Routledge
Taylor & Francis Group

Inquiry
An Interdisciplinary Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/sinq20

Digital suffering: why it's a problem and how to prevent it

Bradford Saad & Adam Bradley

To cite this article: Bradford Saad & Adam Bradley (2022): Digital suffering: why it's a problem and how to prevent it, Inquiry, DOI: 10.1080/0020174X.2022.2144442

To link to this article: https://doi.org/10.1080/0020174X.2022.2144442

Published online: 23 Nov 2022.

INQUIRY
https://doi.org/10.1080/0020174X.2022.2144442

Routledge
Taylor & Francis Group

Digital suffering: why it's a problem and how to prevent it[1]

Bradford Saad[a,b] and Adam Bradley[c,d]

[a]Department of Philosophy and Religious Studies, Utrecht University, Utrecht, Netherlands; [b]Sentience Institute, New York, USA; [c]Department of Philosophy, Lingnan University, Tuen Mun, Hong Kong; [d]The Hong Kong Catastrophic Risk Centre, Tuen Mun, Hong Kong

ABSTRACT

As ever more advanced digital systems are created, it becomes increasingly likely that some of these systems will be *digital minds*, i.e. digital subjects of experience. With digital minds comes the risk of digital suffering. The *problem of digital suffering* is that of mitigating this risk. We argue that the problem of digital suffering is a high stakes moral problem and that formidable epistemic obstacles stand in the way of solving it. We then propose a strategy for solving it: *Access Monitor Prevent* (AMP). AMP uses a 'dancing qualia' argument to link the functional states of certain digital systems to their experiences—this yields epistemic access to digital minds. With that access, we can prevent digital suffering by only creating advanced digital systems that we have such access to, monitoring their functional profiles, and preventing them from entering states with functional markers of suffering. After introducing and motivating AMP, we confront limitations it faces and identify some options for overcoming them. We argue that AMP fits especially well with—and so provides a moral reason to prioritize—one approach to creating such systems: whole brain emulation. We also contend that taking other paths to digital minds would be morally risky.

ARTICLE HISTORY Received 4 June 2022; Accepted 2 November 2022

KEYWORDS Artificial intelligence; machine ethics; consciousness; dancing qualia argument; functionalism; alignment problem

1. Introduction

We are living through a period of rapid progress in the development of artificially intelligent systems. In less than the span of an average human life, the state of the art has advanced from a desktop electronic calculator to self-driving cars, facial recognition devices, and the likes of

CONTACT Bradford Saad t.b.saad@uu.nl; brad@sentienceinstitute.org
[1]This is a thoroughly collaborative work. Author order is arbitrary.

AlphaFold 2, MuZero, and GPT-3, which have respectively made scientific breakthroughs on long-standing protein prediction problems, taught itself how to play strategy games at superhuman levels, and displayed hints of general intelligence in summarizing texts, responding to natural language instructions, and writing both computer code and passable poetry. It is doubtful that any digital system that exists today is conscious, i.e. is such that there is something it is like to be it. But as digital systems' capabilities continue to increase, this will become less doubtful. We will call digital systems with conscious experiences *digital minds*. With digital minds comes the potential for digital suffering, i.e. the potential for experiences with negative valence that are *pro tanto* morally bad to cause.[2] Hence, we face the *problem of digital suffering*: the problem of mitigating the risk of suffering in digital minds.[3]

The problem of digital suffering is acute for three reasons. First, current technological trends render the arrival of digital minds an increasingly live possibility. AI researchers tend to disagree about when digital systems that exceed all human cognitive capabilities will arrive, not whether they will. One recent survey of machine learning researchers found that in aggregate respondents accorded a 50% chance of machines that exceed humans in all abilities arriving within 45 years.[4] Admittedly, artificial intelligence researchers have a track record of erring on the side of optimism for predictions about the pace of technological progress, and they sometimes think of human abilities in a manner divorced from consciousness. Still, it would be perilous to ignore the possibility of digital minds in the face of such forecasts from researchers in the field. Nor would it be appropriate to ignore the possibility of digital minds until their arrival is imminent. For digital minds may arrive in the midst of an intelligence explosion in which digital systems recursively self-improve, leading to accelerating gains in digital capabilities that quickly improve from roughly human to well-beyond human.[5] Under such circumstances, we may have little advanced warning about the arrival of digital minds. And with the arrival of digital systems with capabilities that vastly exceed our own, we may no longer be in a position to prevent digital suffering. Thus, we should regard the arrival of digital minds as a live

[2]We adopt this understanding of suffering for convenience and precision. We take it to be a reasonable precisification of an ordinary notion of suffering that marks out a morally important category, but we do not assume that it is the only such precisification. Those who favor an alternative precisification are invited to substitute their preferred notion into the discussion that follows.
[3]See Gloor (2016), Sandberg (2014b), and Tomasik (2017); cf. Schwitzgebel and Garza (2015).
[4]See Grace et al. (2018).
[5]See Bostrom (2014), Chalmers (2010b), and Good (1965).

possibility for the not-too-distant future, and we have reason to take it seriously well before it occurs.

Second, the moral stakes here are astronomical: digital suffering could quickly come to swamp the amount of suffering that has occurred in biological systems throughout the history of the planet. There are several paths to this outcome. On one, digital minds become cheap to produce. As a result, digital minds come to greatly outnumber biological minds. In this case, even if individual digital minds' capacity to suffer is similar to that of biological minds, the sheer number of digital minds could generate quantities of digital suffering that exceed the amount of biological suffering. Another path exploits differences in processing speed. Digital systems can be expected to surpass brains in processing power by many orders of magnitude.[6] This suggests that a given digital mind might undergo more experiences in less time than any biological subject. In that case, even a relatively small number of digital minds might endure many more suffering experiences than all those had by biological subjects. Yet another path turns on the types of suffering that a digital mind might undergo. Given the vastness of the space of minds, there is little reason to think that the sorts of suffering we are familiar with are among the worst possible kinds. So it would be unsurprising if some digital minds are capable of kinds of suffering that are worse than any endured in the biological realm. This raises the possibility of a short-lived 'disutility monster', a digital system that undergoes a form of suffering so severe that its moral importance outstrips all other suffering that occurs in our world.[7] While the problem of digital suffering is especially acute on some utilitarian views, one does not need to be a utilitarian (or consequentialist) to recognize that suffering is *pro tanto* bad and that we have reason to prevent it. This principle can and should be embraced by a wide range of moral views including deontological, contractualist, and virtue theories; if a theory rejects it, that is a very serious mark against such a theory. The problem of digital suffering is thus a high stakes moral problem regardless of one's choice

[6]Neurons operate seven and a half orders of magnitude more slowly than today's fastest (non-quantum) microprocessors and transmit signals via axons eighteen orders of magnitude more slowly than optical signaling in such processors (Bostrom 2014, 71–72; Berry et al. 2020). While it is debatable whether we should expect non-quantum computational capabilities to continue growing in accordance with Moore's law, there is reason to think quantum computing will be a source of significant further gains in capabilities. For instance, in 2019 a Google quantum computer executed in 200 s a computational task that would have taken 10,000 years to complete on the (then) fastest non-quantum computer (Arute et al. 2019).

[7]Compare the discussion of paths to superhuman welfare in Shulman and Bostrom (2020).

of moral theory, though the precise dimensions of the problem may be sensitive to this choice.[8]

Third, current methods for understanding the inner workings of digital systems will tell us almost nothing about the mental lives of future digital minds. Thus, unless new methods are developed, digital minds will likely be *epistemically inaccessible* to us. As things stand, we have little insight into the inner workings of candidate digital minds. If these systems, or near-future versions of them, are conscious, we may have no idea. And even if we have reason to believe that they are conscious, the character of their experience will elude us. This epistemic obstacle exacerbates the problem of digital suffering because preventing suffering requires some way of determining whether a system is suffering, and that is not something we can know about a system whose mental life is opaque to us. Absent epistemological innovation, this aspect of the problem will become more severe as digital systems become more advanced. As digital systems acquire a wider range of cognitive abilities, our credence that they are conscious should presumably increase. However, given current trends in machine learning and big data, there is little reason to think that these advances will be accompanied by human understanding of how these systems work, much less what their mental lives are like.

The difficulty we face in accessing digital minds bears some similarities to the familiar 'problem of other minds',[9] which calls into question our ordinary beliefs about other minds by noting the apparent compatibility of our evidence with the hypothesis that those beliefs are false, e.g. because other humans are unconscious 'zombies'. But while the problem of other minds is a skeptical puzzle with little real world import, our epistemic access to digital systems is—or at least may soon be—a live issue with important moral ramifications.

In this way, the problem of our epistemic access to digital systems more closely resembles the issue we face in understanding the conscious lives of other animals. We lack a satisfactory answer to the question of how far down the phylogenetic tree states of suffering—and so moral status—descend. The moral status of practices such as factory farming or fishing depends, at least in large part, on which animals suffer. As a result, our lack of epistemic access to the experiences of different types

[8]For example, a utilitarian might treat suffering in digital minds as a basic source of moral badness; a Kantian might instead say we have reason to prevent digital suffering because it interferes with an agent's autonomy, deprives an agent of deserved happiness, or because doing so upholds an imperfect duty of beneficence; and virtue ethicist might say we have reason to prevent digital suffering because that is what an exemplar of kindness and compassion would do.

[9]See Avramides (2019) for an overview.

of animals introduces serious moral concerns about these practices.[10] Digital systems could differ from us mentally as much as or more than the most exotic animals.[11] Since the suffering of digital systems could in principle swamp that of living creatures, we have a strong moral reason to develop strategies for improving our epistemic access to digital systems in order to mitigate the risk of digital suffering. The task of this paper is to develop one such strategy.

One might think that the issue of epistemic access can be easily resolved. If we knew the true theory of consciousness, for instance, we could apply it to arbitrary digital minds in order to determine what types of experience they would have. The problem is that the one true theory of consciousness eludes us, and is likely to elude us for the foreseeable future. Extant theories of consciousness range from panpsychism, on which even electrons are conscious, to illusionism, on which our own (apparent) consciousness is an illusion, with a wide array of theories falling in between. Nor do atheoretical, observational methods solve this problem. Purely behavioral criteria are non-starters: we are only in a position to introspectively corroborate the reliability of behavioral markers for human experiences, and we know from our own case that suffering can exist even when it is not behaviorally manifest.[12] Furthermore, we can no more tell what digital systems are experiencing just by looking at their computer circuitry than we can tell what humans or animals are experiencing just by gazing at their neural circuitry. We might turn to neuroscience, with its technologically enhanced methods for observing the brain. However, while neuroimaging has yielded significant insights into how the brain works, it has not delivered a

[10]See, e.g., Singer (1975). While we maintain that the moral stakes associated with the problem of digital suffering are extremely high, we do not assume that they exceed those of the problem of animal suffering. Which problem has higher stakes is a difficult question, partly because the scale of animal suffering depends on its unknown phylogenetic reach—for example, an estimated 99.9998% of animals are invertebrates and it is unclear whether any or most of these suffer (Bar-On, Phillips, and Milo 2018). Similarly, the expected scale of digital suffering depends on the unclear prospects for proliferation of digital minds. While the proliferation of digital minds could conceivably be thwarted at the design phase, the potential for such proliferation is vast, given the potential energy efficiency, habitable range, and mass reproducibility of digital systems, along with the immense volume of the universe reachable from Earth (Bostrom 2014, 59–60, 99-103; 113–114 Ord 2020: Ch. 8). Fortunately, work on both problems can proceed without settling their comparative significance.

[11]Admittedly, some animals may differ greatly from us in some mental respects. For example, there are empirical grounds for doubting that anything like the unity we take to be characteristic of human consciousness is present in octopus consciousness (Carls-Diamante 2017). However, digital minds could also differ from human minds in these ways. A digital emulation of an octopus nervous system may have consciousness that is no more unified than that of an octopus. And digital minds, say in virtue of not having bodies, may differ from human consciousness in ways that no animal consciousness does.

[12]See, e.g., Aizawa (2007, 23).

purely third-personal method for determining what experiences non-human subjects are having—such a 'consciousness meter' is the stuff of science fiction and seems likely to remain so.

To avoid a digital counterpart to factory farms or worse, then, we need some way of knowing what sorts of experiences candidate digital minds are having. And we need a way to leverage these assessments into recommendations for actions that will bend the trajectory of technological development away from outcomes with digital suffering. Here, a natural suggestion is that we should restrict the class of *advanced* digital systems that may be created, that is digital systems that are live candidates for being conscious. We'll take this suggestion on board, as it will allow us to set aside pocket calculators and iPhones in order to focus on preventing the creation of digital systems that are at significant risk of suffering. We will leave it open exactly how to delineate advanced from non-advanced digital systems. But we'll take it for granted that digital systems that enjoy human-level general intelligence count as advanced while anything one can presently buy at Best Buy does not. This leaves up for grabs whether existing systems such as MuZero and GPT-3 count as advanced.

Our strategy for solving the problem of digital suffering is *Access Monitor Prevent* (AMP). AMP has three parts. The first part proposes a means of gaining epistemic access to some digital minds. In particular, it uses a 'functional connectedness test' to enable us to discern what it's like to be a digital system from what functional states it's in. The test only applies to a limited class of digital systems, those that are functionally connected to ordinary humans via a 'dancing qualia argument'. On AMP, an advanced digital system may be created only if it passes the test. The second part is to monitor created systems to determine what sorts of functional states—and hence what sorts of experience—they are likely to undergo when deployed. The third part is to use information gained from monitoring these systems to prevent them from entering states with the functional markers of suffering. Since we can read off these systems' experiences from their functional states, this ensures that they are not suffering. And since these are the only advanced digital systems whose creation AMP permits, AMP ensures that any path to advanced digital systems will be suffering-free. Or at least this would be so if AMP were universally adopted. However, it should be borne in mind that, even if a global implementation of AMP turns out to be unrealistic, locally implementing AMP may nonetheless serve to mitigate the risk of digital suffering.

Before we proceed, a word about how we are understanding the notion of suffering. Recall that we characterize suffering as experiences with negative valence that are *pro tanto* morally bad to cause. By saying that an experience has negative valence, we mean that it is unpleasant for its subject to undergo, in some way or to some degree. On this characterization, suffering is not identical with (bodily or mental) pain. Indeed, it is plausible that there can be painful experiences without suffering. For instance, consider mildly painful experiences (such as those sought by spicy food enthusiasts) that essentially involve intense pleasure. Plausibly, such forms of pain are not *pro tanto* morally bad to cause and hence do not constitute suffering. That said, pain that is intense, long-lasting, and regarded as bad by its subject is likely to constitute suffering. Thus, while some forms of pain are relevant to the problem of digital suffering, the problem of digital suffering should not be conflated with that of preventing digital pain. Another important point is that not all forms of suffering are necessarily net negative: while we partly define suffering as being *pro tanto morally* bad to cause, this allows for the possibility that certain forms of suffering are morally neutral or positive on the whole. This could be because certain forms of suffering are constitutive of or crucial for attaining morally valuable knowledge or virtues.[13] In light of this, we will regiment the problem of digital suffering by restricting it to forms of suffering that are morally bad or morally catastrophic on the whole. For brevity, we will mostly leave this restriction implicit and write as if the task is to prevent digital suffering in general.

Before implementing AMP in any context, it would be appropriate to consider whether it should be adjusted in order to meet further moral desiderata (for instance, respecting the autonomy of digital agents). Here, our task is simply to develop one strategy for solving the problem of digital suffering.[14] Exploring the prospects for such strategies under further moral constraints is an important project, but one that we must leave for future research.

Here is the plan. In §2, we spell out AMP. In §3, we discuss how AMP would constrain current approaches to developing digital systems with advanced forms of intelligence. One approach turns out to fit especially well with AMP: whole brain emulation. As we will see, this supports AMP's viability and provides moral grounds for prioritizing whole brain

[13]For recent philosophical work on the value of suffering, see Brady (2018, 2019).

[14]Something like AMP could be developed in the service of moral desiderata that do not involve suffering, e.g. the promotion of flourishing. We develop AMP as a way of reducing the risk of digital suffering because a wide range of moral views would sanction that aim.

emulation over other approaches to developing advanced forms of digital intelligence. Unfortunately, AMP turns out to be subject to a major limitation: it only licenses the creation of a very restricted class of advanced digital systems. This limitation precludes benefits that would be realized through the creation of a broader class of digital systems, as there may be ways to architecturally modify digital minds that AMP forbids which would make those minds more capable of solving problems and less susceptible to suffering. Moreover, AMP's restrictiveness makes it less likely to be implemented, and hence less likely to succeed in preventing digital suffering. In §4, we explore options for modifying AMP to permit the creation of a broader class of digital systems while continuing to guard against digital suffering. We won't come down in favor of any one way of developing AMP. Instead, we will introduce several promising avenues for development while highlighting problems and tradeoffs they face. Our hope is that charting this philosophical terrain will put others in a position to make progress on the epistemic, moral, and technical problems that would need to be overcome in order to implement a strategy along the proposed lines and that this will in turn lead to work that will reduce the risk of digital suffering. One consequence of this may be a shift away from paradigms such as machine learning, which threaten to render digital minds epistemically inaccessible to us.

2. AMP

AMP aims to solve the problem of digital suffering by only permitting the creation of digital minds whose experiences are accessible to us, allowing us to use this knowledge to prevent these digital minds from entering suffering states. AMP's key innovation is its use of the *functional connectedness test* to achieve epistemic access to some digital minds. This section will unpack that test, show how it generates access to digital minds, and provide a more precise formulation of AMP.

We begin by describing the functional connectedness test. The test asks whether a candidate digital mind is *functionally connected* with some normally-functioning human. For two systems to be functionally connected is for there to be a gradual transformation of one to the other that preserves fine-grained functional organization. In this context, a transformation is a sequence of mappings from one nomically possible[15] system to another. A transformation can be concretely

[15]That is, allowed by the laws of nature.

implemented by modifying a system through replacement of its parts. Gradual transformations proceed through small steps in the design space composed of nomically possible systems. Crucially, for two systems to be related by such a transformation—and so to be functionally connected—it is not required that the transformation be implemented in any actual system. Thus, whether a digital system passes the functional connectedness test turns on what sorts of systems are nomically possible, not on which sort have in fact been created. *Normally-functioning humans* are conscious, rational adults who are free of interfering factors such as cognitive impairments, psychiatric disorders, drugs, and technological changes to their minds. An advanced digital system passes the functional connectedness test if it is functionally connected with some normally-functioning human and fails otherwise. To pass the functional connectedness test there need not be any actual human to which the digital system is connected. It is enough that there is some possible normally functioning human that the system is connected to in the sense we have here specified.

AMP accords the functional connectedness test the role of securing epistemic access to digital minds. The functional connectedness test yields a substantial form of access: we can know that digital systems that pass the test have experiences like those of the normally-functioning humans with which they are functionally connected. The source of this access to digital minds is the *dancing qualia argument*.[16] This argument requires unpacking. We'll start by giving the gist of the argument. We'll then offer some clarifications and provide a more formal rendition of it.

As an initial illustration of the dancing qualia argument, consider the transformation of your brain as you report on your experiences of a red apple before you. Initially, scientists seamlessly replace one of your neurons with a functionally equivalent silicon chip. In each subsequent step, scientists seamlessly replace another neuron with a functionally equivalent silicon chip, the end result being an isomorph of your brain made from silicon chips rather than neurons. Next, we observe that if that system has very different experiences at the end of this process than at the beginning, we should be able to generate a system with 'dancing qualia', experiences that flip back and forth in their qualitative

[16]See Chalmers (1996: Ch. 7; 2010a, 23–25; 2010b, 45–48). See Chalmers (1996: 253, Ch. 7, note 19) for related arguments and cases previously proposed by other authors. See Mandik (2017) for a superficially similar argument involving the gradual transformation of a subject in pain into a robot and Schneider (2019: Ch. 4) for a related 'chip test' for determining whether an artificially intelligent system is conscious.

character, say between red and green. If this situation were to occur, there would have to be some point in the series at which color experiences flip. We could then make a system with dancing qualia by transitioning back and forth between nearby members of the series which straddle this point. However, since the brain's functional organization is preserved throughout the entire series, this change would go entirely unnoticed. If asked, even as the color experiences are flipping back and forth, whether anything strange was happening the subject would reply, 'No, I'm still just looking at a red apple.' But it's implausible that such phenomenal differences would go unnoticed. It's also implausible that a nomically possible functional isomorph of a rational subject would exhibit such irrationality—after all, if highly irrational isomorphs of rational subjects are nomically possible, it would be a fortuitous coincidence that evolution generated rational subjects like us rather than our functionally equivalent but woefully irrational counterparts. This is in effect a *reductio* of the claim that there can be substantial phenomenal differences between systems that are related by a gradual function-preserving transformation to a normally-functioning human. As a result, the argument shows that the digital brain at the end of the series has experiences like those of the human brain at the beginning.

Some clarifications are in order. First, the dancing qualia argument relies on the notion of a (functional) isomorph of a brain. This notion can be defined relative to different levels of functional organization. Following Chalmers, we can understand isomorphs as entities that are functionally equivalent at the level of *fine-grained functional organization*, the level (whichever it is) that suffices to determine behavioral capacities. This means that a perfect isomorph of a human brain needn't duplicate its atomic structure, which would obviously rule out replacing neurons with silicon chips. For concreteness, we used a case featuring an isomorph at the neuronal level. However, the level of behavior-determining functional organization may in fact be higher or lower than the neuronal level.

Second, the argument appeals to oscillations between *nearby* members in the series. What counts as nearby? We need to answer this question in a way that plausibly precludes transformation between adjacent members in the series from completely destroying one subject and creating another anew, as unnoticed dancing qualia is only clearly a result to be avoided within a single subject. In light of this, we will (again following Chalmers) stipulate that systems are nearby just when each can be turned into the other via a transformation that replaces no more than ten percent of the original system.

Third, the argument requires that the series going from the human to the digital system should be gradual. But gradualness comes in degrees. Just how gradual is the series supposed to be? Well, the series is supposed to be gradual in order to allow for oscillations between nearby members of the series. So we will understand a series as gradual just in case every member in the series is nearby the adjacent member(s) in the just stipulated sense of nearby. In practice, isomorphs that are related by a minimally gradual series are likely to be related by much more gradual series as well.

Fourth, it should be borne in mind that we have opted for the stipulated notions of nearness and gradualness for specificity and because we think that they yield a formulation of the argument that is about as plausible as any. However, perhaps we are wrong about this and other precisifications of these notions yield a substantially more plausible argument. If so, then those notions could be substituted into the argument while leaving the rest of the discussion intact.

Fifth, the dancing qualia argument does not establish *functionalism*, the thesis that if two nomically possible systems are isomorphs, then they are phenomenally identical.[17] Functionalism doesn't follow because it concerns all pairs of nomically possible isomorphs, whereas this argument concerns only those that are functionally connected. To illustrate, consider your antimatter isomorph. Due to matter-antimatter interactions such as particle annihilation, there is no way to gradually transform either of you into the other while preserving functional organization. Plausibly, there are worlds with physics like ours in which matter and antimatter (or the realizers of their roles) have different categorical properties, ones that yield phenomenal differences (e.g. pain-pleasure inversion) between pairs of matter-antimatter isomorphs.[18] Yet if there are such worlds, there seems to be no reason to think we are not in such a world. Similarly, we may one day create digital isomorphs of ourselves that are made from materials that interface in disruptive ways with our biological makeup, rendering them functionally disconnected from us.[19] In that case, the argument would not license functionalism's prediction that such systems share our experiences. More generally, for all the dancing qualia argument shows, differences in realizers between

[17]This marks an important difference between our presentation of the dancing qualia argument and Chalmers's: he construes functionalism (in his terms the 'principle of organizational invariance') as the argument's target conclusion—though see Chalmers (1996: 272).

[18]See, e.g., Alter and Pereboom (2019: §1.1.1), Chalmers (2018, 48), and Lewis (2009).

[19]Compare: replacing an organ with one that realizes the same functional role can be extremely disruptive as a result of biochemical mismatches between the donor and recipient.

functionally disconnected isomorphs may induce phenomenal differences between them. Among these differences could be forms of suffering that we would erroneously conclude to be absent if we relied on functionalism rather than the functional connectedness test.

That the dancing qualia argument does not establish functionalism is important for two reasons. One is that if, contrary to fact, the argument established functionalism, then the class of advanced digital systems that the argument yields epistemic access to would be broader than the class AMP permits: the class would encompass all digital isomorphs of normally-functioning humans, not just those that pass the functional connectedness test. Another reason is that functionalism is a deeply controversial thesis among relevant experts. Hence we should, given the moral stakes, be wary of relying on functionalism, or any argument that entails it, in trying to solve the problem of digital suffering. In contrast, the dancing qualia argument should court less controversy. It is silent on the nature of experience. And, on the face of it, the dancing qualia argument can be used on any prominent theory of experience, including theories in the metaphysics of mind (such as physicalism, dualism, and Russellian monism), the philosophy of perception (such as representationalism and disjunctivism), and scientific theorizing about consciousness (such as the global workspace and recurrent processing theories).[20]

Sixth, the argument would collapse if we supposed for *reductio* that the digital input's experience *differs only slightly* from the human's. For in that case we would not expect oscillations between nearby systems to give rise to noticeable or rationality-wrecking phenomenal differences. This is why the argument supposes that the digital input's experience is *very different* from the human's. As a result, the argument only shows that the digital input's experience is not very different from the humans, not that they are phenomenally identical. But this suffices for AMP. To ensure that a digital system is not suffering, we do not need to know exactly what sort of experience it is having: it's enough for it to be functionally connected with a human whose experiences are very different from suffering experiences.

[20]This is not to say that the dancing qualia argument is entirely theory-neutral—it's not. With suitable adjustments, the argument can be run against at least two theories discussed in the literature: the integrated information theory (Tononi 2008) and tracking intentionalism (Dretske 1995; Tye 1995). But even these theories can embrace the argument if they are combined with certain accounts of phenomenal causation (Saad 2019a, 2019b). We are inclined to reject these theories on other grounds—for objections, see, e.g., Dalbey & Saad (forthcoming), Mendelovici (2018), and Pautz (2019). Insofar as they conflict with the dancing qualia argument, we see that as an additional reason to reject them.

With these clarifications in place, we can put the argument more formally as follows:

1. Let *DS* be a nomically possible digital system such that *DS* is functionally connected with some nomically possible normally-functioning human *H*—thus there is a gradual function-preserving series from *DS* to *H* in nomic space.
2. *DS* is a normally-functioning cognitive system.
3. If *DS* is a normally-functioning cognitive system and has very different experiences from *H*, then dancing qualia are nomically possible in a normally-functioning cognitive system: the series is populated by normally-functioning cognitive systems and function-preserving transformations within the series can induce large phenomenal changes (i.e. ones of a magnitude we would expect to be detected) in a normally-functioning cognitive system that are wholly undetected.
4. So, if *DS* has very different experiences from *H*, a nomically possible normally-functioning cognitive system is subject to dancing qualia. [1-3]
5. It is implausible that a nomically possible normally-functioning cognitive system is subject to dancing qualia.
6. So, it is implausible that *DS* has very different experiences from *H*. [4, 5]
7. So, plausibly, if *DS* is conscious, the experiences of *DS* and *H* are similar. [6]

This argument yields epistemic access to the conscious lives of digital systems that it takes as inputs. Those are precisely the digital systems that pass the functional connectedness test, a bar that digital systems must clear in order for AMP to permit their creation. But here AMP confronts a problem: digital systems that pass will be capable of suffering. This follows from the fact that we are capable of suffering and the argument shows that such systems would have experiences like ours. Indeed, we should expect that putting any permitted system into a state with the functional markers of suffering in humans will induce suffering in that system. AMP solves this problem by regulating the types of states that advanced digital systems may enter. In particular, it allows an advanced digital system to be created only if it is functionally connected to a normally-functioning human *and* would be prevented from entering states with the functional markers of suffering.[21] Given that humans do not

[21] Alternatively, rather than requiring that advanced digital systems be prevented from entering states with the functional markers of suffering, AMP could be formulated to require that digital systems only be allowed to enter states that functionally mark clear non-suffering. We regard the latter

suffer unless such markers are present, this will then ensure that the digital systems do not suffer either. The functional markers of suffering should here be understood broadly, so as to include not only reports and other overt behavioral markers, but also whatever we learn about the neural-functional signatures of suffering from current or future cognitive science.

Different implementations of the strategy would prevent digital systems from entering states with the functional markers of suffering in different ways. One option is to shield digital systems from stimuli that would induce suffering. Another is to modify how subjects process stimuli so as to prevent noxious stimuli from inducing suffering. However, AMP precludes interventions that sever direct functional connections between humans and digital systems by, for instance, just removing the 'pain center' from the digital mind, since the resulting system would not be functionally connected with a normally-functioning human. This rules out a range of architectural modifications that might be used to make digital systems less susceptible to suffering or more capable. We will explore some options for relaxing AMP to allow such modifications in §4.

3. Where does AMP lead?

AMP permits the creation of an advanced digital system only if that system is functionally connected to a normally-functioning human. Since advanced digital systems with non-human cognitive architectures are not functionally connected to normally-functioning humans, AMP therefore prohibits the creation of such digital systems. Among the digital systems whose creation AMP forbids are 'superintelligent' digital systems—ones whose intelligence vastly exceeds human intelligence—that have non-human cognitive architectures. To appreciate the contours and severity of this limitation, we need to explore how AMP constrains the creation of different types of digital minds. That is the task of this section. In later sections, we discuss ways of modifying AMP to partially overcome this limitation.

Many approaches to creating advanced digital systems would, if successful, generate advanced digital systems that would fail the functional connectedness test. For instance, traditional 'Good Old-Fashioned

requirement as morally preferable and believe it could replace the former requirement without much affecting our discussion. We put matters in terms of the former for expositional ease.

Artificial Intelligence' (GOFAI), sought to produce intelligent systems by programming explicit rules for manipulating symbols. An 'intelligent' chess-playing system, say, would be programmed with specific rules for playing chess (control the center of the board, develop your minor pieces early, etc.). In creating such systems, little if any attention is paid to how these problems are solved in the human mind. As a result, it is highly improbable that advanced digital systems created on this approach would be functionally isomorphic with (much less functionally connected to) human minds. Thus, AMP blocks this path to advanced digital systems.

Trends in artificial intelligence (AI) research have moved away from the GOFAI model towards machine learning. Machine learning algorithms automatically improve at a task as they are applied to more inputs ('training data'). These algorithms are often implemented on 'neural networks', which consist of artificial nodes. The connections between nodes have different weights, which evolve over time as the network is trained on more data. Whereas a GOFAI approach to creating a chess-engine involves programming explicit chess strategies, a machine learning approach forgoes use of explicit information about chess in favor of a 'reward function' that scores different outcomes and a 'learning algorithm' for optimizing score. These systems are capable of developing great skill at the game, and some contemporary versions do not even need to have the rules of chess hard coded into them—they can learn those along the way. By playing more games of chess, the system uses machine learning algorithms to maximize cumulative reward (or minimize loss), which in this case comes to winning chess games. Eventually, a highly capable chess playing system is produced, while receiving little or no explicit guidance.

Machine learning approaches have demonstrated a remarkable ability to match or exceed human capabilities in a wide array of domains (chess, Go, image recognition). However, there is no reason to think that systems built on these principles will be functional isomorphs of human minds. While there is an obvious analogy between brains and machine learning systems, their underlying architectures differ greatly: artificial nodes in existing machine learning systems function differently than neurons and differ in number and organization from their intracranial analogs. Even if there are possible machine learning systems with the brain's cognitive architecture, the space of machine learning systems is vast. So there is little chance that advanced machine learning systems will functionally mirror human cognition; *a fortiori*, there is little chance that their creation will be permitted on AMP.

That GOFAI and machine learning approaches to creating advanced digital systems are not conducive to producing digital minds that are functionally connected to human minds raises two worries.[22] One is that AMP entirely precludes the creation of human-like intelligence in digital systems. This would be the case if there is no viable path to creating intelligent digital systems that can pass the functional connectedness test. The second worry is that even if *some* advanced digital systems can pass the functional connectedness test, the test's strictures mean that such minds will, at best, be equivalent to the human mind in cognitive capabilities. That in turn threatens to render the strategy unviable, as one of the main goals of AI research is to create digital systems that exceed human intelligence.

Fortunately, there is at least one approach to creating advanced digital systems that is especially promising by AMP's lights: *whole brain emulation*.[23] The goal of whole brain emulation is to create a digital system that functions like a brain. This digital system is created by scanning a brain (or set of brains), uploading the scan to a computer, and running an emulation algorithm that operates on the scan, yielding brain-equivalent outputs. By construction, whole brain emulations would functionally mirror human brains to a selected level of detail and so be excellent candidates for systems that are functionally isomorphic with humans. Whole brain emulations would be run as software. Whether AMP permits a given whole brain emulation will depend on whether the whole brain emulation is functionally connected with a normally-functioning human.[24]

That AMP meshes well with whole brain emulation would be of idle curiosity if whole brain emulation were not a viable path to digital minds. But, in fact, there is a case to be made that it will become feasible in the next century, and that it will be the first path taken to general AI.[25] After all, the human brain is the closest thing we have to a domain-general superintelligent system at the moment. The requirements for creating whole brain emulations are also relatively well-understood. In

[22]The same goes for other potential approaches to AI—such as 'seed' AI and artificial evolution—since they too are unlikely to produce systems that functionally mirror humans.

[23]For background on whole brain emulation see Sandberg and Bostrom (2008). For skeptical concerns about their prospects, see Mandelbaum (2022); for a rebuttal see Hanson (2022).

[24]For reasons to think that whole brain emulations would be functionally connected with humans, see Chalmers (1996: Ch. 7).

[25]Eth, Foust, and Whale (2013) argue that whole brain emulation will likely arrive by 2063. Sandberg (2014a) offers a model for when whole brain emulations will arrive that accords a 50% probability to their arriving before 2064 and a 75% probability to their arriving by 2080. Kurzweil (2005, 197) predicts the arrival of the tools required for whole brain emulation in the 2020s. Hanson (2016) predicts that whole brain emulation will arrive roughly within the next century and that they will be the first era-transformative form of artificial intelligence.

broad outlines, the task is to scan a human brain (or set of human brains) so as to determine its (their) neuronal circuit diagram. Although the technology to do this at the scale of an entire human brain does not yet exist, there do not appear to be any in-principle barriers to doing this. In addition to knowing the overall connection diagram of a given brain, whole brain emulation will also require an adequate model of the dynamics of neural processing. It must be determined how information is processed by groups of neurons, or even within individual neurons. With this information, the task is then to develop an emulation of these neural processes on digital hardware. If these processes are emulated with enough fidelity, there is no reason why a digital system, when given the same inputs as the human brain it is emulating, should produce relevantly different outputs.

Indeed, whole brain emulation promises to surpass, not merely match, human processing speed and power. According to some estimates, whole brain emulations could be run up to a million times faster than their human counterparts.[26] Whole brain emulation might also be used to generate a collective form of superintelligence. For instance, a whole brain emulation might be tasked with a problem and allowed to create copies of itself to work on sub-problems or plugged into existing proposals for AI safety or intelligence amplification that enlist human agents.[27] Thus, because of its inherent speed advantages, and the opportunities it affords for cooperative cognition, whole brain emulation would likely yield a rapid transition from general (human-level) AI to AI that is superintelligent, at least in terms of processing speed.

Whole-brain emulation would also have other advantages. Since these systems would have minds organized just like human minds, they would initially reason and act in accordance with human values. Hence, this approach to the creation of digital minds would appear to resolve the 'value-loading problem' of instilling our values in advanced digital systems and make headway on the closely related 'value alignment problem' of ensuring that powerful digital systems are reliably aligned with our values.[28] In the event that other types of superintelligent digital systems are created, value-aligned whole brain emulations would likely prove useful in helping us monitor those systems and prevent them from causing undesirable outcomes: the value-alignment of whole brain emulations would render them trustworthy; their

[26]See Hanson (2016, 10) and Bostrom (2014: Ch. 3, note 3).
[27]See Irving, Christiano, and Amodei (2018) and Hubinger (2020).
[28]See Bostrom (2014: Ch. 12), Russell and Norvig (2021: Ch. 1), and Wiener (1960).

superintelligence would put them in a better position to understand other forms of superintelligence than we are; and our shared cognitive architecture would put them in a position to explain other advanced digital systems to us in terms we would understand. Whole brain emulation also offers a plausible bridge to other forms of superintelligence.[29] Any paths to other forms of superintelligence that can be taken without whole brain emulation may be able to be taken much more quickly with it. Finally, with suitable uploading technology, whole brain emulations might offer humans digital afterlives.[30]

To sum up, while AMP severely constrains the development of advanced digital systems by blocking GOFAI and machine learning paths, it does not entirely foreclose their development. Indeed, AMP fits well with the whole brain emulation path, a path that holds independent appeal. Even so, it is worth considering whether AMP can be safely modified so as to partially overcome its limitations. To preview, while our investigation will uncover a range of promising relaxations of AMP, these will mainly serve to enhance degrees of design freedom on the whole brain emulation path, not to open up other independent paths to digital minds. Absent justification for more ambitious relaxations of the epistemic access requirement or the development of other strategies for preventing digital suffering, we should regard paths to digital minds other than whole brain emulation as morally treacherous. This conclusion applies to the currently dominant paradigm of machine learning, which at the moment affords little epistemic access to the inner workings of complex machine learning systems and so *a fortiori* little if any access to any experiences such systems may have now or (more realistically) in the future.

4. How AMP'S limitations might be overcome

There are two main dimensions along which AMP might be improved. First, the strategy might be made *safer* by better guarding against the risk of suffering in systems it permits, conditional on the strategy being implemented. Second, the strategy might be made more *viable*. After all, to have a shot at being implemented and reducing the risk of digital suffering, AMP must not be technologically infeasible or prohibitively costly to implement. Other things equal, a more viable strategy

[29]See Bostrom (2014: Ch. 2) and Chalmers (2010b, 18–19)
[30]See Chalmers (2010b) and Schneider (2019).

will have a better shot at reducing the risk of digital suffering. Since our aim is to develop a strategy that reduces risk of digital suffering overall (not just conditional on its being implemented), it is therefore worth exploring the prospects for improving the strategy along both dimensions.

We can distinguish at least four ways of revising AMP to yield improvements in safety or viability:

- Apply the functional connectedness test to a broader class of humans (§4.1)
- Relax the functional connectedness criterion (§4.2)
- Apply AMP recursively (§4.3)
- Allow advanced digital systems to be created even if they will suffer in certain ways (§4.4)[31]

For simplicity, we will mostly consider these proposals for modifying AMP in isolation from one another, though, in principle, any of them could be combined with any of the others.

4.1. Apply the functional connectedness test to a broader class of humans

To make AMP safer or more viable, we might allow the functional connectedness test to apply to a broader class of humans, thereby potentially yielding a broader class of advanced digital systems whose creation is permitted. There are various ways to accomplish this. For instance, we could allow all humans as inputs, not just those that are normally-functioning. However, there would be no point in running the functional connectedness test on epistemically opaque humans: if a human's experiences are inaccessible to us, then the functional connectedness test cannot leverage our access to its experiences into access to a digital mind's experiences. Thus, we should instead apply the functional connectedness test to a broader class of humans in ways that do not compromise it. We'll consider the following promising approaches for expanding the class of human inputs while meeting this constraint:

- Allow certain unmodified human inputs that are not normally-functioning (§4.1.1)

[31] A fifth approach that would require a paper of its own to address is passing the buck for improving AMP to digital systems. For relevant discussion, see Bostrom (2014: Ch. 13) and Yudkowsky (2004).

- Allow certain pharmacologically modified human inputs (§4.1.2)
- Allow certain technologically modified human inputs (§4.1.3)[32]

4.1.1. Unmodified humans that are not normally-functioning

The functional connectedness test only takes human inputs that are normally-functioning, and so does not apply to individuals with cognitive impairments or psychiatric conditions. As a result, AMP forbids the creation of digital systems that are functionally connected to these humans. But some of these humans have readily epistemically accessible experiences and cognitive or sensory capabilities that normally-functioning humans lack. Extending the functional connectedness test to these humans would permit the creation of digital systems that are functionally connected to these humans and which share their distinctive capabilities. We'll illustrate the potential for improving AMP in this fashion with some examples.

Consider non-normally functioning individuals with cognitive capacities that normally-functioning humans generally lack. These include the many individuals from history with psychiatric impairments who made intellectual breakthroughs. Plausibly, if isomorphs of these individuals were posed with suitably selected contemporary problems, some would make field-advancing contributions—one imagines that a digital isomorph of, say, Cantor or Gödel would be able to find a research niche in contemporary mathematics or logic. In addition, individuals with savant syndrome often pair extraordinary abilities in domains such as calculation and memorization with more general cognitive impairments. Savant syndrome demonstrates that human beings can perform 'superhuman' cognitive tasks such as memorizing *The History of the Decline and Fall of the Roman Empire* and reciting it backwards and forwards. Isomorphs of individuals with savant skills would themselves possess such skills.[33] Since savants' abilities in tasks such as memorization and calculation are orders of magnitude greater than those of ordinary humans, extending the functional connectedness test to these individuals would enable AMP to permit the creation of digital isomorphs of humans with these 'superhuman' cognitive capabilities.

[32] Another approach to revising AMP would be to expand the class of admissible inputs to include non-human animals. We leave the exploration of this approach as a task for future research.

[33] See Treffert (2009, 1352). There are also reported cases of neurotypical individuals acquiring savant skills as a result of brain injury (*ibid*: 1354). Savant syndrome can occur alongside a wide variety of other conditions including (though not limited to) autism and dementia, and its occurrence is not currently well-understood.

There are potential moral drawbacks to this approach. It is morally questionable whether it is permissible to create impaired individuals one knows will be impaired. Such worries would extend *mutatis mutandis* to digital minds. At the same time, impairment is a relative notion. An ordinary functioning human is impaired relative to many possible digital systems. Unless we are willing to accept that it would be wrong to create humans once there is the option of creating cognitively superior beings, we should perhaps reconsider our resistance to creating human or AI subjects with certain impairments who would live worthwhile lives free of suffering.[34]

4.1.2. Pharmacological modifications

Another suggestion for enhancing AMP is to extend the functional connectedness test to humans who have enhanced cognitive capabilities or a diminished suffering capacity while under the influence of certain mind-altering drugs. By extending the functional connectedness test to these individuals, AMP can permit the creation of digital minds that share their advantages over normally-functioning humans, potentially making AMP more viable and safer.

Substances such as caffeine, amphetamine, and modafinil are apparent cognitive enhancers, drugs that enhance capabilities in certain cognitive domains such as attention, memory, or focus.[35] Substances such as psilocybin, LSD, and DMT are potent psychedelics, which appear to improve performance in tasks involving creativity and divergent thinking.[36] Pain-relieving opiates (morphine, oxycodone) and dissociatives (nitrous oxide, ketamine) operate in the central nervous system to relieve or prevent pain and suffering in subjects.[37] These drugs, or future drugs, may extend human cognitive capacities beyond their ordinary boundaries or prevent suffering. And while there is the potential for misuse or abuse of any drug, use of these drugs does not necessarily lead to suffering. By extending the functional connectedness test to subjects who are under the influence of such drugs, AMP may permit the creation of digital minds whose cognitive capacities exceed those of normally-functioning humans or whose suffering capacities are lesser than those of normally-functioning humans.

[34]Cf. Parfit (1984: §122).
[35]See Dresler et al. (2019)
[36]See Kuypers et al. (2016).
[37]See Dickenson et al. (2013) and Hill (2013).

4.1.3. Technological modifications

A third option for enhancing AMP is to allow technologically modified humans to serve as inputs to the functional connectedness test. Technological modifications promise to enhance human cognitive abilities without drastically altering our underlying cognitive architecture. Using such cognitively enhanced humans as inputs to the functional connectedness test could improve AMP's viability by permitting the creation of digital minds with greater capabilities.

The most salient types of technological modifications are those that directly interface with the human brain so as to enhance its capacities. Crude versions of these technologies are currently under development.[38] These include non-invasive brain-to-brain interfaces that allow for direct brain-to-brain communication.[39] In the future, brain-computer interfaces may offer a way to enhance our memory, processing speed, resilience to cognitive decline, and communication abilities. Humans equipped with interfaces that do not directly affect processing underlying consciousness would presumably have experiences quite like our own, but would have greater—potentially much greater—cognitive capabilities. Using them as inputs to the functional connectedness test would enable AMP to permit the creation of digital minds with enhanced capabilities, with little loss of safety. Whether such technologies pan out remains to be seen, but if they do they could enhance the types of digital systems that pass the functional connectedness test and hence improve AMP.[40]

4.2. Relax the functional connectedness criterion

AMP permits an advanced digital system to be created only if it is functionally connected with a normally-functioning human. The next approach proposes to relax the criterion for functional connectedness in order to make AMP safer or more viable by enabling it to permit the creation of a broader class of digital systems. We'll consider two versions of this approach.

4.2.1. Weakening via coarse-graining or restriction

We turn now to the first option for enhancing AMP by relaxing the operative notion of functional connectedness. On this option, functional

[38]For instance, Elon Musk's brain-computer interface company, Neuralink, has implanted computer chips in monkeys that interface with their brains and allow them to play video games via the interface (Kay 2021).
[39]See Jiang et al. (2019).
[40]For discussion of the limitations of such technology, see Bostrom (2014: Ch. 2).

connectedness is defined in terms of a weaker notion of isomorphy. There are two overlapping families of such notions. One consists of functional isomorphy within a restricted range of circumstances. The other consists of functional isomorphy at coarser levels of grain than that of fine-grained functional organization. For instance, one such notion is that of functional isomorphy within the restricted circumstances of playing chess and at the coarse-grain level of chess moves, a level that abstracts away from finer-grained behaviors that the system uses to implement those moves. Any such weaker notion of isomorphy can be plugged into the original definition of functional connectedness to yield a weaker notion of functional connectedness: two systems will be functionally connected in such a weaker sense just when they are related by a gradual transformation that preserves functional organization at a specified coarser level of grain or in a specified restricted range of circumstances.

This proposal can be taken too far. Given a sufficiently coarse-grained or restricted notion of functional isomorphy, we can define a notion of functional connectedness on which virtually any pair of systems will be functionally connected. For instance, on a sufficiently coarse-grained notion of functional connectedness, humans will be functionally connected to laptops and superintelligent systems with radically inhuman cognitive architectures. Of course, we cannot define epistemic access to the experiences of systems into existence. So this approach runs the risk of severing the link between functional connectedness and epistemic access. Since the whole point of AMP is to use that access to mitigate the risk of digital suffering, it is crucial that that link be left intact. Here, the dancing qualia argument offers a useful constraint. To check whether a given notion of functional connectedness preserves epistemic access, we can evaluate a dancing qualia argument that is run on a human and candidate digital mind that are functionally connected in that sense but not in stronger senses. Just how much the notion of functional connectedness can be weakened before the dancing qualia argument ceases to link it with epistemic access is an open question that merits further investigation.

It is difficult to predict in advance just which sorts of digital systems would be permitted by a variation of AMP that weakens the criterion of functional connectedness in this fashion. Still, we can identify several potential advantages of this approach. One is that the broader class of systems whose creation is permitted on this approach may include digital systems that are less susceptible to suffering. For instance, perhaps the resulting extension of AMP would permit the creation of a

coarse-grained isomorph with a dormant 'pain-center'. Another potential advantage is that the broader class of digital systems may include digital systems with cognitive capabilities that exceed those of normally-functioning humans. For instance, perhaps this approach would permit the creation of digital systems that are restricted-isomorphs of normally-functioning humans but which have superhuman memory.

A third potential advantage of this approach is that it may render the strategy more verification-friendly. To implement AMP, we would need to somehow determine that certain advanced digital systems would be functionally connected with humans. Developing verification procedures for this purpose is a formidable technical problem. Empirical procedures of this sort—ones involving the actual construction of a gradual series—are infeasible and would in any case be of no use in implementing AMP: since applying these procedures would entail creating advanced digital systems, AMP would forbid applying these procedures to advanced digital systems whose creation AMP has not already permitted. Thus, theoretical procedures—ones that determine whether an advanced digital system would be functionally connected with humans, but without actually constructing a gradual series—will be needed.[41]

Plausibly, the more complicated a digital system is, the more difficult it will be to theoretically verify that it is functionally connected to a normally-functioning human. Thus, one promising way to make AMP more viable would be to make it more verification-friendly by permitting the creation of simpler digital systems that are restricted or coarse-grained isomorphs, rather than only permitting the creation of unrestricted fine-grained isomorphs. Digital simplification may be achieved by degrading or eliminating features of human cognitive architecture that are irrelevant for the purposes to which the digital minds would be put. Since our minds contain much excess functional structure relative to what we use in many contexts, there is much room for digital simplification.

[41]These procedures need not rely on untutored intuition. They might, for example, rely on experimental results from neuroscience, biochemistry, and material sciences, along with computer modeling and atomically precise manufacturing, to verify that the analogs of neurons in digital systems are functionally isomorphic with their neural counterparts. Automated procedures could use such resources to verify that interactions between neurons and chips would not introduce functional disruptions in intermediate cases between a human and a digital isomorph. Advances in these areas would be precursors to the capacity to create advanced digital systems such as whole brain emulations. This provides grounds for optimism that theoretical verification procedures will be available by the time that we are in a position to create whole brain emulations.

4.2.2. Weakening via extension to modifications

Next, we'll consider the second option for enhancing AMP by relaxing the operative notion of functional connectedness. This option extends the notion of functional connectedness to certain digital systems, courtesy of their being suitably modified versions of digital systems that are (in the original sense) functionally connected to humans. To explore this approach, we distinguish two broad sorts of modifications: those that aim to make digital systems less susceptible to suffering (and so enhance AMP's safety) and those that aim to improve the capabilities of the digital systems (and so enhance AMP's viability).

Let's start with suggestions for modifying digital systems to make them less susceptible to suffering. If any of these suggestions withstand scrutiny, we can use it to extend the notion of functional connectedness, thereby enabling AMP to permit the creation of less suffering-susceptible digital systems.

An obvious suggestion is to make digital systems less susceptible to suffering by rendering them unconscious. Since we do not yet know what the true theory of consciousness is, we cannot look to it to tell us what to deprive digital systems of in order to render them unconscious. Still, we could make it less probable that a digital system suffers by depriving it of live candidates for necessary conditions for consciousness. For illustration, suppose we take the following as live candidate necessary conditions for consciousness: being accessible to a global workspace, having a maximal quantity of integrated information, being the object of a higher-order mental state, or being in a state that bears a certain tracking relation to features of the environment.[42] The revised strategy would then allow digital systems that are functionally connected with normally-functioning humans to be modified in ways that deprive them (the systems) of some or all of these properties, thereby yielding systems that are functionally disconnected from normally-functioning humans. As a precaution, these systems should probably be deprived of these properties throughout their (the systems') existence: while creating systems that will never be conscious seems relatively innocuous, it's plausible that creating a conscious system and then depriving it of consciousness would be bad for it in much the way that death would be. Similarly, caution would be needed when depriving a system of a candidate necessary condition for consciousness is apt to cause suffering—for instance, depriving a system of a global workspace might lower the

[42]For an overview of scientific theories of consciousness, see Seth and Bayne (2022).

probability of consciousness while raising the probability of goal frustration to an event greater extent, resulting in a net increase in suffering risk.

Another suggestion is to make digital systems less susceptible to suffering by reducing or eliminating their pain capacity. Some human beings, for example, lack normal pain experience. In the condition known as *pain asymbolia*, subjects' pain experiences retain their sensory/discriminative dimension, but lose their negative valence.[43] Subjects born with *congenital insensitivity to pain,* meanwhile, lack the ability to feel bodily pain entirely. While this may sound like a blessing, it is in fact a curse, as subjects with this condition invariably lead short, injury-prone lives. Although the empirical details of these conditions are messy, they indicate some flexibility in our brain's ability to experience pain and pain affect. A more complete understanding of the neural correlates of pain experience may allow us to precisely target those brain regions responsible for the experience of pain and suffering. We may then be able to engineer digital minds based on the human brain architecture with an artificial form of pain asymbolia that eliminates pain's unpleasantness without any of the odd side-effects that standardly accompany the condition. Similarly, by studying the brain effects of pain-relieving substances such as morphine and ketamine, we may be able to engineer digital minds that enjoy the analgesic effects of these drugs while avoiding their negative cognitive side effects. Since these interventions would be directly based on conditions in human minds that are systematically related to pain and suffering, changes of these sorts could be expected to diminish digital suffering without compromising other capabilities of digital minds.

A related but more theoretical proposal is to remap the stimuli and responses associated with affective processing in digital systems that are functionally connected with normally-functioning humans. One option would be to contract the range of negative outputs, so that the most noxious stimuli induce only mild pains. This approach might be order-preserving, such that whenever one stimulus induces a better affective response in humans than another stimulus, the former stimulus also induces a better affective response in digital systems. Another option would be to shift the affective scale so that stimuli systematically produce affectively better (more positive or less negative) states in digital minds than they would in humans. For instance, one option would be to shift the scale so that stimuli produce only positive affective states in digital

[43]See Klein (2015).

systems. This sort of shift could be both order-preserving and distance-preserving (i.e. such that the distance between affective states produced by different stimuli in humans is the same as the distance between affective states produced by those stimuli in digital systems). Such modifications would render digital systems less susceptible to suffering while retaining important structural features of affective processing in humans, features that have some appeal in the context of building digital minds that are value-aligned with humans.

Finally, we might make digital systems less susceptible to suffering by modifying their cognitive states. Whether a state counts as suffering arguably depends in part on what sorts of cognitive states accompany it. For instance, the pain and negative affect endured by a person running a marathon or performing a heroic act of self-sacrifice may fail to qualify as suffering because they are reflectively endorsed as means to the subjects' ends. Thus, there may be ways of reducing suffering in digital systems by suitably modifying their cognitive states to prevent would-be suffering states from so qualifying, or by making such states less severe as forms of suffering than they would otherwise be.

One concern about this last sort of modification is that it would involve a morally questionable form of 'digital brain washing'. There is something to this worry: certain forms of cognitive manipulation seem morally problematic—for instance, programming digital systems to falsely believe that they will be rewarded for their suffering. In particular, one might worry that such interventions would violate the autonomy of digital systems and so are morally objectionable on that ground. This is a serious worry. At the same time, however, it should be borne in mind that creating digital minds, by its very nature, involves cognitive manipulation. In creating digital minds and using them for our purposes—the goal of AI research—we are shaping the cognition of such systems from the very start in order to achieve certain goals. This is an unavoidable consequence of the fact that digital minds are artificial creations. Concerns about cognitive manipulation may thus prompt a general argument against the creation of digital minds *tout court,* but such concerns apply *mutatis mutandis* to pretty much any proposal for shaping the cognition of digital minds in accordance with some predetermined human-set goal. Thus, while the issue of what sorts of cognitive modifications should be allowed deserves careful consideration, those who are open to the creation of digital minds at all should not balk in principle at this proposal as a way of enhancing AMP's safety.

We turn now to modifications that aim to enhance digital system capabilities. Ideally, these modifications would not increase the risk of suffering conditional on the strategy being implemented. But even such modifications that increase that risk could nonetheless decrease the unconditional risk of digital suffering. They could do this by making the strategy more viable and so more likely to be implemented over riskier alternatives.

Capabilities that are not tied to consciousness are natural targets on this approach: given such a capability in a digital system that is functionally connected with a normally-functioning human and a way of enhancing the capability without affecting conscious processing, we could create an enhanced digital system that would be free of suffering so long as the unenhanced digital system is free of suffering when in corresponding (*modulo* differences in the targeted ability) states. However, in practice, there may be no such capabilities that we could confer to digital systems. Even mental phenomena such as long-term memory and language processing that are largely unconscious occasionally affect what experiences subjects have, as when they are consciously remembering or experiencing speech.

A potentially more feasible approach would enhance abilities in ways that only realize experiences that humans are capable of. On this approach, we might be permitted to create an advanced digital system that can fluently process dozens of human languages because the system can only have experiences of sorts that humans can have. This approach would require us to identify the functional markers of human experiences—otherwise we will not know whether a given enhancement would only lead to experiences of the sort humans can have. While we do not yet have a systematic catalog of the functional markers of human experiences, it's still early days in neuroscience and we have at least begun to discover impressive neural-phenomenal structural correlations.[44]

Should we find ourselves in the possession of such functional markers, it will be worth trying to use general patterns in functional data along with principles of phenomenology to further extend the range of admissible experiences. To illustrate, suppose that we discover not only the functional markers for human experiences but also general compositional regularities concerning how functional markers of simple experiences

[44]See, for example, Forder et al. (2017), Brouwer & Heeger (2013), and Coghill et al. (1999). For an overview of some phenomenal-neural structural correlations and a discussion of their philosophical implications, see Pautz (2014).

combine to yield functional markers of complex experiences. This could afford us a sort of indirect access to complex experiences that we cannot ourselves undergo. Indeed, it would yield access to whether certain experiences that we cannot ourselves undergo involve suffering. For the following is a plausible principle: if two experiences are valence-neutral (i.e. neither pleasant nor unpleasant), a complex experience consisting of just those experiences does not involve suffering.

To appreciate the plausibility of this principle, consider a non-synesthete who wonders whether sound-color synesthesia is a form of suffering. Intuitively, their curiosity would be misplaced. Given this principle and the noted functional information, we would be assured that states of advanced digital systems that are composed in the relevant way from functional states that mark valence-neutral experiences would themselves mark experiences that do not involve suffering—this would be so even when humans cannot have those experiences. Thus, we arrive at the following extension of the previous approach: allow advanced digital systems' abilities to be enhanced in ways that only realize experiences that humans are capable of *or* combinations thereof that we can certify as free of suffering.

Other functional patterns and principles of phenomenology might be leveraged to similar effect. The following are plausible principles of this sort:

> *Sandwiching*: If two experiences do not involve suffering, then any experience between them in a phenomenal quality space does not involve suffering.
>
> *Neutral extension*: If all the experiences within an extended region of phenomenal quality space are valence-neutral, then no experience that can be reached by expanding that region without adding dimensions involves suffering.
>
> *Positive-direction extension*: If $e_1 \ldots e_n$ is a series of experiences that are correspondingly ordered in a phenomenal quality space and their valence is positive and increasingly so, then any experience e_m that belongs to a linear continuation of that series in phenomenal quality space does not involve suffering.[45]
>
> *Spatiotemporal modulation:* Any purely spatial or temporal modulation of valence-neutral experience is itself valence-neutral. (Purely spatial modulations change only features of experiences, such as phenomenal size, shape, and spatial dimensionality. Purely temporal modulations change only features such as phenomenal temporal ordering and duration.)

[45]We equate the positive valence of an experience with its degree of pleasantness. This is relatively standard in the literature. For instance Jacobson (2021, 481) says that an experience is valenced if 'what it's like to undergo [it] can be pleasant or unpleasant—it can feel good or bad to some degree.'

If these principles are correct, they open the way to safely enhance digital systems that pass the functional connectedness test in ways that realize a wide range of experiences outside our own. As with the phenomenal composition principle introduced above, to use these principles to extend the range of admissible experiences would require accompanying functional information. For instance, to apply Sandwiching, we would need to know when a functional state marks an experience that is between those marked by other functional states.

It remains an open question whether something like this approach to capability enhancement will turn out to be a feasible way to enhance AMP's viability. One source of uncertainty here is whether we will come to possess the requisite functional information. A second is whether, given such information, we could use it to construct systems with abilities that realize experiences that go beyond our own. Given a favorable resolution of these two issues, there would then be a question as to whether some such systems would be useful enough to enhance AMP's viability. Here, there are grounds for cautious optimism. Given the splendid ends to which humans have put our limited experience-enabled sensory discrimination abilities and our three-dimensional visualization abilities, it stands to reason that there would be further useful ends to which a system with otherwise similar cognitive abilities might put greatly enhanced sensory discrimination abilities or higher dimensional visualization abilities.

4.3. Recursion

The next suggestion for enhancing AMP: apply it recursively, allowing any system whose creation is permitted by AMP to serve as a further input to the strategy by way of the functional connectedness test.

By itself, recursion cannot extend the range of digital minds whose creation AMP permits: functional connectedness is transitive, since if there is a gradual functional-organization preserving transformation A and B and there is another such transformation between B and C, then performing the first transformation and then second will yield a gradual functional-organization preserving transformation going from A to C.

Nonetheless, recursively applying otherwise enhanced versions of AMP may extend the range of digital minds whose creation AMP permits. Indeed, this approach lends to powerful forms of amplification. To illustrate, suppose we broaden the class of admissible inputs to the functional connectedness test to include any system that passes the functional

connectedness test as well as modest modifications of such systems, ones that have, say, been given a 1% greater memory capacity. By repeatedly taking digital systems that pass the functional connectedness test, memorially enhancing them, and using them as further inputs to the test, this approach could in principle permit the creation of advanced digital systems with arbitrarily large memories.

This sort of amplification could be used to permit the creation of digital systems with any number of amplified capabilities aside from memory. There are also several stronger forms of amplification in the vicinity. One forgoes using a single sort of modification in favor of a menu of modifications, one item from which is applied on each iteration. Another uses a menu of modifications and applies some combination of those modifications on each iteration.

In addition to enabling AMP to permit highly capable digital minds, recursively applying AMP has another virtue: it allows substantial capability improvements to be generated via modest improvements that are tested piecemeal. This reduces the risk of introducing suffering into digital systems via unnoticed interference effects that might arise when modifications are induced in concert. However, the iterated applications of the functional connectedness test would somewhat weaken our epistemic access to the phenomenology of minds whose creation is permitted, at least by way of comparison with digital minds that may be created on AMP without recursion. Hence, we can expect recursion to yield gains in viability albeit with costs to safety. The extent to which iterated applications of the functional connectedness test diminishes epistemic access and in turn safety is unclear. Given the potential for recursive extensions of AMP to improve its viability, this is a topic that merits further investigation.

4.4. Moral relaxation

We have thus far focused on ways of relaxing AMP by modifying its criteria for what types of advanced digital systems may be created. A different way of relaxing it instead modifies what types of states such systems must be prevented from entering. AMP reflects the moral aim of preventing suffering in digital minds and so forbids creating digital systems that enter states with the functional markers of suffering. However, some advanced digital systems may be such that there is no way to usefully deploy them while ensuring that they never enter a functional state that marks even the slightest amount of suffering. Even if we

could amply compensate these beings and they would endorse our causing them to exist and suffer, AMP would forbid their creation. Since deploying such systems might yield great benefits, it may be worth relaxing the strategy to allow for the creation of advanced digital systems that undergo certain limited forms of suffering.

A modest relaxation would license the creation of minds that suffer only if their lives are worthwhile, the sufferers somehow benefit from the suffering, and there is impersonal benefit to their suffering. Less scrupulous relaxations of the strategy would adopt some but not others of these constraints. However, such relaxations would invite controversy. For instance, while allowing any suffering in a digital mind that is impersonally good would be licensed by some forms of consequentialism, it would be forbidden by some non-consequentialist views, e.g. because it would license using an agent as a mere means. Absent convergence in moral theorizing, it would seem unwise to resort to a highly partisan form of relaxation. A better option would be to identify forms of moral relaxation that are compatible with a wide range of moral theories.[46]

While we will not attempt such development here, it is worth flagging two issues that would need to be addressed in the course of such development. One is that to be applicable in practice, such developments will need to offer functional markers for whatever moral criteria they invoke. After all, an impeccable moral criterion will be of no use if we are at a loss as to which advanced digital systems satisfy it. The other is that allowing some forms of digital suffering may precipitate a descent down a causal slippery slope that ends with us allowing egregious forms of digital suffering. Thus, in evaluating proposed moral relaxations, we should be mindful of their likely effects in practice, not just their moral status under ideal conditions.

5. Conclusion

The problem of digital suffering is a high stakes moral problem. Formidable epistemic obstacles stand in the way of solving it. To overcome these obstacles, we proposed AMP. AMP uses a functional connectedness test to generate epistemic access to a limited class of digital minds, thereby putting us in a position to know what experiences digital systems are having and to prevent digital minds from suffering. We have seen that AMP likely precludes some prominent approaches to advanced digital

[46]Cf. MacAskill and Ord (2020).

systems. Fortunately, AMP fits well with and therefore provides reason to prioritize one approach to developing advanced digital systems, namely whole brain emulation. That approach enjoys independent motivation and offers a path to some forms of superintelligence. Even so, it is worth considering ways to enhance AMP by making it safer or more viable. To that end, we explored four ways of modifying AMP: allowing modified human inputs to the functional connectedness test, relaxing the operative notion of functional connectedness, making AMP recursive, and relaxing AMP's moral aims. While each of these approaches exhibited promise, their prospects are uncertain and require further investigation. There is also the task of developing and evaluating alternative solutions to the problem of digital suffering. In short, while AMP points to a way forward on the problem of digital suffering, much work remains if we are to converge upon and implement an optimal solution to the problem of digital suffering while there is still time.

Acknowledgements

For helpful feedback, we are grateful to Ali Ladak, Andrew Y. Lee, Geoffrey Lee, Nuño Sempere, participants in the Future of Humanity Institute's digital minds reading group, and anonymous reviewers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Future of Humanity Institute at the University of Oxford, the John Templeton Foundation [grant no 61516], the Sentience Institute, and Utrecht University. The views expressed are the authors' and do not necessarily reflect those of funders.

References

Aizawa, K. 2007. "Understanding the Embodiment of Perception." *Journal of Philosophy* 104 (1): 5–25. https://doi.org/10.5840/jphil2007104135.

Alter, T., and D. Pereboom. 2019. "Russellian Monism", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), E.N. Zalta (ed.), URL=<https://plato.stanford.edu/archives/fall2019/entries/russellian-monism/>.

Arute, F., K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, et al. 2019. "Quantum Supremacy Using a Programmable Superconducting Processor." *Nature* 574 (7779): 505–510. https://doi.org/10.1038/s41586-019-1666-5.

Avramides, A. 2019. "Other Minds, The Stanford Encyclopedia of Philosophy (Winter 2020 Edition)", E. N. Zalta (ed.), URL=<https://plato.stanford.edu/archives/win2020/entries/other-minds/>.

Bar-On, Y. M., R. Phillips, and R. Milo. 2018. "The Biomass Distribution on Earth." *Proceedings of the National Academy of Sciences* 115 (25): 6506–6511. https://doi.org/10.1073/pnas.1711842115.

Berry, C. J., B. Bell, A. Jatkowski, J. Surprise, J. Isakson, O. Geva, B. Deskin, et al. 2020. "2.7 IBM z15: A 12-Core 5.2 GHz Microprocessor." 2020 IEEE International Solid-State Circuits Conference-(ISSCC), 54-56.

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

Brady, M. 2018. *Suffering and Virtue*. New York: OUP.

Brady, M. 2019. "Why Suffering is Essential to Wisdom." *The Journal of Value Inquiry* 53 (3): 467–469. https://doi.org/10.1007/s10790-019-09707-3.

Brouwer, G. J., and D. J. Heeger. 2013. "Categorical Clustering of the Neural Representation of Color." *Journal of Neuroscience* 33 (39): 15454–15465.

Carls-Diamante, S. 2017. "The Octopus and the Unity of Consciousness." *Biology & Philosophy* 32 (6): 1269–1287. https://doi.org/10.1007/s10539-017-9604-0.

Chalmers, D. 1996. *The Conscious Mind*. New York: Oxford University Press.

Chalmers, D. 2010a. *The Character of Consciousness*. New York: OUP.

Chalmers, D. 2010b. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17: 7–65.

Chalmers, D. 2018. "The Meta-Problem of Consciousness." *Journal of Consciousness Studies* 25 (9-10): 6–61.

Coghill, R., C. Sang, J. Maisog, and M. Iadarola. 1999. "Pain Intensity Processing Within the Human Brain: A Bilateral, Distributed Mechanism." *Journal of Neurophysiology* 82 (4): 1934–1943. https://doi.org/10.1152/jn.1999.82.4.1934.

Dalbey, B., and B. Saad. forthcoming. "Internal Constraints for Phenomenal Externalists: A Structure Matching Theory." *Synthese*.

Dickenson, A. H., B. L. Kieffer, et al. 2013. "Opioids: Basic Mechanisms." In *Wall and Melzack's Textbook of Pain 6th ed*, edited by S. B. McMahon, 413–428. Philadelphia, PA: Elsevier.

Dresler, M., A. Sandberg, C. Bublitz, K. Ohla, C. Trenado, A. Mroczko-Wasowicz, S. Kühn, and D. Repantis. 2019. "Hacking the Brain: Dimensions of Cognitive Enhancement." *ACS Chemical Neuroscience* 10 (3): 1137–1148. https://doi.org/10.1021/acschemneuro.8b00571.

Dretske, F. 1995. *Naturalizing the Mind*. London: MIT Press.

Eth, D., J. C. Foust, and B. Whale. 2013. "The Prospects of Whole Brain Emulation Within the Next Half-Century." *Journal of Artificial General Intelligence* 4 (3): 130.

Forder, L., J. Bosten, X. He, and A. Franklin. 2017. "A Neural Signature of the Unique Hues." *Scientific Reports* 7: 42364.

Gloor, L. 2016. "Suffering-Focused AI Safety: Why "Fail-Safe" Measures Might be Our Top Intervention. Foundational Research Institute." *Report FRI* 16: 1.

Good, I. J. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers, vol 6*, edited by F. Alt, and M. Rubinoff, 31–88. New York: Academic Press.

Grace, K., J. Salvatier, A. Dafoe, B. Zhang, and O. Evans. 2018. "Viewpoint: When will AI Exceed Human Performance? Evidence from AI Experts." *Journal of Artificial Intelligence Research* 62: 729–754. https://doi.org/10.1613/jair.1.11222.

Hanson, R. 2016. *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. New York: OUP. https://doi.org/10.1093/oso/9780198754626.001.0001.

Hanson, R. 2022. "J. Phil. Critique of Em." URL: https://www.overcomingbias.com/2022/02/j-phil-critique-of-ems.html.

Hill, R. G., 2013. "Analgesic Drugs Under Development." In *Wall and Melzack's Textbook of Pain 6th ed*, edited by S. B. McMahon, et al. 556. Philadelphia, PA: Elsevier.

Hubinger, E. 2020. "An Overview of 11 Proposals for Building Safe Advanced AI." arXiv preprint arXiv:2012.07532.

Irving, G., P. Christiano, and D. Amodei. 2018. "AI Safety Via Debate." arXiv preprint arXiv:1805.00899.

Jacobson, Hilla. 2021. "The Role of Valence in Perception: An ARTistic Treatment." *The Philosophical Review* 130 (4): 481–531. https://doi.org/10.1215/00318108-9263939.

Jiang, L., A. Stocco, D. M. Losey, J. A. Abernethy, C. S. Prat, and R. P. Rao. 2019. "Brainnet: A Multi-Person Brain-To-Brain Interface for Direct Collaboration Between Brains." *Scientific Reports* 9 (1): 1–11. https://doi.org/10.1038/s41598-018-37186-2.

Kay, G. 2021. "Elon Musk says Neuralink could Start Planting Computer Chips in Humans Brains within the Year." URL: https://www.businessinsider.com/elon-musk-predicts-neuralink-chip-human-brain-trials-possible-2021-2021-2?r=US&IR=T.

Klein, Colin. 2015. "What Pain Asymbolia Really Shows." *Mind; A Quarterly Review of Psychology and Philosophy* 124 (494): 493–516. https://doi.org/10.1093/mind/fzu185.

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. London: Penguin.

Kuypers, K. P. C., J. Riba, M. De La Fuente Revenga, S. Barker, E. L. Theunissen, and J. G. Ramaekers. 2016. "Ayahuasca Enhances Creative Divergent Thinking while Decreasing Conventional Convergent Thinking." *Psychopharmacology* 233 (18): 3395–3403. https://doi.org/10.1007/s00213-016-4377-8.

Lewis, D. 2009. "Ramseyan Humility." In *Conceptual Analysis and Philosophical Naturalism*, edited by D. Braddon-Mitchel, and R. Nola. Cambridge: MIT Press.

MacAskill, W., and T. Ord. 2020. "Why Maximize Expected Choice-Worthiness? 1." *Noûs* 54 (2): 327–353.

Mandelbaum, E. 2022. "Everything and More: The Prospects of Whole Brain Emulation." *Journal of Philosophy* 119 (8): 444–459.

Mandik, P. 2017. "Robot pain." In *The Routledge Handbook of Philosophy of Pain*. Abingdon, Oxon: Routledge.

Mendelovici, A. 2018. *The phenomenal basis of intentionality*. New York: OUP.

Ord, T. 2020. *The Precipice: Existential Risk and the Future of Humanity*. New York: Hachette Books.

Parfit, D. 1984. *Reasons and Persons*. Oxford: OUP.

Pautz, A. 2014. "The Real Trouble with Phenomenal Externalism: New Empirical Evidence for a Brain-Based Theory of Consciousness." In *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, edited by R. Brown, 237–298. New York: Springer.

Pautz, A. 2019. "What is the Integrated Information Theory of Consciousness?" *Journal of Consciousness Studies* 26 (1-2): 188–215.

Russell, S. J., and P. Norvig. 2021. *Artificial Intelligence: A Modern Approach (4th ed.)*. Hobokin, NJ: Pearson.

Saad, B. 2019a. "A Teleological Strategy for Solving the Meta-Problem of Consciousness." *Journal of Consciousness Studies* 26 (9-10): 205–216.

Saad, B. 2019b. "In Search of a Tracking Theory of Consciousness." Manuscript. URL: https://docs.google.com/document/d/1zoKLQ-ZHHekuFUZMiG-K789i7Cp0EmgkgoIjATlqqeg/edit?usp=sharing.

Sandberg, A. 2014a. "Monte Carlo Model of Brain Emulation Development.".

Sandberg, A. 2014b. "Ethics of Brain Emulations." *Journal of Experimental & Theoretical Artificial Intelligence* 26 (3): 439–457. https://doi.org/10.1080/0952813X.2014.895113.

Sandberg, A., and N. Bostrom. 2008. "Whole Brain Emulation: A Roadmap." Technical report 2008–3, future for humanity institute, Oxford.

Schneider, S. 2019. *Artificial You: AI and the Future of Your Mind*. Princeton, NJ: Princeton University Press.

Schwitzgebel, E., and M. Garza. 2015. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39: 98–119.

Seth, A. K., and T. Bayne. 2022. "Theories of Consciousness." *Nature Reviews Neuroscience* 23: 439–452.

Shulman, C., and N. Bostrom. 2020. "Sharing the World with Digital Minds." manuscript.

Singer, P. 1975. *Animal Liberation*. New York: Avon Books.

Tomasik, B. 2017. *Artificial Intelligence and Its Implications for Future Suffering*. Basel, Switzerland: Foundational research institute.

Tononi, G. 2008. "Consciousness as Integrated Information: A Provisional Manifesto." *The Biological Bulletin* 215 (3): 216–242. https://doi.org/10.2307/25470707.

Treffert, D. A. 2009. "The Savant Syndrome: An Extraordinary Condition. A Synopsis: Past, Present, Future." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364: 1351–1357.

Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge: MIT Press.

Wiener, Norbert. 1960. "Some Moral and Technical Consequences of Automation." *Science* 131 (3410): 1355–1358. https://doi.org/10.1126/science.131.3410.1355.

Yudkowsky, E. 2004. "Coherent Extrapolated Volition." Singularity Institute for Artificial Intelligence.