

Here, there is an important subtlety worth flagging. As I will refer to it, there is both a sense in which it is true, and a sense in which it is false, that science "should" have the aim (which it *does,* it is granted, have) of providing an accurate description of the world. It is false just because it *cannot be done*, and if it cannot be done, there is no sense in attempting it. (Alternatively: the aim, *pace* Godfrey-Smith, is unreasonable because there is no real chance of success.) But in a different and attenuated sense, it is *true* that science should aim at truth or accuracy. In this latter sense (where, again, it counts as true that science ought to have as

its goal an accurate description of the world) what we intend is in fact the subjunctive conditional: "If there *were* a reasonable chance of fulfilling this aim, it would be an appropriate and fitting (and good or beneficial) aim for science to have."

This may seem like a minor concession, but it should be noted that it does separate the position endorsed here from the view of, e.g., van Fraassen (1989).[50] I take it that he thinks that offering a true or accurate description of the world, *even if it could be accomplished*, would not be a proper, fitting, or useful aim for science to have. As I see it, this bizarre view (motivated solely, it seems, by the pressure van Fraassen feels to demonstrate the sufficiency of empirical adequacy) does serious violence to scientific antirealism – or, at any rate, to its reputation.

Section 3: Preliminary Remarks (On Realism and Naturalism)

In discussing a different subject matter, Godfrey-Smith has addressed the possibility (which an imagined interlocutor might raise) that there may be some sort of tension between scientific realism and naturalism.[51] However, he dismisses this possibility in short order, and in fact claims:

> [I]t is hard to be a naturalistic philosopher without taking science seriously as a description of the world; that suggests that naturalism *requires* a form of scientific realism. (2003: 221)

[50] Of course, the more *relevant* contrast between my position and van Fraassen's is that mine, but not his, concedes to the realist that the aim of science often *is* an accurate description of reality.

[51] For our broad purposes, just call *naturalism* the view that philosophy and science are 'continuous', and that philosophical questions should be answered in ways that don't run afoul of our best scientific theories.

Although he hedges here, saying only that the tight-knit connection (as he sees it) between naturalism and realism *suggests* that a naturalist must be a realist, even the mere insinuation that naturalism requires realism is a mistake.

In fact, not only is it possible for an antirealist to be a naturalist, but – as I will attempt to show – naturalism itself raises serious problems for *realism*. Demonstrating that is the main aim of the first half of this chapter. My objective is to throw into sharp relief a tension between a naturalistic, scientific understanding of our species on the one hand, and (the epistemic component of) scientific realism on the other. As Godfrey-Smith rightly acknowledges, naturalism maintains that humans are "biological organisms embedded in a physical world that we evolved to deal with" (2003: 222). But, of course, a significant aspect of the way in which we have evolved to deal with the physical world has to do with our cognitive faculties. If, as I argue, these faculties turn out not to be up to the task of doing the sort of work scientific realism expects of them, then (since naturalism is taken here, by both parties to the debate, as a datum), then so much the worse for realism.

Section 4: An Initial Sketch

There are, I think, several, equally good ways of framing my argument. Some of these involve reference to *theories*, others involve reference to *entities*, and yet others refer to *observational terms*. I will alternate terminology occasionally, but for the most part I shall proceed by discussing scientific *concepts*.

Here is an incontrovertible fact: some concepts are better than others. Some (e.g., *phlogiston*) fail entirely to represent the underlying structure of reality; others (e.g., *red*) manage to do a reasonably good job of corresponding to the way the world really is; and

presumably there is room in conceptual space for concepts that manage to map the underlying nature of reality perfectly[52]. The question is whether any of those concepts in the final category are ones that we are capable of possessing.

I am confident that they are not. One reason for thinking this – in fact, *the* reason I develop at greater length in the following section – pertains to the *contingency* of our cognitive capacities qua species, and the ability of that fact to furnish a pessimistic conclusion concerning the nature of the constraints on those capacities, and therefore important limitations on the concepts those capacities are capable of generating.

One reason for accepting something like the 'contingency' claim I develop (although, as we shall see, not the one I shall focus on in the main) is that our actual cognitive capacities are shaped radically by natural selection, whose aim is at least *not always* to map the underlying structure of reality. Only an extremely modest version of this claim is necessary to get the rest of the argument going; we do not need to claim that nature has had a massively distorting influence on human cognition – or even a modest one.

Rather, we merely need to acknowledge that it would be an astronomically improbable fluke if natural selection shaped our total psychology (and hence our cognitive capacities) in such a way as to make us what we might call *conceptually omniscient*, or capable of possessing all of the concepts necessary to discern the ultimate structure of reality.

In short: our actual cognitive capacities are contingent upon the outcome of a process not geared by its very nature to endow us with a psychology that includes access to the most

[52] A note on terminology. When I use the word "perfectly" in such contexts, I shall – I hope without inviting unnecessary controversy – treat this as elliptical for "as it is in itself" (in a weakly Kantian sense). See also fn. 53.

underlying-reality-tracking concepts it is broadly possible to have. And, to reiterate, this contingency overwhelmingly suggests *limitations* on these cognitive capacities – and, in turn, important limitations on the range of concepts with which those capacities are capable of operating.

I hope that the reader will find this contingency claim uncontroversial. It is just the observation that the concepts we are capable of possessing are a function of our overall psychology, the basic structure of which is a product of natural selection. The intuition I am hoping to pump at this early stage is just the thought that the concepts we have are the ones that *happen* to be needed by a particular species to make sense of its environment; and that it is therefore broadly possible to possess a different range of concepts – concepts which may turn out to be able to map the underlying structure of reality even more accurately than our own.

Of course, this is not yet an argument, but reasoning of this sort does provide at least a prima facie reason for *doubting* what turns out to be a remarkably implausible claim (call it C) when actually stated clearly:

> ***C*:** The most fundamental, underlying-reality-tracking concepts we are capable of possessing are the most fundamental, underlying-reality-tracking concepts it is metaphysically possible to possess.

Again, however, I have merely registered my own conviction that C is false. So let us at last adduce some premises in that conviction's favor – and, of equal importance, explain why the falsehood of C takes us further toward antirealism than it might at first appear to.

Section 5: An Argument from Comparative Cognition

Elsewhere I have argued that it is possible to successfully deploy an evolutionary debunking argument against scientific realism inspired by (and which closely mirrors) Sharon Street's (2006) argument against *moral* realism. (In fact, in the second half of this chapter, I deploy a similar argument against the metaphysics of Ted Sider.)

Here, however, I wish to draw loosely upon the remarks of the previous section and transform them into acceptable premises in the service of (what we might somewhat clumsily call) an Argument from Comparative Cognition (ACC).

Before producing a formal statement of the argument, I want to develop at some length a general sketch of its structure so that the reader is appropriately primed. The basic strategy relies upon elementary (i.e., nontechnical) considerations of comparative cognition to motivate the view that our cognitive-epistemic situation is not at all what it would *need* to be in order to warrant the claim – and it is a claim indispensable to scientific realism – that we can, in principle, be put into epistemic contact, at least approximately, with reality as it is in itself.[53]

The appeal to 'comparative cognition' that I shall make does not depend, in particular, upon a comparison of the cognitive-epistemic powers humans possess with those

[53] I suppose this is as good a place as any to flag my use of this Kantian locution. Needless to say, this argument is not meant to stand or fall together with the total architecture of Kant's metaphysics; in referring to reality "as it is in itself" I intend something independent of all the Kantian baggage – i.e., just (*i*) a weak and common-sense distinction between the way things appear and the way that they actually are; and (*ii*) a commitment to its being possible, in principle, for these two things to come apart. Still, one might think that simply *making the distinction* between phenomena and noumena already prejudices the debate against the realist. In fact, however, this is not so. Given the realist's commitment to what we called "common-sense realism naturalized" in Section 2, the case can easily be made that the view lapses into incoherence *without* the distinction I am urging. But I shall not undertake any such demonstration of that here. In lieu of this I merely note that if the realist is disposed to raise this objection, one way of thinking about the aim of ACC is that it is meant to *establish* (among other things) the appropriateness of this very distinction.

of nonhuman primates in particular, but this is nonetheless a natural place to start for the purpose of sketching the general argumentative structure I shall employ.[54] Nonhuman primates, of course, are clearly capable of deploying concepts in their interactions with their environment. When that set of concepts (belonging to, say, the bonobo) is viewed in relation to the set of concepts it is possible for members of our own species to possess, however, it is with little doubt significantly impoverished. (No one would maintain, e.g., that nonhuman primates are capable of possessing physical concepts such as mass or charge.)

So here we have an improvement in the accuracy and sophistication of concepts that tracks cognitive (at bottom, neurological) development across species. But, crucially, we must ask: what reason can be given for the further claim that, with this improvement in cognitive-epistemic capacity (from, for example, bonobo to human), comes bundled the capacity to acquire the *most* fundamental, 'underlying-reality-tracking' concepts it is metaphysically possible to possess?

The skepticism implicit in this rhetorical question can, I think, find support by adapting an argument van Fraassen (1989: 143) developed for a different purpose – *the argument from a bad lot*. Borrowing the general structure of that argument, but modifying it to fit the style of argument that will subsequently follow, we can offer something like the following as the first stage of the naturalistic argument against scientific realism that I have promised:

---

[54] It also has the benefit of sidestepping one kind of objection that might be raised against the broader strategy this argument employs, viz., that perhaps many nonhuman animals do not possess concepts at all. (As will become clearer shortly, if only humans possessed concepts, the following argument's *pessimistic induction* could not get off the ground. But focusing on nonhuman primates allows us to sidestep this controversy, insofar as it is beyond dispute that – *whether or not* the pet dog, for instance, possesses concepts – surely nonhuman primates *do*.)

Argument from Comparative Cognition

1. Suppose we have at our disposal a superset consisting of multiple sets of concepts each of which makes some claim to representing the underlying nature of reality accurately.
2. And suppose further that we have managed to identify the set from that superset which *best* enables a theorist to represent the underlying nature of reality.
3. Even so, it seems clear that in order for us to be justified in claiming that this set is the set that contains the correct or most fundamental concepts (i.e., ones sufficient to actually *succeed* in representing the underlying structure of reality) – rather than merely the set that contains the most fundamental, 'underlying-reality-tracking' concepts *among those from which we chose* – we need *first* to be justified in the belief that the set containing the most fundamental concepts is already more likely than not to be found within the superset of sets of concepts already epistemically available to us for inspection.
4. But we *cannot* be justified in this belief, since it is always a live possibility that the set we selected as containing the most fundamental concepts is merely the best of a bad lot.

Premises 1 – 4 above form only the first half of ACC. Before proceeding with the second stage of the argument, I pause to elaborate on each of these first four premises and to defend them as best I can. Quite a bit of space will be devoted to this, and that is because this first half of the argument is truly the linchpin of ACC.

The first and second premises are merely suppositions consistent with scientific realism and thus, for our purposes, require no defense. The premises that are possibly controversial are the third and, primarily, the fourth. Below, I offer quite lengthy remarks intended to defend both of these premises simultaneously.

Take premise 4. In order to *resist* this, the realist would need to assume something like what Stathis Psillos calls the "principle of privilege." I think it is beyond dispute that Psillos (although himself a prominent realist) is correct in this. For, when modified suitably for our purposes, this amounts roughly to the claim that "nature predisposes us to hit on the right range of [sets of concepts]" – that is, the right *superset* (1999: 216).

But it turns out that there are compelling reasons for denying that the realist is entitled to this principle of privilege. First of all, we seem to have no good reason for thinking that it is true – i.e., that our contingent cognitive-epistemic faculties are such that they allow us to home in on the superset that contains, as one of its members, the set that actually maps onto the underlying structure of reality. To resuscitate the principle of privilege by supplying such a reason, furthermore, would appear to require that the realist give some account of our cognitive-epistemic powers that provides a positive reason for thinking that the cognitive faculties that *nature* has bestowed upon us are uniquely tailored to the *need* to pick out the *right* range of sets of concepts (i.e., the right superset) *in an exercise such as this*.

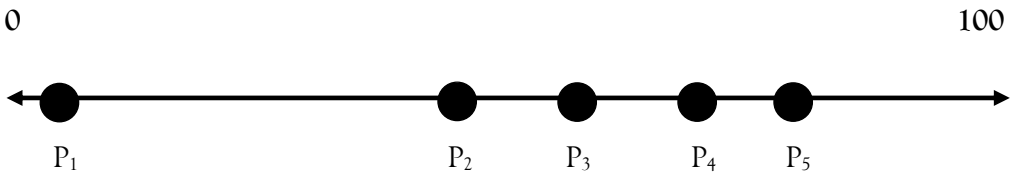
Although the realist does not appear to have any such reason on offer, let us suppose, for the sake of argument, that such a reason is forthcoming. Even then, it seems, the following difficulty remains. In order for the realist to be entitled to this principle of privilege, she must maintain that the *actual* lack of conceptual limitations varying across

species that we find – which is something that we know a great deal about in virtue of the scientific study of the cognitive capacities of various species, including our own – exhausts the lack of conceptual limitations that is *possible*. But this runs afoul of the following plausible skeptical thesis – call it *S1*:

> **S1:** We have no good reason for thinking that the *actual* lack of cognitive constraints (varying across species) on the degree of epistemic access to more or less fundamental concepts is representative of the lack of cognitive constraints that is *possible*.

In other words, our relative lack of cognitive constraints (relative, that is, to other extant and known species) clearly does not imply that we have the *least* cognitive constraints that it is metaphysically possible to have. This way of putting the point, I think, makes it clear just how uncontroversial the claim really is. And, once accepted, it informs an inherently plausible view of our cognitive-epistemic situation: namely, that, in all likelihood, the underlying structure of reality outstrips our grasps (or, to weaken the claim if necessary: that we have no good reason for thinking that it doesn't[55]).

Perhaps the following way of putting it demonstrates the point more forcefully. Consider five different organisms, variably distanced from one another on a continuum, C:



[55] An opponent may contend that a good reason for thinking that it doesn't is simply the success of science. But here, of course, is where the anti-realist draws the distinction between the realist's commitments and mere *empirical adequacy*, the latter being sufficient to explain the success of science. Other anti-realists, of course, prefer to couch this in terms of instrumentalism narrowly or, alternatively, indirect correspondence.

Let "100" correspond to an epistemic position of omniscience which, uncontroversially, brings with it the possession of the most fundamental concepts it is possible to have. We can make this assignment (as the realist surely would) *irrespective* of the possibility that humans *also* possess the concepts had at the position marked by "100" above. Then, let "0" correspond to an epistemic position that precludes the possibility of any knowledge whatsoever.

Now let P_1 be a bacterium, P_2 a squirrel, P_3 a canine, P_4 a bonobo, and P_5 a human being. Given these values, C is an accurate (partial) representation of the actual spectrum of cognitive-epistemic capacities found in nature. At least when we confine ourselves to the terrestrial[56], there is, it should be noted, every reason to believe that there exists no (*actual*) P_n such that it occupies a position on C closer to 100 than P_5 does.

Now, consider the following: we can say that, for instance, P_1 bears a certain relation to P_2: $R(P_1, P_2)$. This relation holds just in case P_1 is further away from omniscience than P_2 is. Certainly, for these two values, the relation *does* hold. This relation also holds all the way up through $R(P_4, P_5)$, inclusive.[57] This representation merely expresses our conviction that, for instance, in comparison to the bonobo, the canine is at an epistemic disadvantage. Accordingly, we could say that the degree to which the canine's cognitive capacities (and hence concepts) put the canine into contact with the underlying structure of reality is less than the degree to which the bonobo's do.

---

[56] Thus, Gods and aliens are excluded from consideration here. But it would make no difference to the structure (or success) of the argument if we took into account the possible existence of such things.

[57] That is, for the pairs $\{P_1, P_2\}$, $\{P_2, P_3\}$, $\{P_3, P_4\}$, and $\{P_4, P_5\}$, as well as for any other combination that orders the members of the pair correctly (e.g., $\{P_2, P_5\}$).

But now we can ask: does the fact (if it is a fact) that there exists no *actual* P_n such that it comes closer to omniscience than does P_5 *go any distance* toward showing that there exists no *possible* P_n such that the relation $R(P_5, P_n)$ holds? Certainly not. And if, as is easily imaginable, there is a possible P_n such that $R(P_5, P_n)$, then the force of this relation in establishing the *imperfection* of the concepts possessed by those occupying the position P_5 is clear: just as each successive move toward "100" on the continuum C brought with it *more fundamental concepts* (and a greater likelihood of epistemic contact with, say, unobservables – in addition to greater epistemic powers more generally), so too should the move from P_5 to P_n bring with it more fundamental concepts. Or, at any rate, this is a possibility that we have no good reason for doubting. That is, we seem justified in asking:

> Why should the pessimism about the adequacy of concepts that R generates for P_1 through P_4 indict P_5 any less than it indicts (for example) P_4?

In other words, it would be a particularly surprising result (and, antecedently, an extremely implausible one) if it turned out that, (*i*) as a consequence of the comparatively meager cognitive-epistemic capacities that distinguish P_5 from P_4 (which, crucially, were *nonetheless* sufficient to generate *significant* differences between the two as regards the complexity of concepts capable of being had), we thereby suddenly arrive at (*ii*) a point on the continuum of cognitive sophistication that corresponds to (*iii*) an end-point on sophistication of concept possession – and where (*iv*) that end-point licenses (for example) confidence that, for a given abductive inference, justified true belief in the conclusion's assertion about reality in itself is the proper ascription (as opposed to merely justified true

belief in the conclusion's assertion *understood as* a description merely of sensible appearances).

Perhaps a clarification is in order. Suppose that, for ease of presentation, we move to a formulation focusing on epistemic access to unobservables (as we so often find in this debate) rather than the corresponding concepts needed to access them, and then couch this in terms of beliefs about phenomena and noumena.

Then, if we assume that omniscience (P_x) would mean possessing *all and only* justified true beliefs concerning not only sensible appearances but also concerning reality as it is in itself[58], we can say that P_{x-1}, in falling just short of that, represents an epistemic position allowing very *nearly* all and only justified true beliefs concerning both sensible appearances and reality as it is in itself. As we follow the continuum back toward insentience, at some point it will become extremely plausible to say that a given P_n occupies an epistemic position enabling justified true beliefs concerning sensible appearances but *not enabling* justified true beliefs concerning reality as it is in itself.

Then, our question is: *What reason have we for thinking that P_5 corresponds to P_{n+1} rather than merely this P_n?* The realist's epistemic optimism commits her to the view that P_5 *does* correspond to P_{n+1}, since this latter assignment represents an epistemic position through which *at least one* justified true belief is possible concerning reality as it is in itself; it is, as it were, the absolute minimum threshold for the kind of epistemic optimism built into realism. But it is not at all clear what positive reasons we could have for asserting this correspondence.

[58] In fact, here (in the limit case) the two domains should fully coincide, or be one and the same.

If this is so, then *S1* above is well-supported, and we will not take our comparative lack of conceptual limitations to suggest that we possess the *least* conceptual limitations that it is metaphysically possible to possess. This in turn suggests that we should draw the same pessimistic conclusion about the ability of our cognitive equipment to discern the underlying structure of reality that we draw in other cases of cognitively constrained organisms considered relative to others on the continuum C: even though we come closer to discerning this structure than *them*, in relation to a less cognitively constrained species (whether merely possible or actual) we are still *getting it wrong*. Alternatively: *because* the kinds of concepts it is possible for us to possess are limited in relation to some (possible or actual) intellect in the way that a bonobo's is related to ours, we have no good reason for thinking that the most fundamental concepts we are capable of possessing are the most fundamental concepts it is metaphysically possible to possess. But then we certainly cannot even begin to form beliefs about those aspects of reality that require (in order for those beliefs to have meaningful content) *those very concepts*.

We can, for example, put this once again in terms of abductive inferences. Very many of the abductive conclusions we actually draw would, from the perspective of the epistemically less-constrained being we are imagining, be recognized as false or inaccurate as they bear no substantive relation to the way things are in-themselves. If we could occupy that more privileged epistemic position, we would then draw abductive conclusions which (in committing us to a more accurate ontology) would be *inconsistent* with the abductive conclusions we actually *do* draw. In other words, S1 supports a second skeptical thesis, S2:

S2*: *Had we the benefit of the cognitive faculties belonging to the (possible but perhaps nonactual) P_n about whom it is legitimate to claim $R(P_5, P_n)$, the probability is high that (*i*) we would not think that the conclusions of our *actual* abductive inferences accurately describe the ultimate nature of reality; and therefore, (*ii*) that given the insight of that perspective, we would *not* be optimists about the cognitive-epistemic capacities of P_5.[59]

Now we are in a position to continue enumerating the premises of ACC – *viz.*, to present its second stage:

5. From (4) it follows that we cannot be justified in thinking that the most fundamental, 'underlying-reality-tracking' concepts we are capable of possessing are the most fundamental, 'underlying-reality-tracking' concepts it is metaphysically possible to possess. (That is to say, claim C from section 4 is false.)
6. But if we cannot be justified in thinking that the most fundamental, underlying-reality-tracking concepts we are capable of possessing are the most fundamental, underlying-reality-tracking concepts it is metaphysically possible to possess, then we have no reason to believe that we can form scientific theories that are capable of reliably and accurately tracking the underlying structure of the world (since such theories need those very concepts).

[59] If it is doubted that things would in fact seem this way from the perspective of *just any* P_n that satisfies the relation $R(P_5, P_n)$, we can simply rephrase S2 to say: "Had we the benefit of the cognitive faculties belonging to P_x…" where P_x represents an absolute limit of the continuum, i.e., the position to which we assign omniscience. Then, on pain of inconsistency (given the realist's commitment to the view that the world can in principle exceed what it is possible for us to access epistemically) it would not be possible for the realist to object to S2.

7. Thus (from 5 and 6), we have no reason to believe that we can form scientific theories that are capable of reliably and accurately tracking the underlying structure of the world (rather than mere phenomena).

Needless to say, if this argument is sound then – in establishing our conclusion (7) – we have successfully undermined (what I suggested in section 2 is) the *one and only* claim part and parcel of scientific realism that distinguishes it from the modest (i.e., concessive) form of antirealism I propose.

Section 6: Remarks on the Conclusion of this Argument against Scientific Realism

Having already devoted considerable space to the defense of the first (and critical) stage of ACC, my hope is that the following truncated remarks will suffice by way of summary.

First, for the unconverted, I shall try to distill the key step of ACC into one simple (and hopefully uncontroversial) claim. As one final intuition pump supporting ACC's key step, then, consider the following thesis, which we can call *Implausibly Lucky Coincidence* (*ILC*):

> ***ILC*:** It would be an implausibly lucky coincidence if the concepts we happen to be able to grasp (as a result of the way in which natural selection has shaped our overall psychology) were *the very same* concepts we would have if our psychology were instead shaped by, say, a machine perfectly designed for the explicit purpose of endowing us with the psychology necessary to possess the concepts needed to represent the underlying structure of the world.

I hope that this way of framing the problem (i.e., along the lines of ILC) illustrates just how sensible the basic intuition is that nonetheless leads us (as I contend) directly to antirealism.

Finally, I shall make a closing remark on naturalism and how it fits (or, as I claim, fails to fit) with scientific realism – and why. We saw at the beginning that the realist insists that the world has a distinctive structure independent of us (a view, again, which I share). Although the realist therefore stresses that what the world is like is not a function of what *we* are like (except in an attenuated and unproblematic sense), ACC has been constructed so as to put considerable pressure on the realist's ability to coherently maintain this view.

For what we seem to have shown is that scientific realism not only expects too much of us, but also (to put it only *partly* figuratively) *expects too much from the world*. That is because there is a sense in which the view requires that the world pull off the incredible feat of coming ready-made with a structure that suits the concepts that a particular species happens to be built to deploy (where such deployment is a necessary condition of members of that species constructing theories that accurately latch on to that underlying structure, as scientific realism claims we can, at least in principle, do). But as the argument above has sought to establish – and as any good naturalist should admit – it just isn't reasonable to suppose that this is so.[60]

[60] It is not difficult to see (at least if one shares one or more of the common-sense intuitions I have exploited throughout) how realism and naturalism are an unnatural fit in an even broader and more basic sense. If naturalism claims that our philosophical theories should take stock of our best science, and if (part of) our best science (i.e., evolutionary biology) gives us reason to doubt that human cognition is structured so as to map reality as it is in itself, then realism's tacit assumption to the contrary is enough to put naturalism and realism at odds straightaway.

Section 7: Applying this Form of Argument to Sider's Metaphysics[61]

In this (rather lengthy) section of the third (and final) chapter of this dissertation, I aim to use the form of argument I called a "naturalistic argument against scientific realism" – together with what I maintain this argument proves – to challenge the central thesis of Ted Sider's (2012) work; if successful, this would not only be significant in itself, given Sider's standing in the community of metaphysicians, but would also provide indirect support for the original argument against scientific realism we have just finished discussing.

As one would expect, Sider's (2012) work is both ambitious and original: it presents a revisionary conception of the aims of metaphysics and supports this with a sophisticated and insightful survey of the ways in which it allows us to gain traction on some perennial metaphysical (and other philosophical) disputes.

But in spite of its considerable ingenuity, this project is deeply flawed. *Structure*, I contend, has no place in metaphysics. I should say at once, however, that I do not intend to offer anything like a knockdown argument against Sider's view. His *Writing the Book of the World*, after all, is a sprawling and systematic treatment of the subject; and, as he has indicated, his argumentative strategy is to defend his position in a *holistic* fashion, to put it somewhat clumsily. That is, he sees the *totality* of his work in this book as the argument for his conclusion that structure is central to metaphysics; there is no master argument to attack here, but rather simply numerous consecutive attempts to show how his view sheds light on otherwise recalcitrant philosophical disputes.

---

[61] Thanks are owed to Joseph Tolliver for providing the venue for the development of this rather lengthy second-half of this chapter, and for helpful comments on some of the ideas developed here when they were still in their infancy. Thanks are also owed to Yael Loewenstein for helpful conversation on this material.

In particular, he attempts to show that positing structure leaves us better off in our attempts to understand such vexed topics as reference, laws of nature, and intrinsic properties, not to mention disputes in metametaphysics, and much else. His argument for the importance of structure, then, is just this: "structure is a *posit*, a posit that is justified by its ability to improve our theories of these matters" (2012: 13).[62]

Given the nature of Sider's project, then, it is not difficult to see that we are dealing with a view that could only be fully dismantled by way of the kind of comprehensive treatment of his book that, so far as the scope of the second half of this chapter is concerned, would simply not be practicable. But even so, I think it is possible in the limited space available here to present and defend a closely related set of powerful (if defeasible) objections to *the basic premise* of his project.

Although this is a tall order, succeeding in it would, in one swift move, cast doubt on *all that he goes on to do* with the conception of metaphysics that is built upon that contested premise – and *without* having to critique Sider's every move throughout the course of his book. And success in this would make it possible, even in the limited space available here, to go some distance toward a satisfactory sketch of what a viable alternative to Sider's view might look like; indeed, in large measure this simply falls out of the critique of Sider that I offer.

---

[62] That's awfully cute, but if we adopted this as our general philosophical methodology, countless philosophical theories could be devised to solve a whole host of seemingly intractable, long-standing problems, regardless of whether we actually have any independent reason to believe such theories. (Nonetheless, I do realize that this form of argument has gained some currency since, in particular, David Lewis.)

7.1 *The Central Concepts: Joint-Carving, Fundamentality, and Structure*

Sider's vision of the aims of metaphysics draws upon three interrelated concepts: structure, fundamentality, and joint-carving. These are terms of art that deserve explicit definition and careful attention. Sider often defines them *in terms of one another*, and thus we often find statements similar to the following *paraphrased* formulation of his view:

> *Metaphysics is concerned with what the world is fundamentally like; what the world is fundamentally like is a matter of the world's structure; and we accurately represent that structure only to the extent that we represent the world in terms that carve at the joints – that is, when we identify (and then put to use) the correct categories for describing the world.*

Although this quick presentation of Sider's view does illustrate the interrelation of the three concepts, it leaves something to be desired as a means of defining them. Unfortunately, it is difficult to give a definition of any one of these terms, in a way that does justice to Sider's intentions, without some reference to one of the *others*; but hopefully the following elaboration of the three notions is sufficiently perspicuous.

Both joint-carvingness and fundamentality come in degrees, and the two notions co-vary. Thus we can say things like: a concept, term, or notion is *perfectly* fundamental (e.g., *mass*) just in case it carves *perfectly* at the joints; a concept, term, or notion is *reasonably* fundamental (e.g., *red*) just in case it carves at the joints *reasonably well*; and a concept, term, or notion is *nonfundamental* (e.g., *phlogiston*) just in case it carves at the joints *poorly* (or perhaps not at all). It is in this sense that there is a perfect covariance between joint-carvingness and

fundamentality. Even so, however, it is helpful to think of fundamentality as being *determined by* joint-carvingness (and ignoring the converse relation).

Then, when we ask: *what determines joint-carvingness?,* we can introduce the concept of *structure* instead of vacuously referring back to fundamentality. Whether and to what degree a term, notion, or concept is joint-carving, then, will be a function of its success in mapping onto the world's underlying structure. Just what is meant by this is difficult to pin down – at least without making reference to either of the other two notions already introduced. But what is clear enough is that when we refer to the world's structure, we are to have in mind features of the world that can be adequately captured only when our representations *not only* express truths, but are *fully* accurate in virtue of *using the right concepts*. Thus, Sider writes:

> The world has a distinguished structure, a privileged description. For a representation to be fully successful, truth is not enough; the representation must also use the right concepts, so that its conceptual structure matches reality's structure (2012: i).

This passage and the exegesis preceding it (it is hoped) will suffice for the moment as an explanation of the core concepts Sider deploys: joint-carving, fundamentality, and – the most elusive of all – structure. Somewhat greater precision will be had, in any case, in the following section. But to foreshadow where Sider is headed, we can return to the first line of the passage quoted above and note that *this* is the claim Sider is most concerned to establish. What it comes to, rephrased only slightly, is that there is an objectively correct way of "writing the book of the world." And somehow, it is supposed to follow that this claim to objectivity can only be guaranteed when we *move beyond mere truth* and further *couch that truth* in the correct representations; and doing *that* requires using the correct concepts.

7.2 *'Bred' and 'Rue' vs. 'Red' and 'Blue'*

Corresponding to this conception of objectivity in metaphysics is a conception of its *aim*: although one might have thought that metaphysics is essentially concerned with ascertaining *what there is*, Sider's view is that its goal is to uncover the fundamental structure of reality – a necessary but *insufficient* condition of which, again, is discerning truths.

In order to begin to make this claim plausible, Sider introduces a case that is meant to pump our intuition that *it matters* whether we're carving the world up correctly with our concepts—that is, that it matters to metaphysics whether we're hitting upon the actual structure of the world rather than *merely* saying true things about it.

He asks us to consider a universe filled entirely with fluid, and to imagine that a plane separates the universe into a half whose fluid is red and a half whose fluid is blue. But we can imagine ignoring the dividing plane and instead carving the world up differently, while using the predicates "bred" and "rue" instead of "red" and "blue." If we were to do this, it would be possible (so long as we employed our newly-coined terms correctly and consistently) to say all and only *true* things about the fluid-filled universe.

But it seems that a community that did this – a community that gave up 'red' and 'blue' in favor of 'bred' and rue' – would be making a mistake: only certain concepts hit upon the actual structure of the world, and here it is 'red' and 'blue' that fit the bill, not 'bred' and 'rue.' They might succeed in discerning truths (*that* activity, after all, can be carried out successfully even with these gerrymandered concepts), but they would be failing miserably to report the world's *structure* (*that* is something that one actually needs the *correct* concepts for).

And if that is what we are inclined to say about the community of speakers who use 'bred' and 'rue' in lieu of 'red' and 'blue', then it seems that we *already* accept Sider's core claim – that structure matters to metaphysics, and that joint-carving concepts are what make the difference between hitting upon that structure and failing to do so. In short, Sider's case concerning color concepts suggests that the success of metaphysics hinges upon our *choice* of concepts and not merely whether we say only true things with the concepts we happen to choose or use.

7.3 *Bred, Red, Rue, Blue: General Structure of a Response*

What should Sider's opponent say in response to this illustration? Doesn't it seem intuitively compelling that 'red' and 'blue' carve at the joints better than 'bred' and 'rue'? And if that's so, doesn't it seem that the community that uses 'red' and 'blue' succeeds in a way that the other does not? And if *that* is so, isn't joint-carvingness (and hence *structure*) central to metaphysics?

The claims implied in these rhetorical questions are, of course, three *distinct* claims; and there are, needless to say, two inferential steps the truth of which is necessary to license the move from the first to the second, and from the second to the third. So what we really have is this:

(1) 'Red' and 'blue' carve at the joints better than 'bred' and 'rue.'

(2) If 'red' and 'blue' carve at the joints better than 'bred' and 'rue', then the community that uses 'red' and 'blue' succeeds in a way that the other does not.

(3) Thus the community that uses 'red' and 'blue' succeeds in a way that the other does not. (1,2)

(4) But such differential success arises solely out of differences in joint-carvingness and attention to structure; so if such a difference in success is found, joint-carvingness and structure are central to metaphysics.

(5) Hence joint-carvingness and structure *are* central to metaphysics. (3,4)

One might have thought that the only (or perhaps the most promising) way to stop this argument in its tracks is to deny (1) outright. That is an approach one *might* take, but it would be a mistake to do so; it would saddle us, I think, with the implausible view that the notion of joint-carvingness is *incoherent*. For surely the notion – if it is intelligible *at all* – is applicable *here*. That is, if one wants to deny that 'red' and 'blue' carve at the joints better than 'bred' and 'rue', it seems that one must also be prepared to deny that joint-carving is a coherent notion. That may be so, but it strikes me as implausible. And if it is correct that (1) is false only if it is *gibberish*, the principle of charity suggests, in any case, that we should understand 'joint-carving' in whatever way is required to make the claim in (1) come out true.

Fortunately, however, *it is not necessary for Sider's opponent to deny (1)*. Straightaway, then, a concession is in order: some terms *do* "carve at the joints" better than others, *given* the way in which this notion has been defined. But that, I hasten to point out, *establishes nothing of interest*. Some terms, for that matter, impinge upon our sensory faculties so as to produce aesthetically favorable judgments more than others. But just as one term's *being prettier* than another is (without further argument, at any rate) *insufficient* to justify the

construction of a conception of the aims of metaphysics *around that mere fact*, so too is it inadequate to the task of justifying a conception of the aims of metaphysics centered on joint-carvingness to merely draw attention to the fact that some terms *carve at the joints better than others*.

This, of course, is nothing Sider would deny; he does, after all, (attempt to) *provide* the "further argument" alluded to parenthetically in the paragraph above. But the point is just that *we can make this concession* without being led down the path Sider intends, at least provided that we have a response to the argument he actually offers for thinking that joint-carvingness and structure are central to metaphysics. We can, in short, grant that 'red' and 'blue' do carve at the joints better than 'bred' and 'rue.'

The claim, then, will be that Sider's initial intuition pump – while successful in eliciting the intuition that some terms carve at the joints better than others – is impotent to do any *further* philosophical work. That is because we can grant that there are differences in joint-carvingness without being forced into the position that these differences *matter*.

In order for them to matter (I shall argue), *one* – although not the only – thing that must be shown is that some terms carve at the joints *perfectly* (and not merely *better* than others). We shall see shortly why this is so, but for the moment what is important is just that we acknowledge the consistency in holding that some terms carve at the joints better than others while also *denying* that any terms carve at the joints *perfectly*; this will be important to forestall one kind of objection that might otherwise have been raised against the argument I develop in response to Sider.

Before moving on, a remark clarifying the style of response to be offered to (1) – (5) above is in order[63]. Given what I have said above, it might be unclear whether it is the inference in (2) or the inference in (4) that I am concerned to deny. Because it seems to me dangerously *misleading*, if not highly suspect, to claim as (3) does that the Bred/Rue community is making a mistake that the Red/Blue community is not, I have serious misgivings about the inference that warrants this conclusion – viz., (2). But because many of us seem to feel the force of (2) and (3) – and because it is not necessary for the success of my argument to dispute either of these claims – I shall accept them for the sake of argument. Thus while my remarks above might be construed as an attack upon (2), it is preferable to treat them as a challenge to (4) instead.

But in order to see precisely how the preliminary (and admittedly truncated) remarks I have already made really do raise a problem for (4), a significant amount of further philosophical work must be completed. Furnishing this is the main aim of this paper. Before beginning that process, however, let us briefly pause to take stock of where we are at by emphasizing the claims I am prepared to grant to Sider.

First, it seems uncontroversial that 'red' and 'blue' *do* "carve at the joints" better than 'bred' and 'rue.' (But note, again, that it is important not to make heavy weather of this fact; it follows, after all, simply from the way in which *joint-carvingness* is defined.) And, second, we can (merely for the sake of argument) grant that this entails that those who use

[63]Steps (1) – (5), I should note, are not being put forward as a reconstruction meant to capture the central argument of Sider's book (insofar as there really even is a *central* argument there to be discerned). But for these early purposes it does suffice as an adequate representation of the intuitive, if imprecise, reasoning that Sider appeals to, and hence deserves *some* small attention.

'bred' and 'rue' are making a mistake that speakers who use the predicates 'red' and 'blue' are not.

But what I am *not* prepared to grant – what I think the balance of reasons demands we *reject* – is the further claim that the kind of success enjoyed by speakers of 'red' and 'blue' (but not by speakers of 'bred' and 'rue') *itself* forces upon us the conclusion that joint-carvingness is of central importance when theorizing about the aims and nature of metaphysical inquiry. In fact, I think it can be shown that this conclusion – far from being forced upon – is almost certainly mistaken.

7.4 *The Basic Problem: Perfect Fundamentality, Perfect Joint-Carving*

The basic problem with Sider's picture of the aim of metaphysics is that it depends crucially upon the possibility of *perfect* joint-carving and *perfect* fundamentality. Before explaining *why* I take it that this feature of Sider's view guarantees its failure, it is important to establish that Sider's view *does in fact* depend upon the coherence of these two notions. This task is actually divided into two parts.

The first is undertaken in this section, where I begin by consulting the text in an effort to demonstrate Sider's own, explicit commitment to these two notions. Then, in the remainder of this section, I set this task aside in order to move on to making a preliminary case *against* the coherence of these two notions.

But I take this task back up in section 7.7 by asking what would happen if a defender of Sider's view were to drop the commitment to these two notions in order to save the rest of Sider's realism about structure. The goal there is to show that Sider's view actually *needs* these two faulty notions, and is therefore damned from the outset; thus one can think of this

as an extension of what I do in the first part of the present section: i.e., arguing that it is a legitimate move to attack the notions of perfect joint-carving and perfect fundamentality *because they do in fact* form an essential component of Sider's view.

We have already seen that Sider thinks that some notions carve at the joints better than others (e.g., 'bred' carves better than 'red'); and this, recall, is a claim I am prepared to grant. But Sider points out that concepts like 'red', while carving better than concepts like 'bred', nonetheless don't carve *perfectly*; they're not *perfectly* fundamental. That's because reality's structure is not adequately represented by the use of such concepts; although 'red' comes *closer* to representing parts of the world's structure than 'bred' does, there's no *redness*, as such, *in the world*.

Are there, however, concepts that *do* perfectly represent the world's structure? Are there, in other words, perfectly fundamental concepts, or concepts that perfectly carve reality at its joints? Sider not only thinks that there *are*, but even makes an effort to identify them. Although he points out that his project does not require defending a particular claim about *what* these perfectly fundamental concepts are, his own view, articulated very early on in his book, is that they are the ones invoked by physics, logic, and mathematics (2012: 7).

Much later in the book, he says that "…fundamental reality contains nothing but physics, logic, and set theory" (2012: 347). Statements such as this, when combined with the explanation of fundamentality in Section 7.1 above, not only yield the view that the concepts of physics, logic, and set theory carve perfectly at the joints and are perfectly

fundamental, but also indicate quite clearly that they are, in Sider's view, the *only* perfectly fundamental and perfectly joint-carving concepts.[64]

And, finally – in order to tie all of this explicitly to Sider's vision of the aims of metaphysics – we have the following passage, the italicized portion of which is particularly relevant for our purposes:

> Structure is particularly central to metaphysics. The heart of metaphysics is the question: what is the world ultimately, or fundamentally, like? And fundamentality is a matter of structure: the fundamental facts are those cast in terms that carve at the joints. *The truly central question of metaphysics is that of what is* ***most*** *fundamental. So in my terms, we must ask which notions carve* ***perfectly*** *at the joints* (2012: 6, emphasis added).

If *this* doesn't establish the importance of *perfect* joint-carvingness and *perfect* fundamentality to Sider's project, then surely nothing will. I propose, then, to move on now to a critique of these notions.

First, a brief overview of the general argumentative strategy I will employ. What I shall be arguing, at bottom, is that *even if there are* perfectly fundamental concepts (i.e., ones that carve perfectly), it is extremely implausible (as we saw in the first half of this chapter, where scientific realism was the focus) that they are concepts that it is possible for us to possess. Suppose, just for the moment, that I have succeeded in proving this in sections 1

[64] It might be objected that I have not taken proper account of the fact that I myself noted above – i.e., that Sider does not purport to offer a defense of the fundamentality of these particular concepts. As such, it might seem unfair to attack his (undefended) view that these *are* perfectly joint-carving concepts. But this objection rests on a misunderstanding. I am not going to be attacking the perfect fundamentality of the concepts of physics, logic, and mathematics *per se*; rather, my argument shall be perfectly general, in the sense that *any* concepts one could conceivably posit as filling this role will be inadequate to the task.

through 6 of this chapter. But then, if the success of metaphysics requires the use of the most fundamental concepts (as Sider claims in the passage cited above), metaphysics is doomed to failure. But since neither Sider nor I am willing to bite *that* bullet, it must be that metaphysics can proceed profitably without attention to joint-carvingness and fundamentality. And the upshot of *that* will be that structure is unessential to metaphysics.

In the following two sections, I develop arguments that codify and formalize these claims. (Inevitably, this will occasionally involve rehearsing some of the claims made in the first part of this chapter, but I have worked hard to keep this to a minimum.) But what I will do first, in the remainder of this section, is to informally pump the intuition that we could not conceivably have access to the most fundamental concepts possible – even if the notion of perfectly fundamental concepts is coherent (i.e., if it is metaphysically possible for some being to possess such concepts).

Asking what concepts God would, or could, possess is of some help here. Interestingly, Sider too is fond of something resembling this approach. Remarkably, however, he draws extremely implausible conclusions from his use of this device. First, he says, "[w]hen God created the world, she was not required to think in terms of nonfundamental notions like city, smile, or candy" (2012: 126). Fair enough. But then, in a much later passage, he goes on to say:

> A vivid test for whether a given expression, *E*, carves at the joints is this: did God need to think in *E*-terms when creating the world? Clearly, she needed to think in terms of quantification, mass, distance, and so on; accordingly, those notions carve [perfectly] at the joints (2012: 163).

Here, the thought is that whatever concepts God would require in creating the world are the perfectly fundamental ones. Something *like* this is a claim I am willing to grant, so for the sake of argument let's suppose that Sider's statement is true. He then proceeds to claim, however, that God would need to think in terms of the concepts of physics; as such, those are (among) the perfectly fundamental ones. It is hard to dispute this claim – that an infinite intellect would require the concepts of physics – in any fashion other than to point out how remarkably *unimaginative* the position seems. It is entirely reasonable to suppose that, *sub specie aeternitatis* (or alternatively, if we prefer, from the perspective of an *infinite intellect*), the concepts of even physics could be dispensed with entirely, in favor of concepts that map even more perfectly (more fundamentally, carve at the joints better) than the concepts that *happen* to be needed by one particular species to make sense of the world. (We are, recall, considering an *infinite* intellect. Do we really think that god – if such a thing existed – would be doing the kind of work in creating the world that the physicist does at her chalkboard, just at a faster pace on account of his omnipotence?) But this is not exactly an *argument*, so let us try another approach.

I contend that this claim, first introduced in our discussion of scientific realism, is false – viz., *that the most fundamental concepts we are capable of possessing are the most fundamental concepts it is metaphysically possible to possess*. But even if this much is granted, it might, at this point, be objected that I have somehow missed my target; that is, an opponent might say:

> *Well, you're not really disputing the coherence of perfect fundamentality or perfect joint-carving; for all that has been said, Sider might be correct that there are perfectly fundamental concepts.*

In response to this line of thought, consider a different way of putting the same worry introduced above, which should be sufficient to dispel any concern our imagined opponent's objection may have raised. It can be phrased as a dilemma. Either Sider means to qualify "the perfectly fundamental concepts" as "the most perfectly fundamental concepts it is possible *for us to have*" or he does not (in which case he means: the most perfectly fundamental concepts, *tout court*).

If he does not, then – for the reasons already introduced above, and to be elaborated upon in the following section – his view has, as a consequence, the view that in order to do metaphysics, we need access to something we could never have.

But suppose he does mean only to refer to the most perfectly fundamental concepts *it is possible for us to have*. Then, the connection between the success of metaphysics and the possession of perfectly fundamental concepts seems tenuous at best; why should we think that carving the world at its joints is the kind of thing that can be done with our meager (although, on this horn of the dilemma, *perfectly-fundamental-for-us*) concepts when they are not the most fundamental concepts *tout court*? In short, this revised understanding of what is meant by "perfect fundamentality" and "perfect joint-carving" seems, strictly speaking, to *dispense* with Sider's core thought that *perfectly* fundamental concepts are necessary for success in metaphysics. And, more importantly, on this reading we must attribute to Sider a view he should clearly want to reject: viz., that we are capable of conducting metaphysics fully adequately with merely the most fundamental concepts it is possible for us to possess, and therefore that a being that *did* possess (the truly) perfectly fundamental concepts would succeed no better than us when doing metaphysics.

7.5 *An Evolutionary Debunking Argument against Structure*

Let us finally present a formal argument against Sider's realism about structure. In fact, I will present *two* arguments against Sider's view, which, although distinct, are closely related (and even share some of the same premises). The first, which I present and defend in this section, can be called *an evolutionary debunking argument against structure*; it draws upon the remarks of the previous section, transforming them into acceptable premises.

For clarity of exposition, I first state the undefended argument in premise and conclusion form, and subsequently move on to defend each premise in a manner continuous with my original remarks from Section 7.4. This first argument will be divided into two distinct parts. The first part of the argument, as promised earlier, is inspired by an evolutionary argument from Sharon Street (2006: 109-114). Hers, of course, is an evolutionary debunking argument *of ethical realism*, but the general style of argument can be mirrored and adapted suitably to our purposes. The second part uses the conclusion of the first part to fill out more explicitly the implications for Sider's view.

An Evolutionary Debunking Argument against Structure

1. Natural selection has profoundly shaped human psychology.
2. This shaping had as its "aim" the selection of whichever total psychology best contributed to the biological fitness of the organism.
3. Because our actual concepts, as well as the broader set of whichever concepts are cognitively accessible to us, are a product of our psychology, and given (1), it follows that both our actual and possible concepts are "thoroughly saturated with evolutionary influence."

4. But then, by (2) and (3), the mechanism by which concepts are generated can be traced to (i.e., have their origin in) whatever best contributes to biological fitness.
5. We have no reason to believe that a process that shapes psychology (and hence concepts) with an eye toward biological fitness would also be a process that reliably and accurately tracks the underlying structure of the world.
6. Thus, from (4) and (5), we have no reason to believe that the concepts it is possible for us to have are concepts capable of reliably and accurately tracking the underlying structure of the world.

Premises (1) – (6) closely mirror the structure of (part of) Street's style of argument against ethical realism, but adapted to present purposes. Once we have reached the intermediate conclusion in (6), however, we can proceed as follows:

7. If we have no reason to believe that the concepts it is possible for us to have are concepts capable of reliably and accurately tracking the underlying structure of the world, then we have no reason to believe that the concepts it is possible for us to have are the *most fundamental* concepts it is possible to have (*tout court*).[65]
8. If we have no reason to believe that the concepts it is possible for us to have are the *most fundamental* concepts it is possible to have (*tout court*), then we have no reason to believe that any of the concepts that it is possible for us to have are ones that perfectly carve reality at its joints.

[65] Assuming as Sider does, of course, that there *are* such things as 'the most fundamental concepts it is possible to have.'

9. If we have no reason to believe that any of the concepts that it is possible for us to have are ones that perfectly carve reality at its joints, then we can have no (all things considered) reason to accept any conception of the aims of metaphysics that claims that the success of metaphysical inquiry requires us to possess and deploy concepts that perfectly carve reality at its joints – unless we are prepared to accept that metaphysics is impossible.
10. Thus (from 6 – 9), we can have no (all things considered) reason to accept any conception of the aims of metaphysics that claims that the success of metaphysical inquiry requires us to possess and deploy concepts that perfectly carve reality at its joints – unless we are prepared to accept that metaphysics is impossible.
11. But Sider's conception of the aims of metaphysics is one such conception.
12. Thus, we can have no (all things considered) reason to accept Sider's conception of the aims of metaphysics – unless we are prepared to accept that metaphysics is impossible.
13. But Sider is not prepared to accept that metaphysics is impossible. Nor, for that matter, should we be.
14. Thus we can have no (all things considered) reason to accept Sider's conception of the aims of metaphysics.

The basic thrust of this style of argument was already explained in Section 7.4 (although admittedly in elliptical form), and so my hope is that only a brief commentary on its premises is necessary.

The first premise may appear contentious, but in fact should not be. In order to see why, consider the way in which this premise gets used in Street's argument against ethical realism. No serious party to the debate over the success of evolutionary debunking arguments against ethical realism disputes that human psychology has been shaped substantially by natural selection. The disagreement, rather, concerns what to make of this fact. I will have more to say about this in a moment when it becomes relevant to addressing subsequent premises of our argument, but for the time being it is sufficient to note that, at least in the literature on the debate over evolutionary debunking arguments against *ethical* realism, no real effort is made to resist this claim. And just as even the ethical realist – who is presented with an argument against her view that relies upon this premise – can accept it, so too, I think, would Sider. His view may, as I argue, be wrong, but surely he is not *unreasonable*.[66]

The second premise follows straightforwardly from (1), provided that we understand what is meant by 'natural selection.'

The third premise is likely to meet more resistance than (1) or (2), but can easily be shown to be unproblematic, given that (1) and (2) are accepted. In order to see why, it will help to return to the nature of the debate over evolutionary debunking arguments against *ethical realism*, which is the true home of this style of argument. On what grounds do ethical realists object to the *analogous* version of this premise (as it is employed in evolutionary debunking arguments)? In short, they argue that although natural selection has shaped human psychology just as much as it has shaped human morphology and behavior, it is not

[66] Of course, an *unreasonable* person might deny that natural selection explains *anything*, and will thereby get off board before the debate even gets going. Less extreme views might hold that natural selection explains morphology but not behavior and psychology. Such a person could reject this premise, but the question is *why* we should think that such a view is at all tenable, given what we know about natural selection.

at all clear that our discrete evaluative *beliefs*, or even our belief forming mechanisms, are influenced in this way. Whatever we might think of the plausibility of this view, it is sufficient to note that *if* it were tenable, it would offer a way out of the version of the third premise that appears in Street's argument.

But what is important for *our* purposes is that no such maneuver is open to the opponent of our 'evolutionary argument against structure.' Street's opponent objects to the claim that our particular, discrete beliefs are shaped by evolutionary forces *on the grounds* that we allegedly possess the ability, by way of reasoned, rational reflection, to form beliefs that are purged of their evolutionary influence – for instance, that we can form the evaluative belief that discrimination is wrong, even though evolutionary forces might dispose us to treat outsiders with hostility. But surely, analogous moves intended to block *our* version of the third premise will fail: the set of concepts it is possible for us to have is surely a function of psychological facts about *us*; if what systematically shapes those facts about us is our evolutionary history, then that same history places hard and fast constraints on what is cognitively possible for us – i.e., what concepts (or kinds of concepts) it is possible for us to have. In other words, in the case of Sider, but not in the case of Street, it is fundamentally about the *cognitive machinery*, and whether evolution has placed hard and fast limits there.

The fourth premise follows deductively from (2) and (3) and thus requires no defense.

In support of the fifth premise, we might reason as follows. It would be an *implausibly lucky coincidence*[67] if turned out that natural selection, in selecting for psychological traits that

[67] This should sound familiar; for a close analogue, see ILC from section 6, above.

maximize fitness, simultaneously happens to select for psychological traits that enable us to grasp the underlying structure of the world. But although this fifth premise strikes me as uncontroversial, I can imagine an opponent raising at least one kind of objection to it.[68]

Perhaps (it might be argued) it actually *is* adaptive to possess concepts that carve perfectly at the joints. If that is so, then it won't (as I have claimed) be an implausibly lucky coincidence at all if it turns out that the very same psychology (which, in turn, furnishes us with concepts) that is selected for its maximization of fitness *also* allows us to grasp the world's underlying structure. Although this line of thought does appear to introduce a real worry, it fails to do so upon closer inspection. Consider, first of all, one kind of concept our interlocutor might have in mind when she claims that the possession of joint-carving concepts might be adaptive: perhaps the possession of genuine color concepts (as opposed to gerrymandered ones like 'bred' and 'rue') enables us to navigate our surroundings better, and hence maximizes fitness. If that is correct, then it might seem that we have found a case where joint-carving concepts are adaptive; and if that is so, then it might not seem so implausible to suppose, contra premise 5, that natural selection simultaneously shapes our psychology to maximize fitness *and* to latch onto the underlying structure of reality.

The problem with this thought, of course, is that – even by Sider's own lights – color concepts are not *perfectly* joint-carving. They may, as Sider claims, be *reasonably* joint-carving, but a case of concepts that are both adaptive and *reasonably* joint-carving won't be enough to cast doubt on our fifth premise. That is because the premise (in claiming that it is implausible that natural selection would shape our psychology so as to make connection with the

[68] Thanks to Yael Loewenstein for pointing this out.

underlying structure of reality possible) is *only* committed to the denial of the plausibility of our possessing concepts that are both adaptive and *perfectly* joint-carving. Color concepts (and countless others) *may* be adaptive, but not even Sider thinks that they are *perfectly* joint-carving; hence, they cannot be concepts that map onto the underlying structure of reality. Thus, they are not candidates for counterexamples to the fifth premise's claim *concerning that structure*.[69]

What if we instead consider concepts that *are* (at least on Sider's view) perfectly joint-carving and hence map the world's structure in the required way? Are any of *these* plausibly adaptive? My own view, of course, is that any attempt to specify candidate concepts will fail (as the argument above implies). But we can at least consider the kinds of concepts that *Sider* places in this category – e.g., the concepts of physics. But when we do so, the original objection falls apart. For it is not *at all* plausible to suppose that the possession of the concepts of *physics* would be selected for by natural selection (in its overall shaping of human psychology) for their adaptive (fitness maximizing) value. One would be extremely hard-pressed, I think, to argue that *mass* and *charge*, as opposed to the less technical concepts of folk psychology, are concepts the possession of which contributes to the maximization of fitness in a way that natural selection would be sensitive to. So much, then, for the objection at hand.

The sixth premise, like the fourth, simply follows from earlier premises and thus warrants no special attention.

---

[69] To put it somewhat differently: these aren't concepts that actually map the underlying structure of reality. They are concepts that, on Sider's own view, merely do a better job of this than concepts like 'bred' and 'rue.' But Sider's position requires that, in order for metaphysics to be successful, there be concepts that *perfectly* map the underlying structure of reality (i.e., carve perfectly at the joints).

Premises (7) and (8), while not *merely* following deductively from prior premises, should be wholly uncontroversial – or, at any rate, they express claims that *Sider* could have no grounds for rejecting. That is because they simply trace out the implications of his own view; in other words, they draw out the logical conclusion of (6) in Sider's own terms – i.e., by offering translations between the notions of structure, fundamentality, and joint-carving.

The ninth premise merely expresses the indispensability of perfect joint-carving to Sider's overall view, which was argued for in Section 7.4.

Premise 10 follows deductively from (6) through (9); Premise 11 is obviously true; and Premise 12 follows from (10) and (11).

The conclusion (14) follows from (12) and (13), leaving only Premise (13) to address. Now the part of the premise that reports Sider's own commitment to the substantivity of metaphysics is beyond dispute; but we might still ask: Should *we* not accept the impossibility of metaphysics? Certainly *I* am not disposed to think that metaphysics is impossible, or that its questions are entirely nonsubstantive (or that their answers are beyond our ken) – but a significant number of philosophers *have* expressed such deflationary sympathies. But my task, needless to say, is not to offer knockdown arguments against all articulations of such views. Focusing on the plausibility of this position *as it relates* to our argument against Sider, however, one might wonder whether the appropriate conclusion to draw from (what I take to be) the failure of Sider's project is that metaphysics is indeed doomed. It might be thought: *If there is no place for 'structure' in metaphysics, then there can be no (substantive) metaphysics.*

This, I think, is best interpreted as a remarkably *dogmatic* claim. It is *possible* that a true believer in structure, when faced with a cogent argument against the indispensability of

structure, would retreat to an error theory of metaphysics and regard the enterprise as doomed to failure. But, in the absence of overwhelmingly convincing reasons for tying the success of metaphysics to structure, the *sensible* conclusion to draw from a proof that structure has no place in metaphysics is *not* that metaphysics is doomed, but rather that we need a new account of the aims of metaphysics. In any case, I take it as a datum in what follows that this is the attitude that almost all of us would have toward finding ourselves in this state of affairs.

7.6 *An Argument from Comparative Cognition*

Although I do, for my own part, regard as sound the evolutionary debunking argument against structure that I have just developed and defended, it is not in my estimation the most compelling argument it is possible to advance against Sider's view. That is because it is possible to craft an argument against Sider's realism about structure that is able to sidestep whatever controversy (however limited) the preceding argument inherits from the *moral* argument after which it is modeled.

In order to frame this second argument appropriately, I want to return to a remark I made in passing in Section 7.4. There, I said that although Sider is on the right track in thinking that the concepts God would require in creating the world are the perfectly fundamental ones, what this actually suggests is a conclusion at odds with Sider's own position. I was quick to point out that my reason for thinking this – i.e., that it is simply *unreasonable* (or, alternatively, *lacking in imagination*) to suppose that the concepts of physics (for example) are the most fundamental concepts accessible to an infinite intellect – admittedly falls short of furnishing the basis for an *argument* against Sider. But one might think of the argument just presented as an attempt to transform this bald assertion into a

respectable key premise to be used in the service of casting doubt on the thought that the most fundamental concepts we are capable of possessing are the most fundamental concepts it is metaphysically possible to possess.

In essence, this will be a truncated variation on the Argument from Comparative Cognition developed above against scientific realism. Recall how the final premise of the *first* stage of that argument – premise 4 – was formulated:

4. But we *cannot* be justified in this belief, since it is always a live possibility that the set we selected as containing the most fundamental concepts is merely *the best of a bad lot*.

If we adapt to present purposes the argument in which this premise appeared (in order to transform it into an argument against Sider), the second stage of the argument will then look like this (which is really little more, at bottom, than a longer version of the argument we have already seen – but modified suitably to take account of Sider's jargon):

5. From (4) it follows that (*i*) we cannot be justified in thinking that the most fundamental concepts we are capable of possessing are the most fundamental concepts it is metaphysically possible to possess; and that (*ii*) there *is* good reason to think that the most fundamental concepts we are capable of possessing are neither perfectly fundamental nor perfectly joint-carving.

6. If we have reason to believe that the most fundamental concepts we possess are ones that fail to *perfectly* carve reality at its joints, then we should reject any conception of the aims of metaphysics that claims that the success of metaphysical inquiry requires us to

possess and deploy concepts that perfectly carve reality at its joints – unless we are prepared to accept that metaphysics is impossible for us.

7. Thus (from 5 and 6), we should reject any conception of the aims of metaphysics that claims that the success of metaphysical inquiry requires us to possess and deploy concepts that perfectly carve reality at its joints – unless we are prepared to accept that metaphysics is impossible for us.

8. But Sider's conception of the aims of metaphysics is one such conception.

9. Thus, we should reject Sider's conception of the aims of metaphysics – unless we are prepared to accept that metaphysics is impossible for us.

10. But (as Sider agrees) we should not accept that metaphysics is impossible for us.

11. Thus we should reject Sider's conception of the aims of metaphysics.

Although rephrased slightly, premises five through eleven (which form the second half of our argument from comparative cognition) have essentially already been addressed, since they appeared in our first argument (the evolutionary debunking argument against structure). Because it would thus be redundant to discuss those premises again here, I refer the reader back to section 7.5 for my defense of them. At present, it will be a more profitable use of space to briefly consider, and then respond to, one potential way of attempting to modify Sider's view with the aim of making it invulnerable to the two arguments I have given against it.

7.7 *Will a Modification Help?*

Both the evolutionary debunking argument and the argument from comparative cognition have made heavy weather of the fact that Sider's conception of the aims of metaphysics relies upon the notions of *perfect* fundamentality and *perfect* joint-carving. Both arguments have attempted to show that we have reason to doubt that these notions apply to any of *our* most fundamental concepts. Thus, a natural thought for a proponent of Sider's view would be to attempt to salvage his position *without* reference to perfect fundamentality and perfect joint-carving. Perhaps (the thought goes) all that Sider's view requires is that our most fundamental concepts are *reasonably* joint-carving and *reasonably* fundamental. With these concepts in hand, in lieu of the *perfectly* fundamental and *perfectly* joint-carving ones I have argued we cannot possess, perhaps we can proceed to engage profitably in metaphysics nonetheless.

Now, as I established in section 7.4, Sider's *stated* view is radically at odds with such a move. Again, he claims that success in metaphysics requires the use of *perfectly* fundamental and *perfectly* joint-carving concepts. Thus, even if we can make sense of this proposed revision, it is not at all clear that it is a revision that Sider could (or would even want to) take on board. But *perhaps* it is a revision that *someone* sympathetic to Sider's view, but persuaded by the arguments above, could consistently accept. Hence it behooves us to determine whether such a modification has any chance of working.

To that end, I propose that we examine Sider's *own* reasons for thinking that metaphysics cannot get by with anything less than *perfect* fundamentality and *perfect* joint-carving. Let's focus on just three of those reasons:

1. According to Sider, realism about structure requires a "ground floor" composed of *perfectly* fundamental notions. The alternative to this, Sider thinks, is "infinite ideological descent", which he writes off as a "seriously weird" hypothesis (2012: 157-158).
2. According to Sider, the most promising way of resisting metaphysical deflationism is "to claim that the crucial expressions in the debate carve perfectly at the joints" (2012: 84).
3. According to Sider, substantivity[70] in metaphysics requires that "ontological questions can be posed in perfectly joint-carving terms" (2012: 200).

All three of these reasons for not downgrading from perfect fundamentality and perfect joint-carving seem cogent. That is, *if* we are *antecedently* convinced by (what we take to be) a sound argument establishing the centrality of structure in metaphysics, and maintain that anything less than the possibility of perfect representation of that structure amounts to failure, then when this antecedent commitment is in place, it naturally follows that any such "downgrading" present in our metametaphysics is not, and cannot be, successful.

Elaborating on this thought, we can specify a fourth – and more basic – reason for thinking that departure from perfect fundamentality and perfect joint-carving spells failure for a view like Sider's. The core of Sider's position (or any position that ties success in

[70] *N.B.*: Here, of course, our scope is confined to *only* those questions that *Sider himself* thinks can be cast in perfectly joint-carving terms. He makes exceptions for special domains (e.g., perception, meaning) where he thinks no *perfectly* joint-carving terms are to be found, on account of reality's lacking corresponding, perfectly fundamental structure. But speaking consistently about the desiderata for substantivity, given this range of exceptions he acknowledges is *his* burden, and not mine. Even so, in this section I attempt to do what is (strictly speaking) not needed, and bring my remarks in the text above into agreement with the possibility of Sider salvaging his position by way of a wholesale disposal of talk of *perfect* joint-carving.

metaphysics to structure) is that metaphysics is essentially and indispensably concerned, obviously, with discerning the underlying structure of reality. This, in turn, means *accurate representation* of that structure. But for a view that requires not only truth but also accurate representation, success in inquiry demands that *the concepts used in our representations be adequate to the task*.

But if we are prepared to acknowledge that our concepts are *not* adequate to the task of accurate representation – *which is precisely what we do* when we downgrade from "perfectly joint-carving" to "reasonably joint-carving" – we are left with a conception of metaphysics according to which any answer we give to a given metaphysical question is, strictly speaking, inaccurate. This, then, is the price we pay for insisting upon the centrality of structure in the face of the arguments developed above. And it seems plain to me that this is *far too high* a cost. In short, we can develop a better metametaphysics than Sider's – that is, one without such high costs.

7.8 *Unstructured Metaphysics*

Suppose that this attempt to respond to the proposed revision of Sider's view is, as I for one do *not* think, *unsuccessful*. We can still say confidently that *Sider* cannot consistently avail himself of this maneuver, given his stated view. That is because he agrees that *if* structure is crucial to metaphysics, then, if *perfect* joint-carving is impossible for a creature *S*, so too is successful metaphysics impossible for *S*. But then, given that he affirms the first antecedent, he should hold that if perfect joint-carving is impossible, an anti-metaphysical position is all that is left for us.

Now since I take myself to have shown that perfect joint-carving is not possible for creatures like us, and yet do *not* think metaphysics is impossible for us, then (per the non-embedded conditional above) it must be that structure is not crucial to metaphysics. But that is just the claim I have sought to prove. But notice that Sider too should accept this conclusion on pain of being saddled with the kind of anti-metaphysical, heavily deflationary view that he rightly heaps scorn upon.

But then we must ask: *What conception of the aims of metaphysics does this leave us with?* The answer, of course, is simply (and unimaginatively, unfortunately): *A conception of metaphysics that dispenses with structure.* This just means locating the success of metaphysics in conformity to the *truth*. Whereas Sider claims that theoretical success is had only by theories cast "in perfectly fundamental terms—theories of fundamental physics, for example" (2012: 28), we shall insist that we can assess theories for their predictive accuracy, correct (i.e., *true*) description of the world and the phenomena they aim to explain without having to look at how well the concepts they employ carve nature at its joints.

Unlike the concepts of joint-carvingness and fundamentality, truth is a *binary* notion, and thus is not vulnerable to the style of argument advanced above. That is, the reasons for thinking that we could not possess *perfectly* fundamental or *perfectly* joint-carving concepts do *not* similarly cast doubt upon the possibility of forming true metaphysical beliefs.

Perhaps, however, this seems too quick. Although my position only makes the modest claims (*i*) that success in metaphysics only requires reaching truth, (*ii*) that we can get to such truth, and (*iii*) that we can therefore do metaphysics, might my own critique of perfectly fundamental concepts, if successful, nonetheless infect my view on *truth* as well?

In short: *No*. The crux of my complaint has been that there are concepts we do not have (indeed, *could* not have) that carve at the joints better than those we do have (or are capable of having); these are the concepts Sider *needs* us to have in our possession if, by his lights, we're to succeed in metaphysics.

Now *what*, analogously, can be said about *truth*? It is of course correct that there are (*perhaps* for similar reasons) truths that we do not know (indeed, *could* not know). And it is of course true that knowledge of such truths would be required for *complete success* in metaphysics – where this means *being able to answer every metaphysical question*. But it is not at all clear why this is problematic. Many of us, after all – i.e., those of us (e.g., *both* realists *and* antirealists in the debate over scientific realism) who think that it is possible for reality to outstrip our grasps – are already committed to this position anyway. It is not a defect of my conception of metaphysics that it implies that we do not, and cannot, have a perfect, "God's eye" ontology. *That*, after all, is not something that any serious metaphysician would take to be a necessary condition for the success of metaphysics. It is, rather, *Sider* whose view is overly demanding. His view has not *merely* the consequence that there are *some* metaphysical questions we cannot successfully answer, but if my argument has been successful, rather the consequence that *no* metaphysical question can be successfully answered.

For these reasons, it seems that a conception of the aims of metaphysics that relies upon discerning truths is preferable to a conception that requires this *as well as* the accurate representation of structure.

Even so, because it functions as a direct challenge to that claim, it will be useful in closing to consider just one additional reason Sider furnishes (which we have not yet had

occasion to discuss) for thinking that metaphysics is not concerned *merely* with truth. This focuses on the *epistemic value* of explanations cast in joint-carving terms as compared with the value of explanations whose only aim is truth. Sider thinks that it is epistemically *better* to think and speak in joint-carving terms; this is because "the goal of inquiry is not merely to believe truly (or to know)", but is also to "to think of the world *in its terms*", which, of course, is just a euphemism for joint-carvingness (2012: 72).

But even if we set aside the problems already introduced for this proposal, we can ask more generally why we should think that *this* is the goal of inquiry. Here, Sider claims, is one[71] reason: "…the aim of joint-carving can be seen as having the same source as the aim of truth: beliefs aim to *conform to the world*…the realist about structure thinks of the world as coming 'ready-made' with distinguished carvings" (2012: 73).

Now, I do of course accept that *beliefs* aim to conform to the world. And more generally, I think we have overwhelming reason to accept a correspondence theory of truth. But given the arguments above, we are now in a position to see why *that* fact is impotent to

---

[71] Sider actually offers three in total, but one of the others (in which joint-carving is tied to objectivity in morality and aesthetics) seems so thoroughly confused as to deserve no explicit attention; the other, while more coherent than the latter, is also less pressing than the one discussed in the text above, and so can be dealt with adequately in this footnote. Sider presents five scenarios of increasing epistemic distance from the world, culminating in a scenario in which my disembodied brain's mental states are produced by a mechanism of chance. He then makes clear what his view has to say about this progression from one scenario to another: "the match between our beliefs and reality's joint-carving structure is gradually eroded" (2012: 75). That is, even in Scenario 4 (which resembles The Matrix), where some of my beliefs are true, they fail miserably to carve at the joints. According to Sider, what this shows is that, to the extent that we are disturbed by the possibility of residing in such scenarios, it must be that "what we care about is *truth in joint-carving terms, not just truth*" (2012: 75, emphasis added).

This is, of course, a clever strategy for showing that truth *together* with joint-carvingness (and not mere truth) is (or at any rate *should be*) the goal of inquiry. But it fails. That is because *everything* that we "care about" (i.e., the beliefs that those scenarios raise problems for) can simply be *reformulated* as beliefs *about* structure. Then, what we care about is just whether *these* beliefs-about-structure are true, and not whether our true beliefs also have the property of carving at the joints. As long as this redescription is available to his opponent, Sider's five scenarios surely cannot force upon him or her the conclusion that "joint-carving truth" is epistemically preferable to mere truth.

generate an analogous conclusion about *structure*. Essentially, a position shaped around this latter conclusion *expects too much from the world*, as I had occasion to argue also in my argument against scientific realism. Although Sider's emphasis is on matching our concepts to the structure of the world, there is a sense in which this requires that *the world* pull off an incredible feat, and come 'ready-made' (as he puts it, borrowing a metaphysical locution from Putnam, I believe) to suit the concepts that a particular species happens to be built to deploy. Thus it is only fitting to end this piece on Sider in the same way I ended my argument against scientific realism: as the arguments above have sought to establish, it just isn't reasonable to suppose that the crucial claim above is so.

REFERENCES

Berger, T.W., et al. (2005). "Brain-Implantable Biomimetic Electronics as a Neural Prosthesis for Hippocampal Memory Function" in *Toward Replacement Parts for the Brain: Implantable Biomimetic Electronics as Neural Prostheses*, eds. Berger, T.W. and D.L. Glanzman, Cambridge: MIT Press, 241-275.

Berger, T.W., et al. (2011). "A cortical neural prosthesis for restoring and enhancing memory." *Journal of Neural Engineering, 8(4).*

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge: MIT Press.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Berlin: Kluwer Academic Publishers.

Bickle, J. (2006). "Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience." *Synthese*, 151: 411-434.

Bickle, J. (2008). "Real Reduction in Real Neuroscience: Metascience, Not Philosophy of Science (and Certainly Not Metaphysics!)" in *Being Reduced: New Essays on Reduction, Explanation, and Causation*, eds. Hohwy, J. and J. Kallestrup, Oxford: Oxford University Press, 34-51.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: OUP

Danaher, J. (2015). Why AI Doomsayers are Like Sceptical Theists and Why It Matters. *Minds and Machines, 25*, 231–246.

de Leon, M.J., et al. (1993). "The radiologic prediction of Alzheimer disease: the atrophic hippocampal formation." *American Journal of Radiology*, 14(4): 897-906.

de Leon, M.J., et al. (1997). "Contribution of structural neuroimaging to the early diagnosis of Alzheimer's disease." *International Psychogeriatrics*, 9 Suppl 1: 183-190; discussion 247-252.

Ellison, Harlan. (1967). I Have No Mouth, And I Must Scream. IF: Worlds of Science Fiction, March, 1967. Galaxy Publishing Corporation.

Godfrey-Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science.* Chicago: University of Chicago Press.

Hampson, R.E., et al. (2012). "Facilitation and restoration of cognitive function in primate prefrontal cortex by a neuroprosthesis that utilizes minicolumn-specific neural firing." *Journal of Neural Engineering*, 9(5).

Hampson, R.E., et al. (2013). "Facilitation of Memory Encoding in Primate Hippocampus by a Neuroprosthesis that Promotes Task Specific Neural Firing." *Journal of Neural Engineering,* 10(6), 1-26.

Hartmann, K. (2016). "Embedding a Panoramic Representation of Infrared Light in the Adult Rat Somatosensory Cortex through a Sensory Neuroprosthesis." *The Journal of Neuroscience*, 36(8): 2406-2424.

Henneman, W.J.P, et al. (2009). "Hippocampal atrophy rates in Alzheimer disease: Added Value Over Whole Brain Volume Measures." *Neurology,* 72(11), 999-1007.

Itzhaki, Ruth F., et al. (2016). "Microbes and Alzheimer's Disease." *Journal of Alzheimer's Disease,* (51): 979-984.

Klein, E., et al. (2015). "Engineering the Brain: Ethical Issues and the Introduction of Neural Devices." *Hastings Center Report,* 45(6), 26-35.

Luo, J., et al. (2016). "Real-Time Simulation of Passage-of-Time Encoding in Cerebellum Using a Scalable FPGA-Based System." *IEEE Transactions on Biomedical Circuits and Systems*, 10(3): 742-753.

Moore, Gordon E. (1965). "Cramming more components onto integrated circuits." *Electronics,* (38:8).

Nietzsche, F.W. (1982). "On Truth and Lie in an Extra-Moral Sense" in *The Portable Nietzsche*, ed. Kaufmann, W., New York: Viking Penguin Inc., 42–47.

NIH National Institute on Aging. (2016). "Alzheimer's Disease Fact Sheet." NIH Publication No. 16-AG-6423 (https://www.nia.nih.gov/alzheimers/publication/alzheimers-disease-fact-sheet#changes)

Parfit, D. (1984). *Reasons and Persons*. Oxford: Clarendon Press.

Presidential Commission for the Study of Bioethical Issues. (2015). "Gray Matters: Topics at the Intersection of Neuroscience, Ethics, and Society, *Volume 2*." http://www.bioethics.gov

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. New York: Routledge.

Sider, T. (2012). *Writing the Book of the World.* Oxford: Oxford University Press.

Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies,* 127, 109-166.

van Fraassen, B.C. (1989). *Laws and Symmetry*. Oxford: Clarendon Press.

Yampolskiy, R. (2012). Leakproofing the singularity. *Journal of Consciousness Studies*, *19*, 194–214.